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

Anonymous Author(s)
Affiliation
Address
email

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. However, their integration into multimodal models remains a challenge, especially considering their high dimensionality. In this work, we present a novel bimodal model, MOJO, that jointly learns representations of bulk RNA-seq and DNA methylation leveraging self-supervision from Masked Language Modeling. We use 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.

1 Introduction

The growing availability of high-throughput technologies has revolutionized molecular research, generating extensive genomic, transcriptomic, and epigenomic data that hold immense potential for personalized medicine [18, 35, 12]. The integration of these diverse data sources remains a significant challenge, especially when modalities may be missing in clinical applications. The high dimensionality of each modality makes classic machine learning ineffective. Consequently, there is a growing tendency to first learn data representations using self-supervised approaches. Foundation models have emerged as powerful tools to learn effective embeddings for biological and clinical tasks [13, 6]. These models often leverage the transformer architecture [40], which is limited by the quadratic memory scaling of its attention mechanism. To handle long-range sequences, recent models have integrated convolutional blocks [3] or state-space models [30]. In this paper, we introduce MOJO (Multi-Omics Joint representation learning), a model that learns joint embeddings of bulk RNA-seq and DNA methylation from The Cancer Genome Atlas (TCGA, `https://portal.gdc.cancer.gov/`)) through bimodal masked language modeling. We show that MOJO's embeddings lead to state-of-the-art performance in pan-cancer classification, survival analysis, and subtype clustering. We also present a framework that uses an auxiliary mutual information loss to preserve performance when a modality is absent at test time. *Code will be made available upon acceptance.*

2 Related Works

Omics representation learning has evolved from statistical methods like PCA [19] to deep learning architectures such as Masked Auto-Encoders [16] and Mixture-of-Experts [28]. In line with foundation models for single-cell transcriptomics [11], [15] developed a transformer-based model for

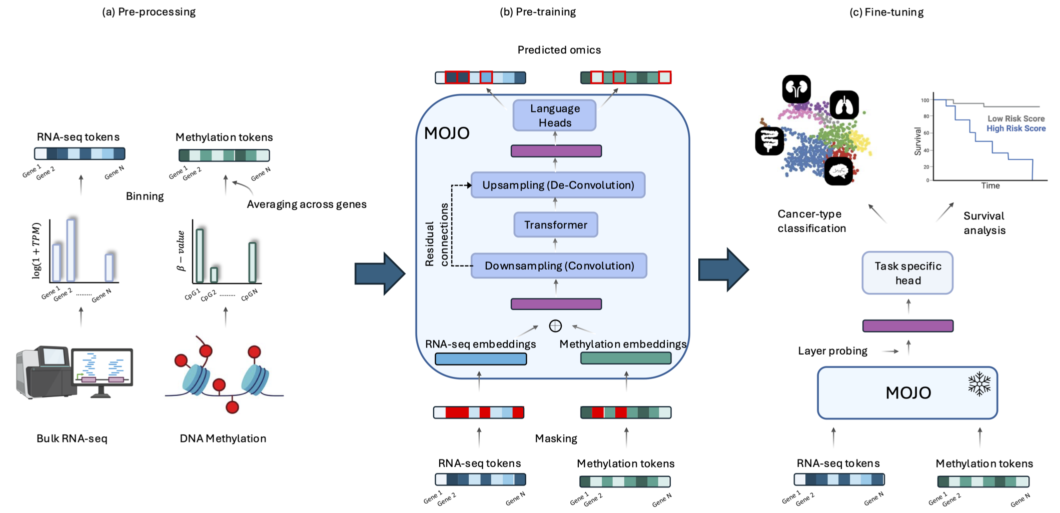

Figure 1: MOJO pipeline. (a) RNA-seq and DNA methylation are processed and tokenized. (b) MOJO, a hybrid convolution-transformer model, is pre-trained via bimodal masked language modeling. (c) The learned embeddings are used to fine-tune downstream models.

bulk RNA-seq. Multi-modal integration is often performed using late integration, where sources are encoded separately before being aggregated via concatenation, element-wise operations [38], or cross-attention [14]. Variational auto-encoders [22] have also been widely used for multi-omics integration, either for single-cell omics [7, 2, 37] or bulk omics [4].

