PLOS Computational Biology

RESEARCH ARTICLE

Transcriptomic-guided whole-slide image classification for molecular subtype identification

Weiwen Wang[1]*, **Xiwen Zhang**[2], **Yuanyan Xiong**[3]

1 Department of Mathematics, College of Information Science and Technology, Jinan University, Guangzhou, Guangdong, China, **2** Department of Bioinformatics, College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou, Guangdong, China, **3** Department of Biochemistry, Key Laboratory of Gene Engineering of the Ministry of Education, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China

* wangww29@jnu.edu.cn

OPEN ACCESS

Citation: Wang W, Zhang X, Xiong Y (2026) Transcriptomic-guided whole-slide image classification for molecular subtype identification. PLoS Comput Biol 22(2): e1013950. https://doi.org/10.1371/journal.pcbi.1013950

Editor: Virginie Uhlmann, University of Zurich Faculty of Mathematics and Science: Universitat Zurich Mathematisch-Naturwissenschaftliche Fakultat, SWITZERLAND

Received: August 28, 2025

Accepted: January 27, 2026

Published: February 9, 2026

Data availability statement: The three datasets CRC-DX, CRC-KR, STAD-DX were downloaded from https://zenodo.org/records/2530835 and

Abstract

Recent advancements in computational pathology have greatly improved automated histopathological analysis. A compelling question in the field is how morphological traits are associated with genetic characteristics or molecular phenotypes. Here we propose TEMI, a novel framework for molecular subtype classification of cancers using whole-slide images (WSIs), augmented with transcriptomic data during training. TEMI aims to extract molecular-level signals from WSIs and make efficient use of available multimodal data. To this end, TEMI introduces a patch fusion network that captures dependencies among local patches of gigapixel WSIs to produce global representations and aligns them with transcriptomic embeddings attained from a masked transcriptomic autoencoder. TEMI achieves superior performance compared with existing methods in molecular subtype classification, owing to its effective integration of transcriptomic information achieved by the two developed alignment strategies. Guided by discriminative transcriptomic data, TEMI learns invariant WSI representations, while morphological features also enhance gene expression prediction. These findings suggest that histological features encode latent molecular signals, highlighting the interplay between the tumor microenvironment and cancer transcriptomics. Our study demonstrates how multimodal learning can bridge morphology and molecular biology, providing an effective tool to advance precision medicine.

Author summary

Cancer's intrinsic heterogeneity poses significant challenges to effective treatment. Therefore, molecular subtyping that stratifies patients into subgroups based on molecular and genetic distinctions serves as a cornerstone of precision medicine and personalized healthcare. Owing to the strong capability of artificial intelligence (AI) particularly deep learning in discovery of patterns, developing AI models to identify molecular subtypes directly from routinely available haematoxylin–eosin

https://zenodo.org/records/2532612. The well-trained ResNet18 for deep feature extraction and tumor detection was obtained from https://github.com/jnkather/MSIfromHE. The HTseq-FPKM profiles for all cancer cohorts and whole-slides images of GBM-DX were accessed via https://portal.gdc.cancer.gov. Molecular subtypes of GBM-DX were retrieved from https://xenabrowser.net/datapages/. The implementation of TEMI is available at https://github.com/wangyuanhao/TEMI.

Funding: This work was supported by the Science and Technology Planning Project of Guangzhou (Grant 2024A04J4225 to WW), the Natural Science Foundation of Guangdong Province(Grant 2025A1515011628 to YX), and the Fundamental Research Funds for the Central Universities (Grant 21623341 to WW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

(H&E)-stained histopathology slides can improve the efficiency of patient stratification, enable timely treatment, and reduce medical costs. A major challenge lies in uncovering how morphological features revealed in H&E-stained slides relate to variation in molecular signals, bridging the gap between phenotypes at different layers of the biological hierarchy. To address this issue, we proposed TEMI, a method for inferring molecular subtypes from H&E-stained histopathology slides, in which the discriminative representation learning of morphological characteristics is guided by transcriptomic profiles. The experiments showed that the proposed method achieved superior performance compared with existing tools, and the guidance of transcriptomic profiles helped to learn stable representations of morphological features in change of imaging techniques. Moreover, morphological features were shown to benefit gene expression prediction. These results suggest that the developed tool is effective for molecular subtype identification from histopathology slides and underscore the links between cancer morphology and transcriptomics.

Introduction

Histopathological images depict morphological and architectural details consisted of thousands of cells from tumor tissue sections in situ, and have been routinely used for cancer diagnostics. Computational pathology, which leverages computational methods for histopathological image analysis, has advanced rapidly with the rise of machine learning, particularly deep learning [1]. Deep learning methods have shown their strong capabilities of learning imperceptible patterns from whole-slide images (WSIs) of tumor sections for various downstream tasks, including classification of cancer types [2,3], prediction of gene mutation [4,5], and survival analysis [6,7].

Particularly, predicting molecular subtypes of cancers from histopathological images using deep learning has become a major area of interest, as it may accelerate diagnosis and enable more precise and timely clinical care. For example, Kather et al. showed that a ResNet18 model can detect microsatellite instability (MSI)—a phenotype arising from deficiencies in the DNA mismatch repair (MMR) system—at a satisfactory level in gastrointestinal cancers from WSIs [8]. Their study also estimated that deep-learning–based MSI screening could substantially reduce medical costs compared with the standard immunohistochemistry workflow. More recently, Saillard et al developed MSIntuit, a clinically approved deep-learning based pre-screening tool for MSI detection from haematoxylin-eosin (H&E)-stained slides [9]. In a related study, Chang et al. proposed a self-attention-based convolutional neural network for MSI classification in colorectal adenocarcinoma using a multicenter Chinese cohort [10]. In [11] and [12], researchers developed a clinically applicable ResNet18 model to distinguish four multi-omics integrative subtypes of hormone receptor-positive (HR^+)/human epidermal growth factor receptor 2-negative ($HER2^-$) breast cancer and demonstrated the superiority of subtyping-directed precision treatment strategies.

A significant challenge in applying deep learning methods to digital pathology lies in learning a compact representation of giga-pixel-sized histopathological images. Traditionally, regions of interest are selected by experienced pathologists, and then convolutional neural networks are trained by these well-curated data [7,13]. This approach is highly dependent on annotated labels and is labor-intensive. Semi-supervised learning methods aim to address this issue by leveraging limited local annotations more efficiently [14–16]. Meanwhile, multiple instance learning methods offer a more comprehensive solution [17] by capturing dependencies among local patches within a WSI through various attention mechanisms [18–21], requiring only global annotations. For instance, the pathology foundation model Prov-GigaPath [22] takes a WSI as input and employs the transformer-based LongNet [23] to model the long-range dependencies between local patches. Similarly, recently developed general-purpose foundation models [24–27] are mainly built upon a self-supervised framework DINOv2 [28] with Vision Transformer (ViT) backbones [29] for representation learning of patches and slides. However, these models demand highly expensive computational resources. To mitigate distributional shifts in datasets, self-supervised learning and meta-learning techniques have also been incorporated in learning highly generalized representations [30,31].

Delving into the tumor microenvironment (TME) is crucial for developing potential cancer therapies. Histopathological images that reveal the spatial tissue context of the TME in combination with spatial transcriptomic profiles is a promising forefront to achieve this goal [32,33]. However, the associated costs may limit accessibility for a broad patient population [34]. In this study, we take a step back and focus on integrating histopathological images with bulk transcriptomic data. It has been demonstrated that combining histopathological images with genomic or transcriptomic data significantly improves cancer diagnostics, prognosis, and the identification of biomarkers [6,7,35–37]. Additionally, research has shown that the morphological traits presented in histopathological images are closely associated with gene expression [38,39]. Inspired by these findings, we explore the potential of leveraging WSIs to differentiate molecular subtypes of cancers driven by gene expression changes. This approach could reveal the relationships between TME and molecular subtypes, paving the way for personalized targeted therapies at lower medical costs.

In contrast to previous studies that train models directly on WSIs and may suffer from limited generalization owing to the intricate morphology–molecular relationship [5,8,30], we propose TEMI (**T**ranscriptomic **E**xpression from **M**orphological **I**mages), a method designed to detect molecular subtypes of cancers using WSIs of tumor sections by jointly incorporating morphological features and transcriptomic profiles during training to improve its performance. There are two main challenges: the aforementioned representation learning of giga-pixel images, and the seamless integration of heterogeneous data with significant gaps. We tackle these issues by introducing a patch fusion network and aligning heterogeneous data under inexact conditions. The patch fusion network employs a multi-head dot-product attention mechanism to learn low-dimensional representations of WSIs. It automatically weighs patches within an images through their scaled similarities that describe pairwise dependencies, then linear combination of patches follows to yield a global representation. We also build a masked transcriptomic autoencoder to obtain compact representations of transcriptome in a low-dimensional space via self-supervised learning. To exploit the discriminative ability of transcriptome, we align paired data from two sources in a shared low-dimensional space by orthogonal decomposition of representations of transcriptomic data with the error term to capture the unnecessary parts to achieve inexact alignment, aiming to mitigate the disparities between sources. We also establish partial alignment between representations by using representations of WSIs for masked transcriptomic data reconstruction.

To evaluate the capability of TEMI and investigate the relationship between morphological features and transcriptomic profiles, we conducted a comprehensive analysis across three TCGA cancer cohorts: colorectal cancer (CRC; including colon and rectal adenocarcinomas), stomach adenocarcinoma (STAD), and glioblastoma multiforme (GBM), in which TEMI showed superior performance compared with existing approaches.

Results

Overview of TEMI

Our method comprises three key components: (1) a patch fusion network, (2) a masked transcriptomic autoencoder (MTA), and (3) heterogeneous data alignment. In our framework, WSIs are divided into non-overlapping patches, and deep features from these patches are extracted using a pretrained convolutional neural network. The patch fusion network integrates the deep features of patches from the same patient into a low-dimensional representation by a stacked multi-head dot-product attention, which automatically weighs the importance of deep features and patches. A multi-layer perceptron follows for decision. The architecture of TEMI is presented in Fig 1.

The MTA learns latent representations of bulk transcriptomic data through a masked autoencoder framework [40]. Specifically, it reconstructs expression signals of masked genes using information from unmasked genes aiming to capture the underlying relationships between genes in a low-dimensional space. Due to the randomness of masking, multiple masked inputs can be generated for each patient. To ensure compact and meaningful representations, contrastive learning [41,42] is employed so that data generated from the same patient are mapped closely together, while those from different patients are well-separated.

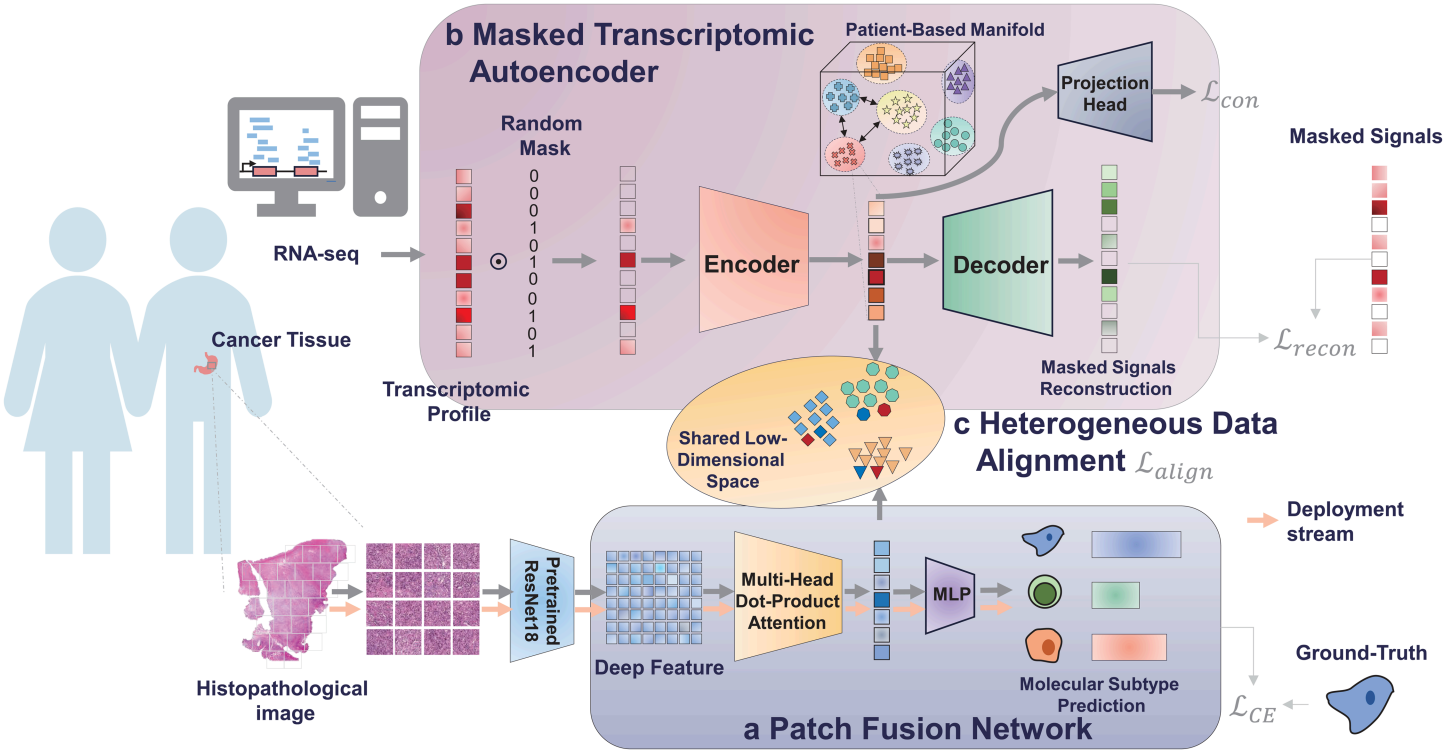

Fig 1. **Overview of TEMI. a** Patch fusion network generates representations of whole-slide images (WSIs) through a stacked multi-head dot-product attention. First, deep features of tiled patches are exacted by a pretrained convolutional neural network. Then the stacked multi-head dot-product attention integrates a collection of deep features into a low-dimensional representation by adaptively weighing deep features and patches. Predictions are made by a multi-layer perceptron (MLP). **b** Masked transcriptomic autoencoder reconstructs masked transcriptomic signals from the expression of unmasked genes, implicitly revealing the underlying relationships between genes. Using contrastive learning, the masked transcriptomic data are projected into a patient-based manifold, where compact, patient-oriented groups emerge, facilitating structured representations. **c** Heterogeneous data alignment maps the representations of WSIs and transcriptomic data into a shared low-dimensional space via orthogonal decomposition and partial reconstruction. This process injects discriminative information from transcriptomic signals into the patch fusion network while addressing the inherent modality gap between WSIs and transcriptomic data through inexact alignment. During deployment, only WSIs are required for molecular subtype prediction.

https://doi.org/10.1371/journal.pcbi.1013950.g001

To incorporate discriminative information from transcriptomic data, we propose heterogeneous data alignment. The representations of WSIs and transcriptomic data are projected into a shared low-dimensional space, ensuring that corresponding features from both modalities are closely aligned. However, exact alignment may not always be achievable due to the inherent gap between WSIs and molecular expression signals. To address this issue, we introduce orthogonal decomposition and partial reconstruction for inexact alignment, denoted by AOR (**A**lignment by **O**rthogonal **D**ecomposition) and APR (**A**lignment by **P**artial **R**econstruction), respectively, which allows us to separate modality-specific components while still aligning shared information across the two modalities. In deployment, only patch fusion network is employed from molecular subtype classification.

Molecular subtype classification

We verified our method using three cancer cohorts from TCGA—colorectal cancer, stomach adenocarcinoma, and glioblastoma multiforme—to identify molecular cancer subtypes based on formalin-fixed paraffin-embedded (FFPE) H&E-stained histopathology slides, and for each sample, bulk RNA-seq gene-expression data served as its transcriptomic profile. These cohorts are referred to as CRC-DX, STAD-DX, and GBM-DX, respectively. In both CRC-DX and STAD-DX, we aimed to distinguish microsatellite instable (MSI) from microsatellite stable (MSS) subtypes, whereas in GBM-DX, the goal was to classify tumors into the Proneural (Pron.) and Mesenchymal (Mese.) subtypes (see the Data and preprocessing section for details of datasets). We compared our method with two groups of approaches: (1) majority-voting based methods, where predictions are made using local patches of slides and the final decision is obtained through majority voting, and (2) one-slide-as-a-whole based methods, where low-dimensional representations of the entire slide are learned using a patch fusion method and used for prediction by a classifier. The first group includes ResNet18 [8] and MetaCon [30], while the second group includes ABMIL (Attention-based Multiple Instance Learning) [43] and 1Dconv (Convolutional One-Dimensional Layers) [44]. We also compared with TEMI trained without transcriptomic data indicated by TEMI w/o G. The models ABMIL and 1Dconv trained with transcriptomic data are denoted by ABMIL w/G and 1Dconv w/G, where we employed alignment by orthogonal decomposition (see the Methods section). We evaluated the models using the area under the receiver operating characteristic curve (AUC) calculated at the patient level. The median and 95% confidence intervals from 1000-fold bootstrapping for test samples are presented Table 1. Three of TEMI variants are indicated by TEMI w/AOD, TEMI w/APR, and TMEI w/AOD+APR (see the Methods section for details of these variants).

In CRC-DX, TMEI w/AOD+APR achieved the best median AUC (92.32%), while 1Dconv w/G outperformed the rest with the median AUC 84.24% in STAD-DX. In GBM-DX, TEMI w/AOD had the best median AUC 80.03%. Overall, variants of TEMI show comparative performance in the three cancer cohorts. For TEMI, ABMIL, and 1Dconv, incorporating transcriptomic profiles during training yields incremental gains for slide-based molecular subtype classification.

Compared to other methods without transcriptomic integration, TEMI w/o G consistently achieved superior performance across all three cancer cohorts, indicating the effectiveness of the proposed patch fusion network in capturing global representations of WSIs.

Transferable representation learning of different sources.

To assess whether transcriptomic data contributes to learning transferable features, we trained variants of TEMI using transcriptomic profiles and FFPE slides from the CRC-DX dataset and evaluated them on the CRC-KR dataset, which consists of snap-frozen colorectal cancer samples, for MSI vs. MSS classification.

Compared to TEMI trained without transcriptomic data, variants of TEMI showed improved performance when transcriptomic profiles were incorporated. We also compared our models with ResNet18 and MetaCon, which were directly trained and evaluated on the CRC-KR dataset. TMEI w/AOD+APR achieved the best AUC of 80.04% in the transfer learning setting, close to ResNet18 (82.15%) and MetaCon (84.09%). Table 2 summarizes these results, demonstrating that transcriptomic data enhances transfer learning in our models.

Table 1. Performance of compared methods in three cancer cohorts. We use AUC as evaluation metric. The median and 95% confidence intervals (in parentheses) from 1000-fold bootstrapping for test samples are reported.

Methods	CRC-DX MSI vs. MSS	STAD-DX MSI vs. MSS	GBM-DX Pron. vs. Mese.
ResNet18	0.7513	0.8107	0.7390
	(0.6169-0.8727)	(0.6898-0.9061)	(0.7242-0.7532)
MetaCon	0.9008	0.7291	–
	(0.8101-0.9599)	(0.5892-0.8446)	
1Dconv	0.8924	0.8322	0.7325
	(0.7980-0.9546)	(0.7326-0.9101)	(0.5827-0.8544)
ABMIL	0.8551	0.8141	0.7097
	(0.7470-0.9386)	(0.7101-0.8935)	(0.5657-0.8374)
TEMI w/o G	0.9043	0.8251	0.7755
	(0.8259-0.9608)	(0.7243-0.9076)	(0.6467-0.8913)
1Dconv w/G	0.8944	**0.8424**	0.7401
	(0.7971-0.9643)	(0.7491-0.9122)	(0.5938-0.8600)
ABMIL w/G	0.8862	0.8386	0.7333
	(0.7833-0.9582)	(0.7414-0.9156)	(0.5995-0.8581)
TEMI w/AOD	0.9148	0.8304	**0.8003**
	(0.8377-0.9661)	(0.7278-0.9089)	(0.6653-0.9174)
TEMI w/APR	0.9116	0.8181	0.7728
	(0.8412-0.9634)	(0.7202-0.8930)	(0.6353-0.8813)
TMEI w/AOD+APR	**0.9232**	0.8322	0.7893
	(0.8624-0.9720)	(0.7319-0.9022)	(0.6547-0.9019)

https://doi.org/10.1371/journal.pcbi.1013950.t001

Table 2. Evaluation on transfer learning. Variants of TEMI trained FFPE slides were evaluated by snap-frozen colorectal cancer samples. The median and 95% confidence intervals (CI) of AUC from 1000-fold bootstrapping for test samples are reported.