Handling missing modalities is crucial for clinical applicability. Common approaches include data-level imputation [8] and model-level adjustments like fusion or knowledge distillation [32]. Training strategies such as modality dropout [23] are also employed to simulate missing data scenarios. Our work adapts a test-time-adaptation technique from [31] that uses mutual information to improve the robustness of their model to missing modalities.

3 Multi-Omics Joint Representation Learning

Modalities Alignment Bulk RNA-seq provides gene expression estimates ($X_{rna} \in \mathbb{R}^{N_{genes}}$, with typically $N_{genes} \sim 10^4$), to which we apply an $x \mapsto \log_{10}(1+x)$ transformation. DNA methylation data consists of beta values for numerous CpG sites ($X_{sites\_meth} \in [0,1]^{N_{sites}}$), obtained through the Illumina Infinium HumanMethylation450 (450K) BeadChip array [5] (so $N_{sites} \sim 450,000$). We align these modalities by averaging the methylation beta values of all sites associated with a given gene (e.g., within its promoter region or gene body) to obtain a single methylation value per gene, $X_{meth} \in \mathbb{R}^{N_{genes}}$. A bimodal sample is thus a vector $X = (X_{rna}, X_{meth}) \in \mathbb{R}^{2N_{genes}}$.

Tokenization We tokenize each component of the feature vector X by binning its values on linear scales. The token ID for a given value is its corresponding bin ID. After tokenization, a sample is represented as a vector of integers $\widetilde{X} = (\widetilde{X}_{rna}, \widetilde{X}_{meth})$.

Model Architecture and Pre-training To learn representations, we propose a model combining convolution and transformer blocks, inspired by architectures for long-range genomic dependencies [3, 25]. As shown in Figure 1, each omic token is passed through embedding layers and summed with a shared gene embedding (initialized with the *Gene2Vec* method [43]), which acts as a positional encoding. This bimodal embedding is downsampled by a convolutional tower before being fed to a transformer block, significantly reducing computational cost. The original sequence length is restored using a deconvolutional tower with residual connections. Separate language modeling heads predict the binned gene expression and methylation values. The model is pre-trained through self-supervision using multimodal masked language modeling. For each sequence, 15% of tokens are corrupted (80% masked, 10% randomized, 10% unchanged). We optimize a multimodal negative log-likelihood loss, $\mathcal{L}_{\text{multimodal MLM}} = -\sum_{m \in \mathcal{M}} \sum_{i \in \mathcal{M}_m} \log(p_m(i)_{\bar{X}_m(i)})$, where $\mathcal{M}_m$ is the set of masked token indices for modality m. We pre-trained *MOJO* on 9,252 paired samples from TCGA over 17,116 genes. Further pre-training details are in Appendix A.

4 Evaluation on Downstream Tasks

We evaluate MOJO's representations on cancer-type classification and survival analysis, comparing against unimodal models (*BulkRNABert* [15] for bulk RNA-seq and its counterpart for DNA methylation that we developed in our work and called *MethFormer*), late integration schemes (aggregating embeddings of two aforementioned encoders, see figure in Appendix B for more details), *CustOmics* [4] (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), and *MOFA* [1]. The quality of MOJO's embeddings is further confirmed in zero-shot classification and clustering tasks especially on breast cancer sub-typing (see Appendix C.4).

4.1 Cancer-Type Classification

We fine-tune MOJO's embeddings (extracted from the last attention layer and averaged across the sequence dimension) with a small MLP for 33-way pan-cancer classification. *BulkRNABert*, *MethFormer*, and *MOJO* are further fine-tuned in addition to training the MLPs using the parameter-efficient method IA^3 [26]. Table 1 presents the cancer-type classification results on the pan-cancer TCGA dataset, split into 80% for training and 20% for testing (averaged over 5 seeds). *MOJO* achieves state-of-the-art results with both modalities, outperforming *CustOmics* and Late Integration methods. *MOJO* also exceeds unimodal transformers (*BulkRNABert* and *MethFormer*). Furthermore, probing MOJO's last attention layer with an SVM (*MOJO (probing)*) shows a clear performance increase over *CustOmics(probing)*, indicating stronger predictive capacity from its masked language modeling representations.

Table 1: Cancer type classification

Model	Test weighted-F1
BulkRNABert	0.943 ± 0.004
MethFormer	0.931 ± 0.006
MOFA	0.852 ± 0.007
Late integration (concatenation)	0.945 ± 0.007
Late integration (cross-attention)	0.945 ± 0.002
CustOmics (probing)	0.911 ± 0.088
MOJO (probing)	0.945 ± 0.006
CustOmics (end-to-end)	0.946 ± 0.006
MOJO (no pre-training)	0.891 ± 0.006
MOJO	**0.952 ± 0.006**

4.2 Survival Analysis

Table 2: Pan-cancer survival analysis

Model	C-index
BulkRNABert	0.749 ± 0.003
MethFormer	0.736 ± 0.006
MOFA	0.648 ± 0.037
CustOmics	0.686 ± 0.018
Late integration	0.756 ± 0.004
MOJO	**0.771 ± 0.006**

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 for individuals who have cancer, specifically the time from diagnosis until death from right-censored datasets. We use adaptations of Cox proportional model [10] to the deep learning setting [21, 9] and thus employ negative partial Cox-log-likelihood as loss for model training. Table 2 reports the test C-indexes [17] of the benchmarked models and shows that *MOJO* outperforms other methods, demonstrating the strength of its learned representations for prognosis. Kaplan-Meier curves are also provided in Appendix C.5, showing better patient stratification with *MOJO*.

5 Robustness to Missing Modalities

In clinical settings, modalities can be missing. *MOJO* can inherently handle missing data by replacing a modality's input with `<MASK>` tokens.

Missing modalities: fine-tuning In the context of Ovarian (OV) cancer-subtyping in TCGA (4 classes: differentiated, immunoreactive, mesenchymal, and proliferative [41]), one only gets access to RNA-seq samples. A bimodal pre-trained *MOJO* model is thus fine-tuned on this task with $(X_{rna}, None)$ as input and gets better performance than *BulkRNABert* while being faster to train (Figure 2). We also conduct this experiment by pre-training another *MOJO* model 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). This further improves the performance on OV sub-typing.

Missing modalities: test-time We aim for a model trained on bimodal data to maintain performance when one modality is absent at test time (we simulate the absence of either RNA-seq or methylation by dropping it from x% of test pairs). To improve robustness, we adapt a technique from [31] and incorporate an auxiliary mutual information (MI) loss during fine-tuning. The goal is to make the model's prediction $f_\theta(x; m)$ for an input x with modality m independent of the modality $m \in \mathcal{D}_{modality} = \{rna + meth, rna, meth\}$ seen as a random variable. We achieve this by minimizing the MI between the model's output and the modality set: $\mathcal{L}_{aux} = \mathbb{E}_{m \in \mathcal{D}_{modality}} [MI(f_\theta(x, m), m)]$. The total loss becomes $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{aux}}$ (a detailed algorithm is available in Appendix D). Results when RNA-seq is dropped are provided in Figure 3 (similar results when dropping DNA methylation are provided in Appendix D). When tested on the cancer-type classification task, a standard *MOJO* model's performance drops significantly when a modality is removed from 100% of the test samples (e.g., weighted-F1 from 0.952 to 0.538 when RNA-seq is dropped). Fine-tuning with the MI auxiliary loss largely mitigates this drop (recovering to 0.916), achieving performance close to that of a unimodal model trained only on the available data (0.943), without sacrificing performance in the bimodal case.

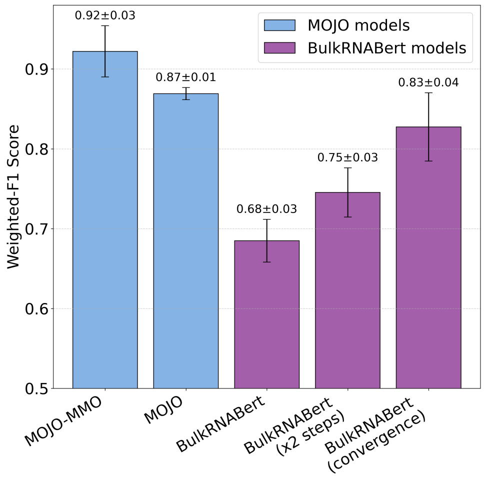

Figure 2: Ovarian cancer sub-typing: MOJO outperforms *BulkRNABert* while being faster to fine-tune. (*BulkRNABert* models bars from left to right: same fine-tuning budget as *MOJO*, $\times 2$ fine-tuning steps, and until convergence).