Learning paradigms	Methods		AUC	95% CI
Supervised learning	ResNet18		0.8215	0.7246-0.9109
	MetaCon		**0.8409**	0.7475-0.9193
Transfer learning	TEMI	w/o G	0.7638	0.6392-0.8642
		w/AOD	0.7690	0.6464-0.8685
		w/APR	0.7757	0.6549-0.8706
		w/AOD+APR	**0.8004**	0.6927-0.8945

https://doi.org/10.1371/journal.pcbi.1013950.t002

For each variant of TEMI, we extracted the representations of WSIs from the test samples of CRC-DX and CRC-KR. Principal component analysis was employed to project the representations into a three-dimensional space, thereby reducing the impact of high dimensionality on the distance metric. Fig 2 displays 200 randomly selected samples from each subtype. For each subtype, we then computed the pairwise cosine distance between the representations from the two sources. The average cosine distance, normalized by the average pairwise cosine distance of the same subtype within the CRC-KR dataset, is reported in Table 3. For all variants of TEMI, the normalized average pairwise inter-source distance was less than 1, indicating that representations of samples from CRC-DX and CRC-KR are more similar to each other than representations of samples within CRC-KR. This may explain why all TEMI variants achieved fair AUCs in the transfer learning setting. Leveraging transcriptomic profiles generally reduced the gap between representations of snap-frozen and FFPE samples in TEMI, except for TEMI w/APR on the MSI subtype. Overall, TEMI w/AOD+APR benefited the most, consistent with the results shown in Table 2.

The results suggest that integrating WSIs and transcriptomic profiles through co-training enhances the learning of invariant discriminative features for MSI vs. MSS classification in colorectal cancer, contingent on the alignment strategy used.

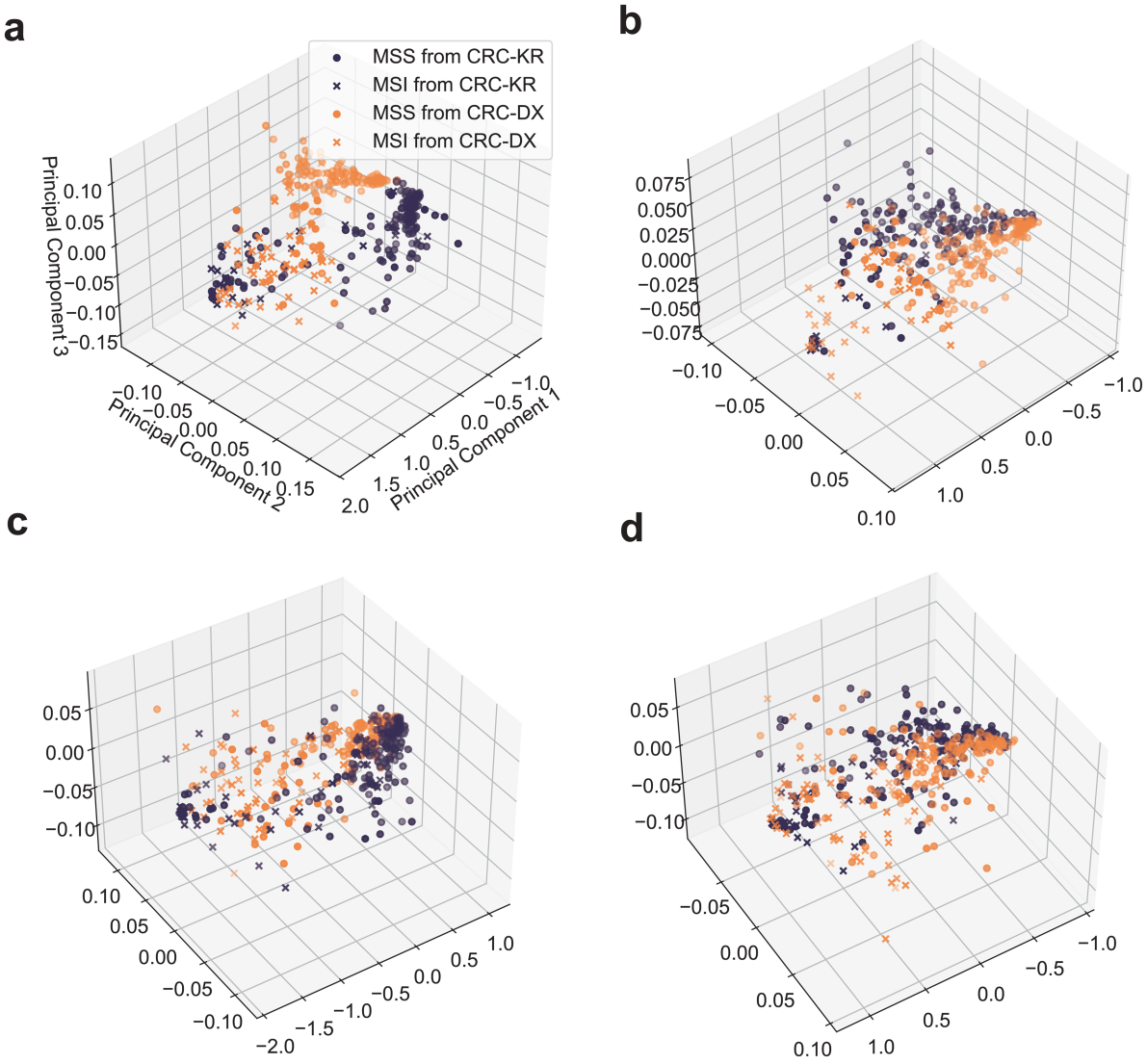

Fig 2. **Visualization of low-dimensional representations of snap-frozen and FFPE samples from colorectal cancer cohort. a** TEMI trained using whole-slide images only (TEMI w/o G). **b** TEMI incorporating alignment by partial reconstruction (TEMI w/APR). **c** TEMI incorporating alignment by orthogonal decomposition (TEMI w AOD). **d** TEMI combining both AOD and APR modules (TEMI w/AOD+APR).

https://doi.org/10.1371/journal.pcbi.1013950.g002

Table 3. **Representation gap between snap-frozen and FFPE samples.** The normalized average cosine distance between representations of snap-frozen and FFPE colorectal cancer samples was computed for each subtype. The ΔGap% denotes the relative change in percentage compared to the results obtained from TEMI w/o G, with arrows marking the change direction.

Variants of TEMI	MSI	MSS	Δ Gap%	
w/o G	0.8379	0.9733	–	–
w/AOD	0.8277	0.9308	1.22% ↓	4.37% ↓
w/APR	0.8491	0.9467	1.33% ↑	2.73% ↓
w/AOD+APR	0.8108	0.9414	3.23% ↓	3.27% ↓

https://doi.org/10.1371/journal.pcbi.1013950.t003

Contribution of morphological features to gene expression prediction.

To investigate the connections between morphological features and transcriptomic profiles, we evaluated MTA and TMEI w/AOD+APR for masked gene expression prediction and reported their mean squared errors (MSEs) on CRC and STAD

test samples. Distributions of errors are shown in Fig 3a–3d. A one-sided Mann-Whitney *U*-test was conducted to determine whether MSEs of TMEI w/AOD+APR were significantly lower than that of MTA. The results suggest morphological features benefit gene expression prediction for MTA, except in the MSI subtype of CRC samples. It provides evidence of connections between the morphological characteristics of tissues and their molecular profiles.

We performed gene ontology (GO) enrichment analysis using ClueGO (version 2.5.10) for the top 200 genes with the smallest MSEs in MSS and MSI subtypes of STAD samples, respectively. A Benjamini–Hochberg corrected *p*-value threshold of ≤ 0.05 was applied, GO terms were required to contain at least three genes(≥ 3). The enriched GO terms are presented in Fig 3e and 3f (also see Fig A in S1 Appendix). Most of the enriched GO terms are related to cell morphology. For the MSS subtype, the GO terms also involve immune regulation, specifically processes related to immune activation, antibody production, or immune tolerance.

The enrichment reflects genes whose expression levels were better predicted by TMEI w/AOD+APR, which are not necessarily those with higher expression levels and therefore does not conflict with the established knowledge that the MSI subtype is generally characterized by higher immune infiltration and immune activation in stomach adenocarcinoma [45,46]. Although MSI tumors typically exhibit stronger immune activation and infiltration, immune cells in MSS tumors are often sparse and localized, which enhances the morphological contrast between immune and non-immune regions and facilitates model recognition. Consequently, immune-related genes in the MSS subtype may exhibit lower prediction errors, reflecting the model's ability to capture morphologically distinct immune features rather than indicating stronger overall immune activity.

Ablation study on alignment strategies.

To validate the effectiveness of our heterogeneous data alignment methods—alignment by orthogonal decomposition (AOD), alignment by partial reconstruction (APR), and their combination (AOD + APR)—we compared various approaches for integrating transcriptomic data and WSIs. The methods compared include: (i) Mean Squared Error (MSE); (ii) Similarity Consistency (SimC); (iii) Hilbert-Schmidt Independence Criterion (HSIC); (iv) Maximum Mean Discrepancy (MMD); and (v) Gaussian Wasserstein Distance (GWD). These methods are commonly employed in multi-modal learning and transfer learning for data alignment (see the Methods section for details).

Results are presented in Table 4. In CRC-DX, MSE achieved the highest AUC, while in STAD-DX AOD + APR performed best, and in GBM-DX MMD yielded the highest score. Although no single alignment method can be regarded as universally dominant, the consistent top-3 performance of AOD + APR across all three datasets highlights the effectiveness of the proposed alignment strategies. Notably, AOD consistently outperformed APR, which indicates soft alignment may be preferred for seemingly disparate modalities.

Analysis of attention scores

Interpretability is of great concern for computational models applied to histopathological images, as it provides users with supporting evidence for the model's decisions. We demonstrated the interpretability of TEMI using patch-level attention scores on the CRC-DX cohort, with computational details provided in the Methods section. Two test samples were randomly selected: TCGA-AA-3837 from the MSS tumor and TCGA-WS-AB45 from the MSI tumor. Top 20 patches of the two samples with the highest attention scores are presented in Fig 4. Additional results, including the top 20 patches ranked by attention scores from CRC-DX, STAD-DX, and GBM-DX test samples, are shown in Figs B-D in S1 Appendix. The results demonstrate clear morphological distinctions between the molecular subtypes and consistent characteristics within each group. Heatmaps of patch-level attention scores from four GBM-DX test samples are also presented in Fig E in S1 Appendix.

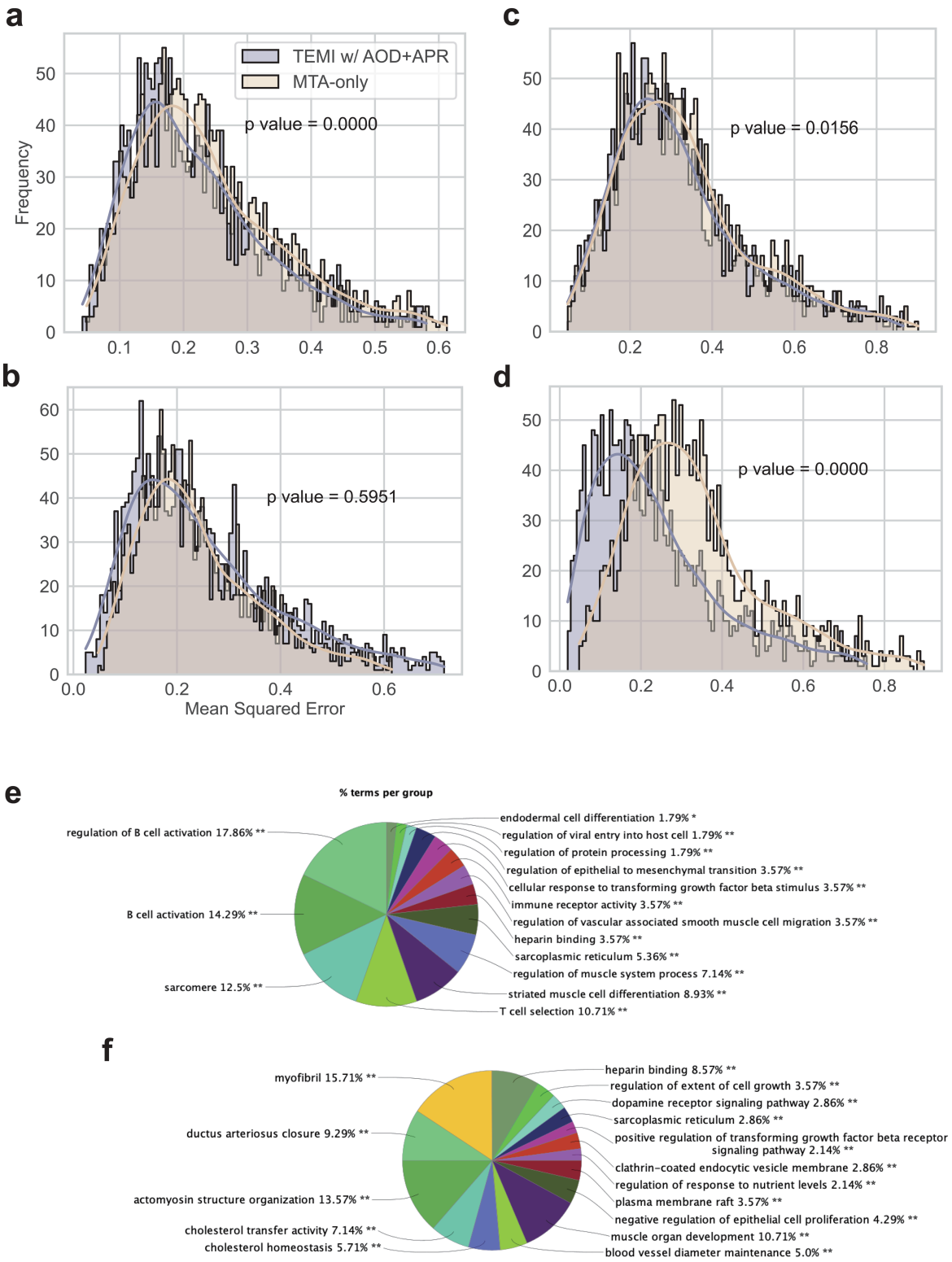

Fig 3. Distributions of mean squared errors (MSEs) of predicted masked genes and pathway enrichment analysis of the top 200 genes with the smallest errors. a, c MSS groups of colorectal cancer (CRC) and stomach adenocarcinoma (STAD) samples. **b, d** MSI groups of CRC and STAD samples. **e, f** Enriched pathways of the top 200 low-error genes in the MSS and MSI subtypes of STAD, respectively.

https://doi.org/10.1371/journal.pcbi.1013950.g003

Table 4. Comparison of methods for data alignment. We use AUC as evaluation metric. The median and 95% confidence intervals (in parentheses) from 1000-fold bootstrapping for test samples are reported.

Alignment	CRC-DX	STAD-DX	GBM-DX
MSE	**0.9260**	0.8169	0.7844
	(0.856-0.9739)	(0.7193-0.9044)	(0.6350-0.8954)
SimC	0.9182	0.8067	0.7685
	(0.8564-0.9661)	(0.7077-0.8947)	(0.6310-0.8950)
HSIC	0.9168	0.8069	0.7825
	(0.8481-0.9668)	(0.6947-0.8954)	(0.6454-0.8880)
MMD	0.9024	0.8151	**0.8120**
	(0.8233-0.9574)	(0.7019-0.9000)	(0.6787-0.9204)
GWD	0.9174	0.8155	0.7787
	(0.8456-0.9665)	(0.7138-0.9027)	(0.6303-0.8990)
AOD	0.9148	0.8304	0.8003
	(0.8377-0.9661)	(0.7278-0.9089)	(0.6653-0.9174)
APR	0.9116	0.8181	0.7728
	(0.8412-0.9634)	(0.7202-0.8930)	(0.6353-0.8813)
AOD+APR	0.9232	**0.8322**	0.7893
	(0.8624-0.9720)	(0.7319-0.9022)	(0.6547-0.9019)

https://doi.org/10.1371/journal.pcbi.1013950.t004

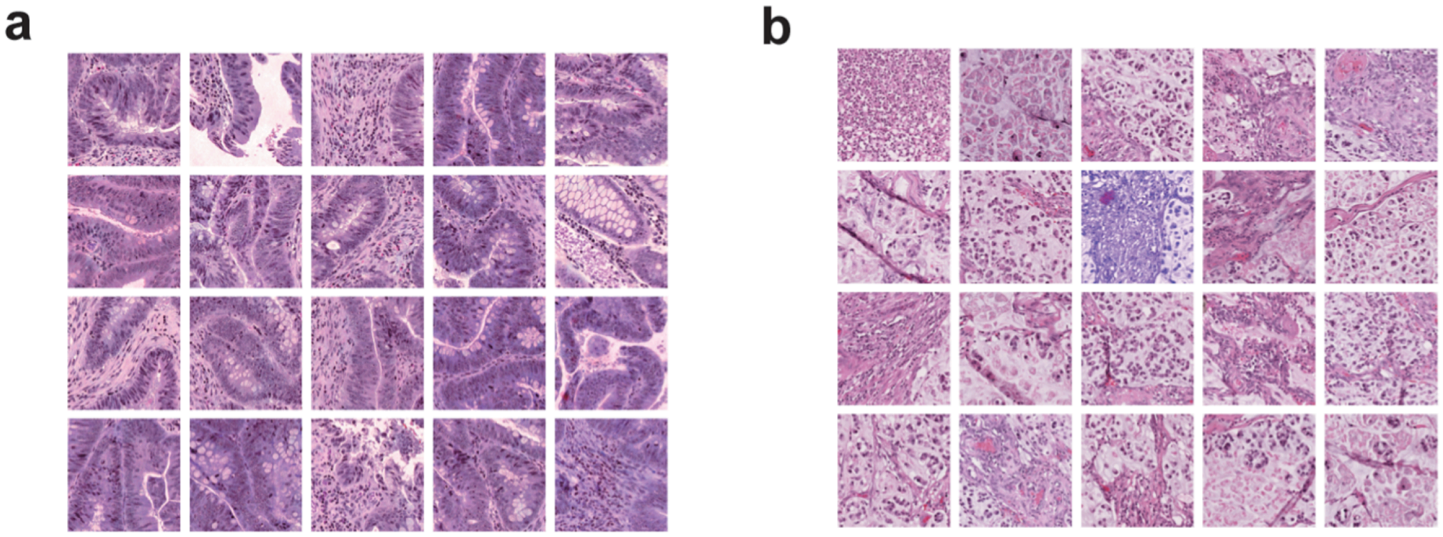

Fig 4. Top 20 patches ranked by attention scores of two samples from CRC-DX. a TCGA-AA-3837 from the MSS group. **b** TCGA-WS-AB45 from the MSI group.

https://doi.org/10.1371/journal.pcbi.1013950.g004

To further explore the effectiveness of the selected patches, we treat patches demonstrated in Fig 4 as templates and analyze their associations with MSS and MSI samples. We define similarity between patches $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ by the following formula, where patches are represented by their deep features.

$$\mathtt{sim}(\boldsymbol{r}_i, \boldsymbol{r}_j) = \exp\left(\frac{\boldsymbol{r}_i^T \boldsymbol{r}_j}{\|\boldsymbol{r}_i\|_2 \|\boldsymbol{r}_j\|_2} \Big/ \tau\right).$$

In the experiment, the parameter τ was set as 0.07.

For each patch in a WSI, its subtype-specific similarity score was computed by averaging its similarity scores between the 20 templates of the respective subtype. We visualized patches of a WSI by UMAP [47] using their deep features, with

brightness of each point indicating its subtype-specific similarity. Two test samples, the MSS sample TCGA-CM-5864 and the MSI sample TCGA-AZ-6598, were selected for demonstration. The results are shown in Fig 5. Patches from TCGA-CM-5864 exhibit higher similarity scores with the MSS templates than with the MSI templates, whereas the opposite is observed for TCGA-AZ-6598.

These findings suggest that attention scores may help reveal subtype-relevant morphological characteristics.