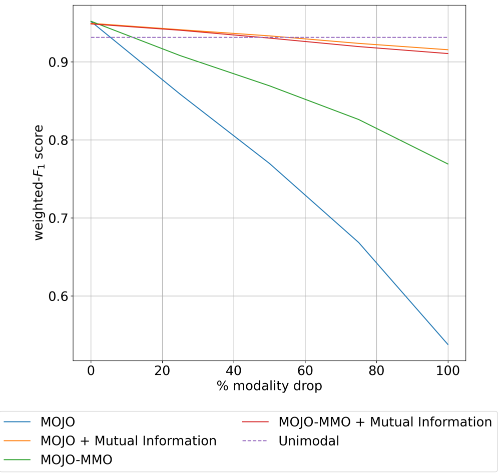

Figure 3: Performance when dropping RNA-seq. Test weighted-F1 score is reported as a function of the percentage of dropped RNA-seq samples in the test set.

6 Conclusion

We introduced MOJO, a novel architecture for learning joint representations of bulk RNA-seq and DNA methylation via bimodal masked language modeling. Its hybrid convolution-attention design efficiently handles high-dimensional omics data. The learned embeddings achieve state-of-the-art performance on cancer-type classification and survival analysis, outperforming unimodal and late-integration approaches. Furthermore, by incorporating a mutual information-based auxiliary loss during fine-tuning, we demonstrate that our model can maintain robust performance even when a modality is missing at test time, enhancing its clinical applicability.

References

[1] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[2] Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.

[3] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[4] Hakim Benkirane, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. Customics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, 19(3):e1010921, 2023.

[5] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

[6] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pages 2025–02, 2025.

[7] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[8] Qianqian Chen, Jiadong Zhang, Runqi Meng, Lei Zhou, Zhenhui Li, Qianjin Feng, and Dinggang Shen. 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.

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

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

[11] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

[12] Xiaofeng Dai and Li Shen. Advances and trends in omics technology development. *Frontiers in Medicine*, 9:911861, 2022.

[13] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

[14] Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Maša Roller, Bernardo de Almeida, Christopher Blum, Lorenz Hexemer, Stefan Laurent, Maren Lang, Thomas Pierrot, et al. Multi-modal transfer learning between biological foundation models. *Advances in Neural Information Processing Systems*, 37:78431–78450, 2024.

[15] Maxence Gélard, Guillaume Richard, Thomas Pierrot, and Paul-Henry Cournède. 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*, pages 384–400. PMLR, 15–16 Dec 2025.

[16] Baptiste Gross, Antonin Dauvin, Vincent Cabeli, Virgilio Kmetzsch, Jean El Khoury, Gaëtan Dissez, Khalil Ouardini, Simon Grouard, Alec Davi, Regis Loeb, 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.

[17] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[18] Won Jin Ho, Rossin Erbe, Ludmila Danilova, Zaw Phyo, Emma Bigelow, Genevieve Stein-O'Brien, Dwayne L Thomas, Soren Charmsaz, Nicole Gross, Skylar Woolman, 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.

[19] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[20] Ameya Joshi, Raphael Boige, Lee Zamparo, Ugo Tanielian, Juan Jose Garau-Luis, Michail Chatzianastasis, Priyanka Pandey, Janik Sielemann, Alexander Seifert, Martin Brand, et al. A long range foundation model for zero-shot predictions in single-cell and spatial transcriptomics data. 2025.

[21] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] Gautam Krishna, Sameer Dharur, Oggi Rudovic, Pranay Dighe, Saurabh Adya, Ahmed Hussen Abdelaziz, and Ahmed H Tewfik. 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)*, pages 8240–8244. IEEE, 2024.

[24] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[25] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pages 1–13, 2025.

[26] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[27] Shihao Ma, Andy GX Zeng, Benjamin Haibe-Kains, Anna Goldenberg, John E Dick, and Bo Wang. Integrate any omics: Towards genome-wide data integration for patient stratification. *arXiv preprint arXiv:2401.07937*, 2024.

[28] Xiangyu Meng, Xue Li, Qing Yang, Huanhuan Dai, Lian Qiao, Hongzhen Ding, Long Hao, and Xun Wang. Gene-moe: A sparsely gated prognosis and classification framework exploiting pan-cancer genomic information. *arXiv preprint arXiv:2311.17401*, 2023.

[29] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, 2009.