Impact of feature extractors

In our experiments, we employed ResNet18 as the feature extractor for image patches. However, with advanced architectures for visual representation available, such as the ViT, it is natural to question whether ResNet18 is sufficiently effective for TEMI. To address this, we further incorporated DINOv2, a prevailing backbone in foundation models for

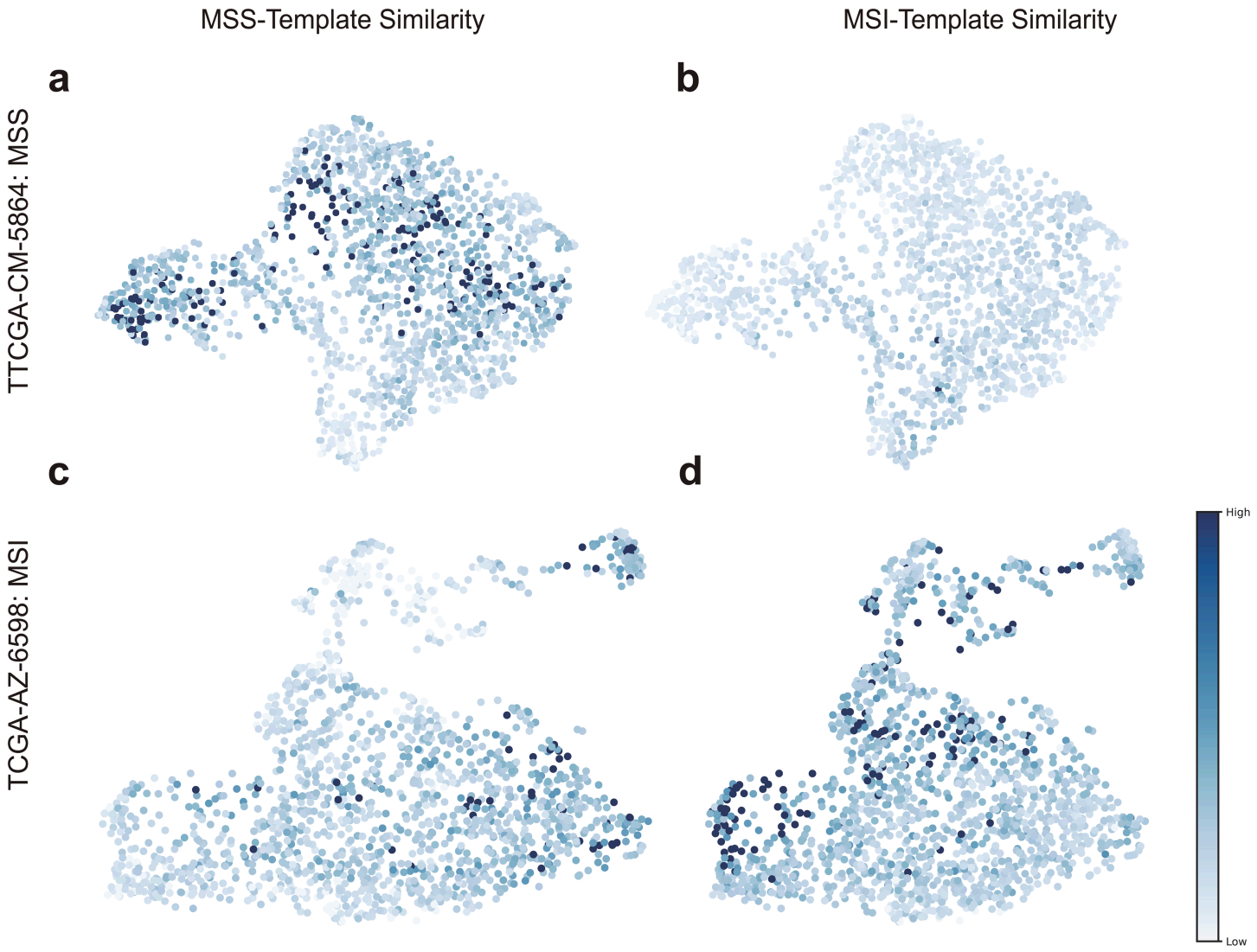

Fig 5. **Subtype-specific similarity scores of patches.** **a**, **b** The MSS TCGA-CM-5864. **c**, **d** The MSI TCGA-AZ-6598. **a**, **c** Similarity with respect to the MSS templates. **b**, **d** Similarity with respect to the MSI templates.

https://doi.org/10.1371/journal.pcbi.1013950.g005

histopathological image analysis, as a benchmark of comparison. We utilized a pretrained DINOv2 model on ImageNet and subsequently fine-tuned it for tumor detection, where the task was to classify patches as originating from tumor, dense tissue, or loose tissue. The fine-tuning dataset, provided by [8], comprised 11,977 patches, and it was split into 70% for training and 30% for validation. During the fine-tuning stage, only the last ten layers of DINOv2, as well as an additional linear layer for classification, were updated by AdamW with a learning rate of 1×10^{-4}. After 20 epochs, the accuracy reached 99.82% on the training set and 99.64% on the validation set. The fine-tuned model was then used as a feature extractor for TEMI.

As foundation models for computational pathology have emerged in recent years, we also included several of these models as feature extractors for comparison. Specifically, we evaluated: (1) Prov-GigaPath, trained on approximately 170,000 WSIs from the Providence Health Network [22]; (2) H-optimus-1, trained on over 1 million proprietary WSIs [26]; and (3) Virchow, trained on 1.5 million slides from Memorial Sloan Kettering Cancer Center [27]. Additional details of these foundation models are provided in Table A in S1 Appendix.

Fig 6 presents the performance of TEMI variants on the CRC-DX cohort using ResNet18 and the four advanced models as feature extractors. The best median AUC (93.13%) was achieved by TEMI w/AOD+APR using Prov-GigaPath as the feature extractor, followed closely by the variant using ResNet18 (median AUC: 92.32%). The median AUCs achieved by using different feature extractors are summarized in Table B in S1 Appendix. Interestingly, using advanced foundation-model extractors did not yield substantial performance improvements for TEMI. This contrasts with ABMIL, which benefitted considerably from these extractors, with its median AUC increasing from 85.51% to over 90%. For TEMI, the performance differences among simple extractor (ResNet18) and advanced extractors (e.g., Prov-GigaPath) were marginal. This suggests that the architecture of TEMI may already be sufficiently expressive for this task. Meanwhile, these results do not diminish the strong representation-learning capabilities of pathology foundation models, as they require far fewer training epochs to achieve satisfactory performance before overfitting in the downstream classification.

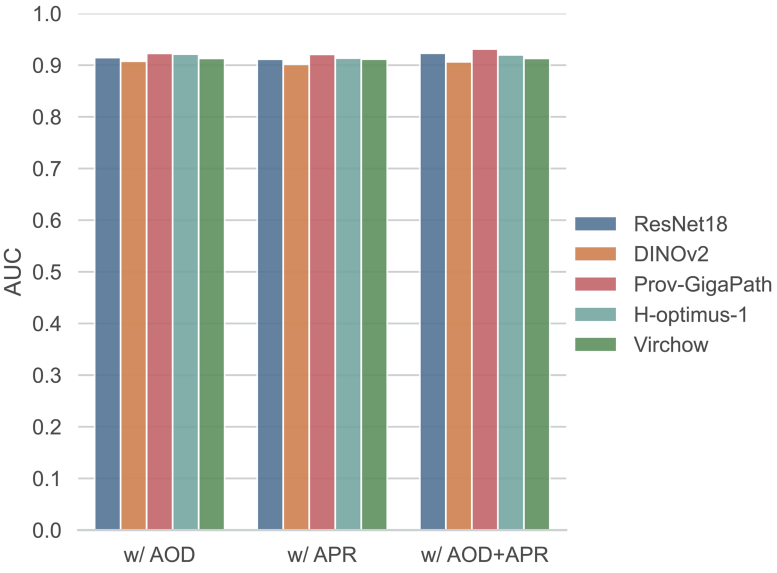

Fig 6. Performance of variants of TEMI on CRC-DX using different feature extractors. Employing the advanced feature extractor does not lead to significant improvement in TEMI.

https://doi.org/10.1371/journal.pcbi.1013950.g006

Discussion

We developed TEMI, a framework for identifying molecular subtypes of cancer from whole-slide images (WSIs), enhanced with transcriptomic data during training. TEMI makes three main contributions. First, it adopts a multimodal-training-with-unimodal-inference paradigm, enabling molecular subtype classification from WSIs while leveraging transcriptomic profiles only at the training stage. Second, to address the challenge of representation learning on giga-pixel WSIs, TEMI develops a patch fusion network that leverages a multi-head dot-product attention mechanism to adaptively weigh features and patches, thereby generating unified representations through linear combinations. For the transcriptomic modality, TEMI employs a masked transcriptomic autoencoder to derive compact representations, where masked reconstruction effectively captures underlying gene relationships. Finally, TEMI presents two alignment strategies that bridge morphological features and molecular signals in a shared low-dimensional space, guiding WSI representations through alignment with discriminative transcriptomic embeddings.

Experiments demonstrate that TEMI achieves superior performance in molecular subtype classification compared with existing methods. In a transfer learning setting, TEMI benefits from the incorporation of transcriptomic data, which guides the learning of invariant representations through heterogeneous data alignment. The proposed alignment strategies—alignment by orthogonal decomposition and alignment by partial reconstruction—prove effective against established alignment approaches. Analysis of attention scores shows that TEMI highlights discriminative local regions associated with different molecular subtypes, supporting the design of the multi-head dot-product attention mechanism. Further, our experiments reveal that morphological features enhance gene expression prediction for the masked transcriptomic autoencoder, confirming associations between WSIs and transcriptomic profiles. TEMI relies on patch-level deep features, and notably, even a relatively simple backbone (ResNet18) is sufficient to yield strong performance. Overall, TEMI provides an effective and efficient framework for molecular subtype classification from WSIs.

This study has several limitations that should be addressed. First, we integrated WSIs only with bulk RNA-seq gene-expression data, whereas additional molecular modalities—such as protein expression and DNA methylation—should be considered in future work. Even advanced foundation-model feature extractors with over a billion parameters trained on large-scale WSIs—such as Prov-GigaPath and H-optimus-1—did not provide substantial improvements over the much simpler feature extractor ResNet18 for TEMI. This suggests that, while foundation models may be necessary for molecular subtype classification due to their ability to achieve satisfactory performance in fewer training epochs before overfitting (see the section *Impact of feature extractors*), they are not sufficient on their own, highlighting that opportunities remain for designing efficient and task-specific architectures in computational pathology.