[30] Matvei Popov, Aymen Kallala, Anirudha Ramesh, Narimane Hennouni, Shivesh Khaitan, Rick Gentry, and Alain-Sam Cohen. Leveraging state space models in long range genomics. *arXiv preprint arXiv:2504.06304*, 2025.

[31] Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] Pramit Saha, Divyanshu Mishra, Felix Wagner, Konstantinos Kamnitsas, and J Alison Noble. Examining modality incongruity in multimodal federated learning for medical vision and language-based disease detection. *arXiv preprint arXiv:2402.05294*, 2024.

[33] Alex Sánchez, José Fernández-Real, Esteban Vegas, Francesc Carmona, Jacques Amar, Remy Burcelin, Matteo Serino, Francisco Tinahones, M Carmen Ruíz de Villa, Antonio Minãrro, 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*, pages 29–41. Springer, 2012.

[34] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[35] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[36] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[37] Xinming Tu, Zhi-Jie Cao, Chen-Rui Xia, Sara Mostafavi, and Ge Gao. Cross-linked unified embedding for cross-modality representation learning. In *Advances in Neural Information Processing Systems*, 2022.

[38] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021.

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Roel GW Verhaak, Pablo Tamayo, Ji-Yeon Yang, Diana Hubbard, Hailei Zhang, Chad J Creighton, Sian Fereday, Michael Lawrence, Scott L Carter, Craig H Mermel, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *The Journal of clinical investigation*, 123(1), 2012.

[42] Xiaoyu Zhang, Yuting Xing, Kai Sun, and Yike Guo. Omiembed: a unified multi-task deep learning framework for multi-omics data. *Cancers*, 13(12):3047, 2021.

[43] Quan Zou, Pengwei Xing, Leyi Wei, and Bin Liu. 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 3: 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

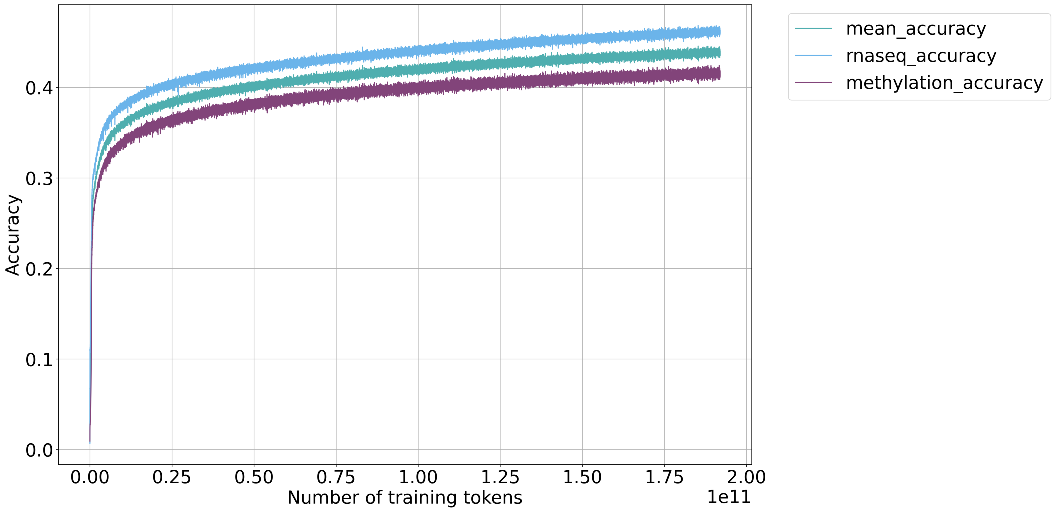

Figure 4: Bimodal masked language modeling pre-training curves of the *MOJO* architecture. The training reconstruction accuracy is represented of each omic separately as well as the average reconstruction accuracy among the different omics.

B Late integration

We refer to *Late integration* as the bimodal integration resulting from the fusion of embeddings extracted from unimodal models. More precisely, we refer to *Late Integration (concatenation)* as the concatenation of the embeddings from *BulkRNABert* (for RNA-seq) and *MethFormer* (for DNA 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 5.

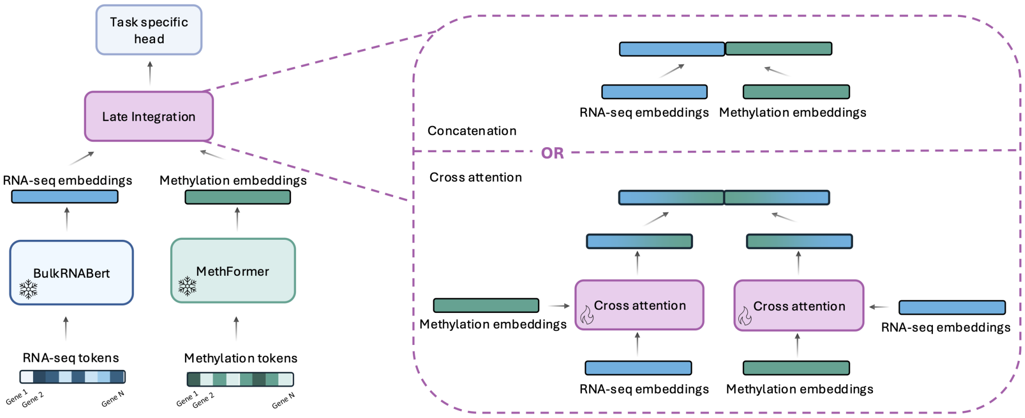

Figure 5: Late integration architecture. RNA-seq and Methylation embeddings are obtained from pre-trained transformer based encoders (respectively *BulkRNABert* and *MethFormer*) and are fused either by concatenation or by a two-steps cross-attention mechanism.

C Downstream tasks dataset and benchmarks

C.1 Pan-cancer classification dataset

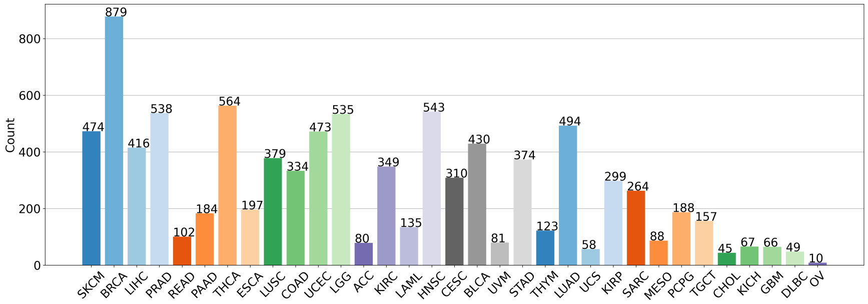

Figure 6: Pan-cancer classification label distribution.

C.2 Classification and survival analysis exhaustive benchmarks

In addition to Table 1 (cancer-type classification) and Table 2, a more exhaustive benchmark including other representation models for RNA-seq and DNA methylation has been performed:

- Multiple Factor Analysis (MFA) [33], using a latent space of dimension 256.
- Non-negative Matrix Factorization (NMF) [24], with the same latent space dimension as for MFA.
- *OmiEmbed* [42]: a unified multi-task deep learning framework for multi-omics data based on Variational Auto-Encoders [22] from early integrated omics.
- *IntegrAO* [27]: an unsupervised framework based on Graph Neural Networks [34] 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 4 and Table 5.

For the classification task, in addition to the weighted F_1 score, we also report the macro F_1 score. For the survival analysis task, 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.

C.3 Fine-tuning training times

We report in Table 6 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 $100\times$ 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.

Table 4: 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 5: 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 6: 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**

C.4 Zero-shot pan-cancer and breast cancer sub-typing and clustering

To further evaluate *MOJO*'s learned embeddings in a fully unsupervised manner, we assess their zero-shot classification and clustering capabilities on PAM50 breast cancer sub-typing (Luminal A, Luminal B, Basal, and HER2) [29] and the Pan-cancer dataset from section **??**. First, zero-shot classification uses a *k*-nearest neighbors model ($k = 5$), evaluated by accuracy, to assess embedding quality without fine-tuning, inspired by [20]. Second, Leiden clustering [36] is performed in the embedding space, with Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) as metrics. We primarily compare the effectiveness of *MOJO*'s joint modeling against late integration embeddings for bimodal data.

Results Zero-shot classification and clustering results are shown in Table 7, showing better performance when using *MOJO* embedding than late integration and *CustOmics*. We present in Figure 7 t-SNE [39] plots of both embeddings in the pan-cancer setting, reflecting that *MOJO* embeddings more effectively separate the cohorts.