Although alignment methods are well developed for modalities with clear logical correspondence (e.g., images, audio, and text), aligning WSIs with transcriptomic profiles remains challenging due to the significant modality gap and potentially undiscovered associations. Our results further indicate that discoveries may be highly dependent on the alignment strategies employed; since no single method consistently dominates, future efforts should focus on developing alignment approaches tailored to these modalities.

Another important avenue for future research lies in elucidating the relationship between morphological context and underlying molecular functions or biological processes. Notably, the benefit of morphological features for gene expression prediction was absent in the GBM-DX cohort (see Fig F in S1 Appendix), raising the question of whether such associations are universal across cancers. For the Proneural vs. Mesenchymal classification in the GBM-DX cohort, all methods achieved only fair performance compared with the CRC-DX and STAD-DX cohorts, with the best AUC being 80.03% from TEMI w/AOD. These results suggest that inferring gene expression from H&E morphology is more difficult in GBM than in CRC and STAD. Therefore, the lack of improvement from morphological features is likely due to weaker associations between GBM morphology and its transcriptomic profiles. Moreover, such conclusion may also be affected by model bias as only fair AUC was achieved, and therefore it may require more sophisticated models for representation learning before morphology can meaningfully enhance gene expression prediction.

Methods

There are two main obstacles in the formulation of our methods: (1) representation learning of giga-pixel-level images, and (2) data alignment of heterogenous data. This section describes how we resolve these challenges.

Patch fusion network (PFN)

A whole-slide image is tiled into non-overlapped patches. Mathematically, image $\mathbb{I}_i$ can be represented as

$$\mathbb{I}_i = \bigcup_{k=1}^{N_i} R_k^{(i)}, \tag{1}$$

and for any $k, k' \in \{1, 2, \ldots, N_i\}$, we have $R_k^{(i)} \bigcap R_{k'}^{(i)} = \varnothing$ when $k \neq k'$. Here $R_k^{(i)}$ is a fixed-size region of $\mathbb{I}_i$ for all k.

To simplify training process, we extract the deep feature of $R_k^{(i)}$ denoted by $\boldsymbol{r}_k^{(i)} (\in \mathbb{R}^m)$ using a pretrained deep neural network. Let $\mathcal{S}$ represent the space consisted of any finite set of elements in $\mathbb{R}^m$. The goal of PNF is to learn a mapping $\mathcal{A}_{\boldsymbol{\Psi}}$ that transforms elements from $\mathcal{S}$ to a low-dimensional space $\mathcal{Z} \subset \mathbb{R}^d$,

$$\mathcal{A}_{\boldsymbol{\Psi}} : \{\boldsymbol{r}_k^{(i)}\}_{k=1}^{N_i} \mapsto \boldsymbol{z}_i, \ \mathcal{S} \to \mathcal{Z}, \tag{2}$$

where $\mathcal{A}_{\boldsymbol{\Psi}}$ is characterized by a learnable-parameter set $\boldsymbol{\Psi}$.

Linear transformation. We first transform an input $\boldsymbol{H}_{\text{in}} \in \mathbb{R}^{c_{\text{in}} \times p}$ by a linear mapping $\boldsymbol{W}^{p \times c_{\text{out}}}$ for dimensionality reduction, i.e.,

$$\boldsymbol{H}_{\text{out}} = \boldsymbol{H}_{\text{in}} \cdot \boldsymbol{W}, \tag{3}$$

and hence $\boldsymbol{H}_{\text{out}} \in \mathbb{R}^{c_{\text{in}} \times c_{\text{out}}}$. Note that $\boldsymbol{W}$ can be regarded as an ensemble of one-dimensional convolution.

Multi-head dot-product attention. We then propose a multi-head attention to aggregate columns of $\boldsymbol{H}_{\text{out}}$. Consider M heads attention, we characterize the similarity relationship between columns of $\boldsymbol{H}_{\text{out}}$ by a scaled inner product

$$\alpha_i^m = \sum_{j=1, j \neq i}^{c_{\text{out}}} \frac{\langle \boldsymbol{h}_{\text{out},i} \odot \theta^m, \boldsymbol{h}_{\text{out},j} \odot \theta^m \rangle}{c_{\text{out}}}, \tag{4}$$

where $\boldsymbol{h}_{\text{out},\cdot}$ stands for the columns of $\boldsymbol{H}_{\text{out}}$ and $\theta^m \in \mathbb{R}^{c_{\text{in}}}$. The softmax operator follows for normalization

$$\hat{\alpha}_i^m = \exp(\alpha_i^m) \Big/ \sum_{j=1}^{c_{\text{out}}} \exp(\alpha_j^m). \tag{5}$$

We attain linear combination of element-wise scaled columns as

$$\boldsymbol{h}_a^m = \sum_{i=1}^{c_{\text{out}}} \hat{\alpha}_i^m (\boldsymbol{h}_{\text{out},i} \odot \theta). \tag{6}$$

where $\theta \in \mathbb{R}^{c_{\text{in}}}$. By varying m from 1 to M, we have a multi-head attention representation $\boldsymbol{H}_a = [\boldsymbol{h}_a^1, \boldsymbol{h}_a^2, \ldots, \boldsymbol{h}_a^M] \in \mathbb{R}^{c_{\text{in}} \times M}$.

Building blocks of PFN. Let $C_{\boldsymbol{W}}$ and $A_{\boldsymbol{\Theta}}$ denote the linear transformation and the multi-head dot-product attention, respectively. The building block of PFN can be represented as a composition operator $FB = A_{\boldsymbol{\Theta}} \circ \text{Tanh} \circ C_{\boldsymbol{W}}$, where W and $\boldsymbol{\Theta}$ are learnable parameters in the linear transformation and the multi-head dot-product attention, respectively.

By stacking multiple building blocks, we express PFN $\mathcal{A}_{\Psi}$ as

$$\mathcal{A}_{\Psi} = FB^{L} \circ (FB^{L-1})^{T} \circ \cdots \circ FB^{1}. \tag{7}$$

where $(\cdot)^T$ represents transposition. For the final block, we flatten its output in row-major order, and achieve a low-dimensional representation, denoted by $\boldsymbol{h}$.

To accommodate with various number of patches of whole-slide images, we perform uniformly sampling with replacement on deep features of patches with a fixed sample size for each image and use the random samples as inputs of PFN. Hence the input of PFN is represented as $\boldsymbol{H}^i = [\boldsymbol{c}_1^i, \boldsymbol{c}_2^i, \ldots, \boldsymbol{c}_N^i]^T \in \mathbb{R}^{N\times m}$ for the image $\mathbb{I}_i$, where $\boldsymbol{c}_{\cdot}^i$ represents a random sample from $\{\boldsymbol{r}_k^{(i)}\}_{k=1}^{N_i}$ and N is the sample size.

Attention score. We define attention score of $\boldsymbol{c}_{\cdot}^i$ by the output of the first building block.

Let

$$\widetilde{\boldsymbol{H}}^i = FB^1(\boldsymbol{H}^i)^T \in \mathbb{R}^{M_1\times N},$$

where M_1 denotes the number of heads of attention using in the FB^1. The attention score of $\boldsymbol{c}_k^i$ is defined as

$$\alpha_k^i = \frac{\exp\left(\sum_{r=1}^{M_1} |\widetilde{\boldsymbol{H}}_{r,k}^i|\right)}{\sum_{j=1}^{N} \exp\left(\sum_{r=1}^{M_1} |\widetilde{\boldsymbol{H}}_{r,j}^i|\right)}, \quad k \in \{1, 2, \ldots, N\}. \tag{8}$$

Memory cost of dot-product attention. Given the input $\boldsymbol{H} \in \mathbb{R}^{K\times D}$, the memory cost of single-head dot-product attention is $O(K^2 + KD)$, where the K^2 term arises from computing pairwise inner products between scaled embeddings, and the KD term accounts for storage of the scaled embeddings. If $K<D$, the quadratic term K^2 can be omitted, reducing the memory cost to $O(KD)$. Under this condition, dot-product attention can be more memory-efficient than ABMIL and comparable to LongNet. A summary of the comparison with ABMIL and LongNet is provided in Table C in S1 Appendix.

Masked transcriptomic autoencoder (MTA)

To learn high quality compact representations of transcriptomic data, we propose a modified masked autoencoder, which is implemented by multi-layer perceptrons (MLP). Let $\boldsymbol{g}_i \in \mathbb{R}^p$ denote the transcriptomic data of patient i corresponding to the whole-slide image $\mathbb{I}_i$. We randomly mask elements of $\boldsymbol{g}_i$ and reconstruct the unmasked elements by a MLP-based autoencoder.

Let $\boldsymbol{m}_i \in \mathbb{R}^p$ be a random vector, where each of its elements is generated by a Bernoulli distribution Bernoulli(ξ). The masked $\boldsymbol{g}_i$ is written as

$$\hat{\boldsymbol{g}}_i = \boldsymbol{g}_i \odot (1 - \boldsymbol{m}_i), \tag{9}$$

The transcriptomic data is reconstructed by a MLP,

$$\tilde{\boldsymbol{g}}_i = \mathrm{MLP}(\hat{\boldsymbol{g}}_i), \tag{10}$$

where the l-th layer of MLP is implemented as

$$\boldsymbol{z}_i^{(l)} = \mathrm{Tanh}(\boldsymbol{W}^{(l)}\boldsymbol{z}_i^{(l-1)} + \boldsymbol{b}^{(l)}). \tag{11}$$

Tanh$(\cdot)$ is employed for its symmetric and bounded characteristics, which avoid activation explosion and promote stable optimization. For positively skewed inputs, the preceding linear projections can adaptively rescale their magnitudes during training, thereby reducing the risk of saturation.

In Eq (11), $\boldsymbol{W}^{(l)} \in \mathbb{R}^{p_l \times p_{l-1}}$ and $\boldsymbol{b}^{(l)} \in \mathbb{R}^{p_l}$ are learnable parameters and we denote the dimension of the l-th layer by p_l. Note that $\boldsymbol{z}_i^{(0)} = \hat{\boldsymbol{g}}_i$.

The reconstruction loss of masked transcriptomic data is defined by the squared loss

$$\ell_{recon} = \|\tilde{\boldsymbol{g}} \odot \boldsymbol{m} - \boldsymbol{g} \odot \boldsymbol{m}\|_2^2, \tag{12}$$

where we have omitted the subscript for brevity. The reconstruction loss varies depending on the random mask. Hence each randomly masked version of a patient's transcriptomic profile can be regarded as an independent sample for data reconstruction.

Given a mini-batch of $\mathcal{B} = \{(\boldsymbol{g}_i, \boldsymbol{m}_i)\}_{i=1}^B$ with size B, the reconstruction loss is defined by the mean squared error (MSE)

$$\mathcal{L}_{recon} = \frac{1}{B}\sum_{i=1}^{B} \|\tilde{\boldsymbol{g}}_i \odot \boldsymbol{m}_i - \boldsymbol{g}_i \odot \boldsymbol{m}_i\|_2^2. \tag{13}$$

MTA learns implicit connections between genes via masked reconstruction. With the random vector $\boldsymbol{m}_i$, we are able to generate different versions of masked transcriptomic expression signals for patient i. Hence, we also adopt InfoNCE loss [41,42] in MTA. For simplicity, let $\boldsymbol{z}$ be the output of the middle layer for MLP as the latent representation of masked transcriptomic data $\hat{\boldsymbol{g}}$. We define a projection head as

$$\tilde{\boldsymbol{z}} = \boldsymbol{W}_p^{(2)}\text{Tanh}(\boldsymbol{W}_p^{(1)}\boldsymbol{z} + \boldsymbol{b}_p^{(1)}) + \boldsymbol{b}_p^{(2)} \tag{14}$$

Shapes of the linear transformation matrices $\boldsymbol{W}_p^{(i)}$ and biases $\boldsymbol{b}_p^{(i)}$ $(i = 1, 2)$ are predefined.

Given a mini-batch of projected latent representations $\mathcal{B} = \{\tilde{\boldsymbol{z}}_k\}_{k=1}^B$with size as B, let $\mathcal{B}_{-k} = \mathcal{B}\backslash\{\tilde{\boldsymbol{z}}_k\}$ and $\mathcal{P}_k$ be the collections of projected latent representations generated from the same patient as $\tilde{\boldsymbol{z}}_k$. Let $\hat{\boldsymbol{z}} = \tilde{\boldsymbol{z}}/\|\tilde{\boldsymbol{z}}\|_2$, where subscripts are omitted for brevity. The InfoNCE loss is defined as

$$\mathcal{L}_{con} = \sum_{\tilde{\boldsymbol{z}}_k \in \mathcal{B}} \frac{-1}{|\mathcal{P}_k|} \sum_{\tilde{\boldsymbol{z}}_p \in \mathcal{P}_k} \log \frac{\exp(\hat{\boldsymbol{z}}_k^T \hat{\boldsymbol{z}}_p/\tau)}{\sum_{\tilde{\boldsymbol{z}}_a \in \mathcal{B}_{-k}} \exp(\hat{\boldsymbol{z}}_k^T \hat{\boldsymbol{z}}_a/\tau)}, \tag{15}$$

where τ is the temperature.

Heterogenous data alignment

Alignment by orthogonal decomposition. Given the low-dimensional representation $\boldsymbol{h}$ of a whole-slide image and the corresponding latent representation of transcriptomic data $\boldsymbol{z}$, both of which are originated from the same patient, we assume the orthogonal decomposition

$$\boldsymbol{z} = \boldsymbol{h} + \boldsymbol{e} \quad \text{and} \quad \boldsymbol{h}^T\boldsymbol{e} = 0, \tag{16}$$

where $\boldsymbol{e}$ denotes the irrelevant part to $\boldsymbol{h}$.

A straightforward way to align $\boldsymbol{h}$ and $\boldsymbol{z}$ is to minimize $\|\boldsymbol{h}-\boldsymbol{z}\|_2^2$. But it may lead to a trivial solution $\boldsymbol{e}=\boldsymbol{0}$. Instead, we relax the orthogonal decomposition as the loss function

$$\ell_{algin} = |\boldsymbol{h}^T(\boldsymbol{h}-\boldsymbol{z})| + \frac{\delta}{2}(\|\boldsymbol{z}-\boldsymbol{h}\|_2^2 - C)^2. \tag{17}$$

where δ and C are hyper-parameters. Empirically, δ is set to balance the scales of each term and C is set to 1.

By minimizing ℓ_{algin}, the first term decomposes the irrelevant part to $\boldsymbol{h}$ from $\boldsymbol{z}$ and the second term achieves alignment between the whole-slide image and the transcriptomic data while avoiding a trivial solution that leads to exact alignment.

Alignment by partial reconstruction. We can also establish implicit alignment between $\boldsymbol{h}$ and $\boldsymbol{z}$ by using latent representations of images for transcriptomic reconstruction. Specifically, we transform the output of the middle layer of MTA by a bilinear operator as

$$\dot{\boldsymbol{z}} = (\boldsymbol{W}_G\boldsymbol{z} + \boldsymbol{b}_G) \odot (\boldsymbol{W}_I\boldsymbol{h} + \boldsymbol{b}_I), \tag{18}$$

where $\odot$ indicates element-wise product. Shapes of $\boldsymbol{W}_\cdot$ and $\boldsymbol{b}_\cdot$ in Eq (18) depend on the dimensions of $\boldsymbol{z}$ and $\boldsymbol{h}$. Then $\dot{\boldsymbol{z}}$ passes the next layer in MTA for reconstruction of masked transcriptomic data.

Data augmentation

In our study, each patient is provide with a whole-slide image and transcriptomic data at the training stage. Denote $\left\{\left(\{\boldsymbol{r}_k^{(i)}\}_{k=1}^{N_i}, \boldsymbol{g}_i, y_i\right)\right\}_{i=1}^{M}$ as the training set, where y_i indicates the molecular subtype of patient i. We sample with replacement from $\{\boldsymbol{r}_k^{(i)}\}_{k=1}^{N_i}$ with a fixed sample size N and get $\boldsymbol{H}^j = [\boldsymbol{c}_1^j, \boldsymbol{c}_2^j, \ldots, \boldsymbol{c}_N^j]^T \in \mathbb{R}^{N\times m}$ where $\boldsymbol{c}_\cdot^j$ represents the random samples. Simultaneously, we generate a random mask $\boldsymbol{m}_i$ for the transcriptomic data $\boldsymbol{g}_i$. We repeat the procedure multiple times for one sample in the training set. Eventually, we obtain an extended training set denoted by $\left\{(\boldsymbol{H}^j, \boldsymbol{g}_i, \boldsymbol{m}_i, y_i)\right\}_{i=1}^{M'}$. We have abused using the index i here, but the meaning of i should be clear from the context. Note that in the tuple $(\boldsymbol{H}^j, \boldsymbol{g}_i, \boldsymbol{m}_i, y_i)$, the transcriptomic data $\boldsymbol{g}_i$ and the molecular subtype y_i are copied from the patient that generates $\boldsymbol{H}^j$ and $\boldsymbol{m}_i$.

Total loss

Given a mini-batch of augmented data $\mathcal{B} = \left\{(\boldsymbol{H}^j, \boldsymbol{g}_i, \boldsymbol{m}_i, y_i)\right\}_{i=1}^{B}$. The latent representation $\boldsymbol{h}_i = \mathcal{A}_\Psi(\boldsymbol{H}^j)$ and the ground-truth subtype $y_i \in \{1, 2, \ldots, K\}$ are used to construct the cross-entropy loss for training. A linear classifier generates the predicted subtype $\hat{\boldsymbol{y}}_i \in \mathbb{R}^K$, i.e.,

$$\hat{\boldsymbol{y}}_i = \text{softmax}(\boldsymbol{W}_c\boldsymbol{h}_i + \boldsymbol{b}_c). \tag{19}$$

The symbol K indicates the number of subtypes. The cross-entropy loss of the batch $\mathcal{B}$ is

$$\mathcal{L}_{CE} = \frac{1}{B}\sum_{i=1}^{B} -\log \hat{\boldsymbol{y}}_{i,y_i}. \tag{20}$$

By Eq (17), the alignment loss of whole-slide images and transcriptomic data is

$$\mathcal{L}_{align} = \frac{1}{B}\sum_{i=1}^{B}\left(|\boldsymbol{h}_i^T(\boldsymbol{h}_i-\boldsymbol{z}_i)| + \frac{\delta}{2}(\|\boldsymbol{z}_i-\boldsymbol{h}_i\|_2^2 - C)^2\right). \tag{21}$$

With different combinations of loss functions, we get three variants of our method.

- Combing Eqs (13), (15), (20), and (21) yields the total loss

$$\mathcal{L}_{AOD} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{align} + \mathcal{L}_{con} + \mathcal{L}_{recon}. \quad \text{(TEMI w/AOD)}$$

- Combing Eqs (13), (15), (18), and (20) yields the total loss

$$\mathcal{L}_{APR} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{con} + \mathcal{L}_{recon}. \quad \text{(TEMI w/APR)}$$

- Combing Eqs (13), (15), (18), (20), and (21) yields the total loss

$$\mathcal{L}_{AOD+APR} = \mathcal{L}_{CE} + \lambda \left(\mathcal{L}_{align} + \mathcal{L}_{con} \right) + \mathcal{L}_{recon}. \quad \text{(TEMI w/AOD+APR)}$$

The hyperparameter λ is set for balancing the scales of individual terms. All learnable parameters are learned by minimizing the total loss.

Datasets and preprocessing

Three datasets were used to classify microsatellite instable (MSI) and microsatellite stable (MSS) subtypes, with MSI arising from deficiencies in the DNA mismatch repair (MMR) system:

- **CRC-DX:** 358 colorectal cancer (CRC) patients with diagnostic slides (indicated by the *DX* suffix) and bulk transcriptomic profiles retrieved from TCGA colon and rectal adenocarcinomas.
- **CRC-KR:** The same patients as CRC-DX, with *cr*yosections generated from snap-frozen tissues, denoted by the suffix *KR.*
- **STAD-DX:** 284 stomach adenocarcinoma (STAD) patients with diagnostic slides and bulk transcriptomic profiles obtained from TCGA.

Additionally, one dataset was used to classify Proneural (Pron.) and Mesenchymal (Mese.) subtypes of glioblastoma multiforme (GBM), two highly distinct transcriptomic subtypes characterized by differential gene expression [48]:

- **GBM-DX:** 182 GBM patients with diagnostic images and bulk transcriptomic profiles retrieved from TCGA.

All datasets used HTSeq-FPKM records as bulk transcriptmoic profiles. For while-slide images, CRC-DX, STAD-DX, and GBM-DX utilized diagnostic slides from formalin-fixed paraffin-embedded tissues, while CRC-KR used cryosections from snap-frozen tissues.