Table 7: 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

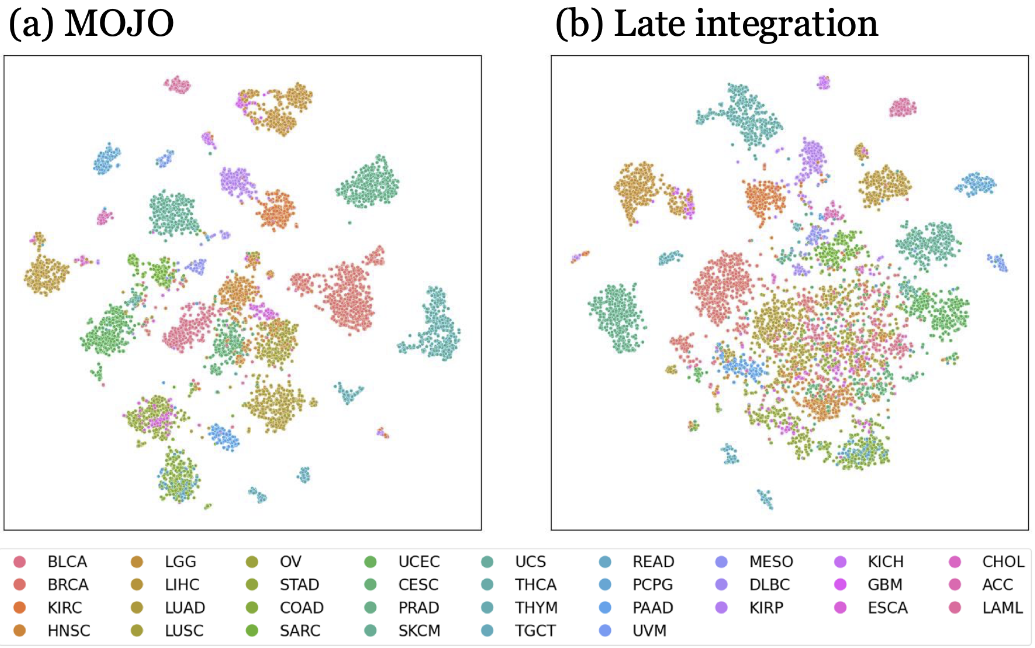

Figure 7: Pan-cancer version of the t-SNE representation of *MOJO* and *Late integration* embeddings, colored by cancer-type.

329 ## C.5 Kaplan-Meier curves

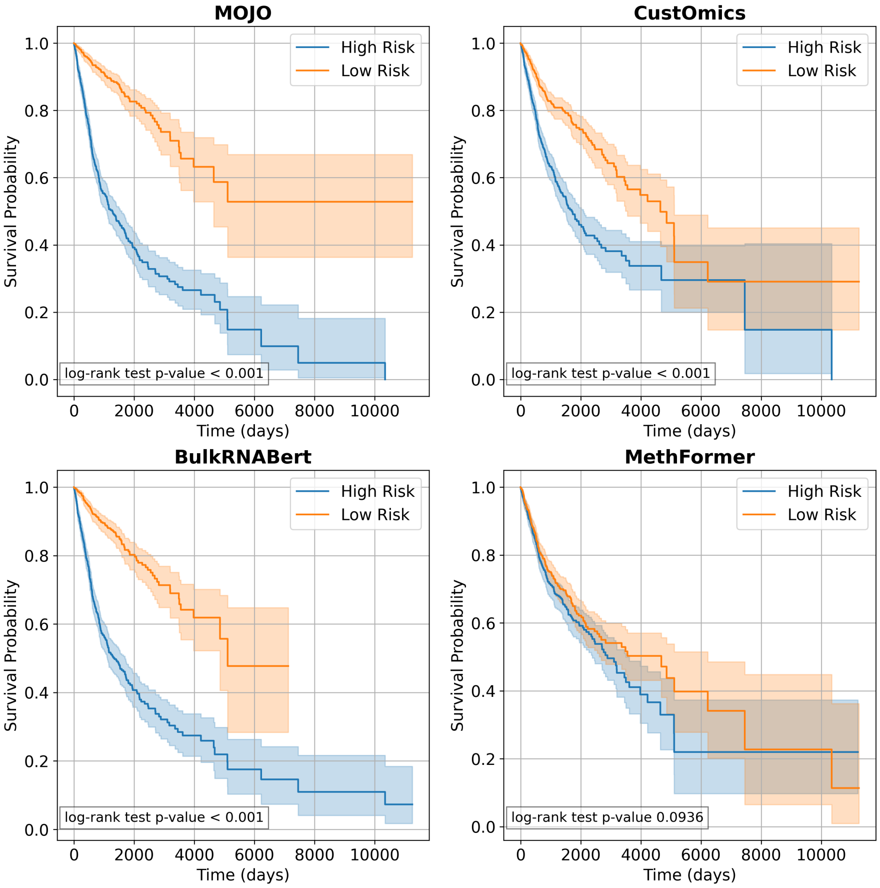

Figure 8: Kaplan-Meier curve for pan-cancer survival models for four models: *MOJO*, *CustOmics*, *BulkRNABert*, *MethFormer*.

D Missing modalities experiments

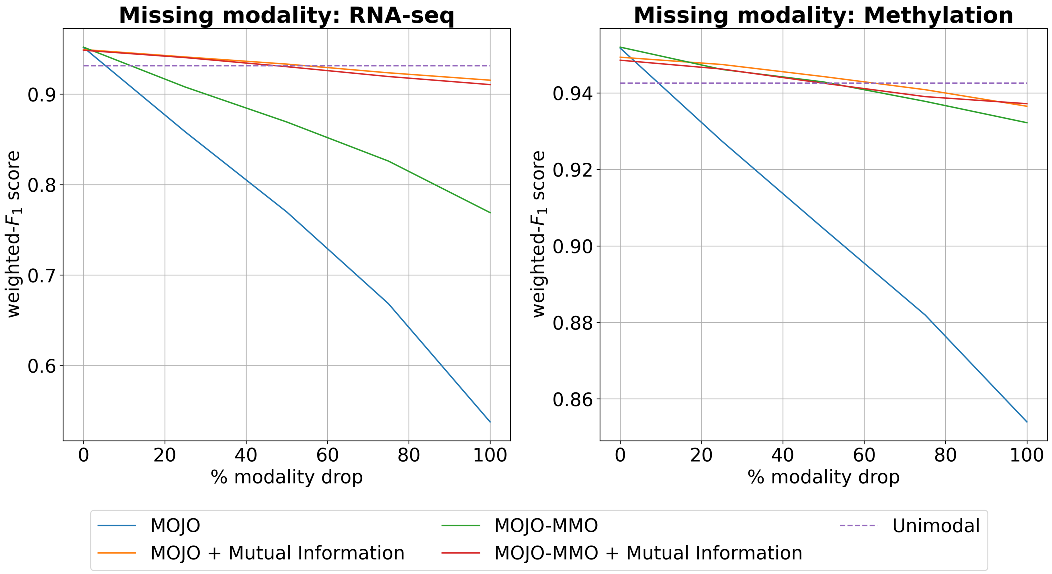

Figure 9: Missing modalities experimental results. Test weighted-F_1 score for the pan-cancer classification is reported for different methods to handle the absence of a modality in x% of the samples (left: RNA-seq, right: Methylation). Unimodal models are respectively *MethFormer* and *BulkRNABert* when RNA-seq or Methylation is missing.

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 = $CrossEntropy(f_\theta(X), y) + \lambda * MILoss$
```

Table 8: Missing modalities experiment: cancer type classification

Model	Add mutual information	Drop modality (test time)	test 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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract introduces our model *MOJO* for RNA-seq and DNA methylation representation learning which is then benchmarked against other models on various tasks, showing SOTA performance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Although the papers discusses clinical applicability through the problem of missing modalities, further validation on external datasets (other than TCGA), could also help validating the model for clinical applications.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Even though the paper tackles the clinical applicability of the model considering possible missing modalities, no external dataset other than TCGA has been used to fully validate its clinical relevance.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The model and training hyperparameters are provided either in the main body of the paper or in the appendix. The dataset (TCGA) is open-access and thus allows for reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be made available upond acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide *MOJO* hyperparatemers as well as pre-training setting. For fine-tuning, we detail how the dataset is split and how the metrics are computed for each task.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each downstream task, we randomly split the dataset 5 different times and report the mean test metric as well as the standard deviation.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As stated in the pre-training paragraph of the paper, we used a TPU v4-8 for our experiments. Also, training times and maximum batch sizes are reported in the Appendix in the case of classification models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No deviation from the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discussed the application of the model in real clinical applications when considering the possibility of missing modalities.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are credited in this manuscript

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.