The three datasets for MSI vs. MSS classification were sourced from [8]. The whole-slide images had been tiled into non-overlapped patches with size 224 pixel $\times$ 224 pixel at a resolution of 0.5μm/pixel and had been conducted color normalization with the Macenko method. The tumor patches were identified by a well-trained ResNet18. We extracted 512-dimensional deep features of tumor patches from the last pooling layer of this pre-trained ResNet18 as inputs of our methods.

In GBM-DX, the diagnostic slides at 20x magnification were tiled into 512 pixel $\times$ 512 pixel patches and the non-informative regions were filtered by the Otsu's method. Each patch was resized as 224 pixel $\times$ 224 pixel to be compatible with the input size of ResNet18. Notably, a resize-free alternative pipeline is to tile WSIs directly into patches of size 224 pixel $\times$ 224 pixel and store them in memory-efficient formats (e.g., OME-Zarr), which can be accessed via a PyTorch DataLoader [49]. Patches from different slides but the same patient were merged. The rest procedure followed the same pipeline as the three datasets used for MSI vs. MSS classification.

For HTSeq-FPKM records, genes with more than 20% missing values were excluded, and the top 2000 most variable genes were selected. Subsequently, a $\log(1+x)$ transformation was applied to the data.

We fixed the sample size $N = 200$ and repeated 100 times for collections of deep features of patches for one patient to generate the augmented dataset. See the Data augmentation section for details.

The training set and the test set were randomly splitted at patient-level for each cohort. In CRC-DX, 258 patients were used for training and 100 patients were used for test. After data augmentation, there were 25800 samples in the training set and 10000 samples in the test set. In GBM-DX, 127 patients were used for training and 55 patients were used for test. Similarly, we built a training set including 12700 samples and a test set including 5500 samples through data augmentation. There were 185 patients for training and 99 patients for test in STAD-DX. In CRC-KR, 278 patients were used for training and 109 patients were used for test. Through data augmentation, training sets and test sets were 100 times larger than the number of patients in each cancer cohort. A summary of datasets is presented in Table D S1 Appendix.

Technical details

Network architectures. We implemented PFN with two stacked building blocks. We summarized the shapes of parameters of each block in Fig G in S1 Appendix. The backbone of MTA consisted of six linear layers. The dimensional structure of layers is 2000-512-128-32-128-1024-2000. MTA's projection head was implemented by a two-layer fully connected network with structure 32-64-128. Batch normalization was applied before hidden features went through activation functions. The network architecture in a pytorch-style was presented Fig H in S1 Appendix.

Label smoothing. We applied label smoothing [50] in the cross-entropy loss Eq (20). Suppose $\boldsymbol{y} \in \{0,1\}^K$ is a one-hot coding of label $y \in \{1,2,\ldots,K\}$ and $\hat{\boldsymbol{y}} \in \Delta_K$ is the predicted likelihood that assigns to each class, where Δ_K is the K-dimensional simplex. Then the individual loss of the pair $(\boldsymbol{y}, \hat{\boldsymbol{y}})$ is rewritten as

$$\sum_{k=1}^{K} \tilde{\boldsymbol{y}}_k \log \hat{\boldsymbol{y}}_k \tag{22}$$

when label smoothing is applied, where $\tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k(1-\alpha) + \alpha/K$ and α is a hyperparameter in the range of [0,1].

Parameter settings. We trained our model by AdamW with 1e-3 as learning rate. The batch size was 2048. The temperature τ in Eq (15) was 0.05. The δ in Eq (21) and the λ in various total losses were set as 0.1 and 0.5, respectively, for balancing loss scales. The number of epochs varied in different datasets and total losses. It depended on the model's performance over training sets. In (TEMI w/AOD), the number of epochs when training CRC-DX and STAD-DX was set as 15 and it was 50 in training GBM-DX. In (TEMI w/APR) and (TEMI w/AOD+APR), the numbers of epochs were 15, 25, and 50 for CRC-DX, STAD-DX, and GBM-DX, respectively. We used a larger learning rate 5e-3 when training GBM-DX with (TEMI w/APR) and (TEMI w/AOD+APR). The parameter α for label smoothing was set as 0.05, 0.1, and 0.01 for CRC-DX, STAD-DX, GBM-DX, respectively. The masked ratio ξ was 0.7 in MTA. For scenarios that advanced foundation models were used as feature extractors in CRC-DX, the training parameters—including learning rate, optimizer, and numbers of epochs—are summarized in Table E in S1 Appendix.

Patch-level attention scores. We computed the final attention score of a patch by averaging its scores obtained from Eq (8) over the 100 repetitions for sampling.

Evaluation

Metrics. We evaluated models using the area under the receiver operating characteristic curve (AUC). The metric was computed at the patient level, with prediction for samples from the same patient aggregated by majority voting.

Bootstrapping. To ensure sufficient data for both training and testing, we did not create a separate independent validation set. To address this limitation, we performed 1,000-fold bootstrapping. The median and 95% confidence intervals for AUC are reported.

Alternative heterogenous data alignment

Data alignment is prevalent in multi-modal learning and transfer learning [51,52]. There are serveral typical criteria for data alignment. Suppose we are given a min-batch of pair-representation from two different data sources $\mathcal{H}$ and $\mathcal{Z}$ denoted by $\{(\boldsymbol{h}_i, \boldsymbol{z}_i)\}_{i=1}^{B}$.

Mean squared error. The mean squared error describes an exact alignment between different data sources.

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \|\boldsymbol{h}_i - \boldsymbol{z}_i\|_2^2. \tag{23}$$

Similarity consistency. Similarity consistency assumes similarity as an invariant metric in different representation spaces. Define the similarities between data points in $\mathcal{H}$ and $\mathcal{Z}$, respectively, as

$$s_{\mathcal{H},ij} = \frac{\exp(\boldsymbol{h}_i^T \boldsymbol{h}_j/\tau)}{\sum_{k=1}^{B} \exp(\boldsymbol{h}_i^T \boldsymbol{h}_k/\tau)}, \tag{24}$$

and

$$s_{\mathcal{Z},ij} = \frac{\exp(\boldsymbol{z}_i^T \boldsymbol{z}_j/\tau)}{\sum_{k=1}^{B} \exp(\boldsymbol{z}_i^T \boldsymbol{z}_k/\tau)}. \tag{25}$$

The similarity consistency is defined by Frobenius norm of the difference between similarity matrices

$$\mathcal{L} = \|\boldsymbol{S}_{\mathcal{H}} - \boldsymbol{S}_{\mathcal{Z}}\|_F. \tag{26}$$

where $\boldsymbol{S}_{\cdot} = [s_{\cdot,ij}]_{B\times B}$.

Hilbert-Schmidt independence criterion (HSIC). HSIC measures independence between two data sources $\mathcal{H}$ and $\mathcal{Z}$ in a reproducing kernel Hilbert space [53]. Its empirical estimator is written as

$$\text{HSIC}(\mathcal{H}, \mathcal{Z}) = \frac{1}{(B-1)^2}\text{trace}(\boldsymbol{K}_{\mathcal{H}}\boldsymbol{\Lambda}\boldsymbol{K}_{\mathcal{Z}}\boldsymbol{\Lambda}), \tag{27}$$

where $\boldsymbol{K}_{\mathcal{H}} = [k_{\mathcal{H}}(\boldsymbol{h}_i, \boldsymbol{h}_j)]_{B\times B}$ and $\boldsymbol{K}_{\mathcal{Z}} = [k(\boldsymbol{z}_i, \boldsymbol{z}_j)]_{B\times B}$, defined by reproducing kernels $k_{\mathcal{H}}(\cdot,\cdot)$ and $k_{\mathcal{Z}}(\cdot,\cdot)$, respectively, and $\boldsymbol{\Lambda} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/B \in \mathbb{R}^{B\times B}$. The symbol $\boldsymbol{I}$ indicates a B-by-B identity matrix, and $\boldsymbol{1}$ is a column vector with all its elements as 1. If $\text{HSIC}(\mathcal{H}, \mathcal{Z}) = 0$, then data points from two different data sources are independence. Thus, we maximize the criterion to achieve alignment. The loss function is

$$\mathcal{L} = -\text{HSIC}(\mathcal{H}, \mathcal{Z}). \tag{28}$$

Maximum mean discrepancy (MMD). MMD is a well-known divergence that measures the distance between distributions [54]. It is commonly used for minimizing the gap between data sources in domain adaptation. Let $P_{\mathcal{H}}$ and $P_{\mathcal{Z}}$ denote

distributions. A biased MMD estimator is

$$\mathcal{L} = \left[\frac{1}{B^2} \sum_{i,j=1}^{B} k(\boldsymbol{h}_i, \boldsymbol{h}_j) - \frac{2}{B^2} \sum_{i,j=1}^{B} k(\boldsymbol{h}_i, \boldsymbol{z}_j) + \frac{1}{B^2} \sum_{i,j=1} k(\boldsymbol{z}_i, \boldsymbol{z}_j) \right]^{\frac{1}{2}}. \tag{29}$$

where $k(\cdot, \cdot)$ is a kernel defined on the support set of $P_{\mathcal{H}}$ and $P_{\mathcal{Z}}$.

Gaussian Wasserstein distance. Gaussian Wasserstein distance is a divergence that measures the difference between Gaussian distribution [55]. Given Gaussian distributions $\mathcal{N}(\boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Sigma}_{\mathcal{H}})$ and $\mathcal{N}(\boldsymbol{m}_{\mathcal{Z}}, \boldsymbol{\Sigma}_{\mathcal{Z}})$ on the support set $\mathcal{U} \subset \mathbb{R}^d$, the 2-Wasserstein distance is written as

$$\|\boldsymbol{m}_{\mathcal{H}} - \boldsymbol{m}_{\mathcal{Z}}\|_2^2 + \text{trace}\left(\boldsymbol{\Sigma}_{\mathcal{H}} + \boldsymbol{\Sigma}_{\mathcal{Z}} - 2\left(\boldsymbol{\Sigma}_{\mathcal{H}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\mathcal{Z}} \boldsymbol{\Sigma}_{\mathcal{H}}^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \tag{30}$$

From simplicity, we assume that $\boldsymbol{\Sigma}_{\mathcal{H}}$ and $\boldsymbol{\Sigma}_{\mathcal{Z}}$ are commuting. The Eq (30) can be simplified as

$$\mathcal{L} = \|\boldsymbol{m}_{\mathcal{H}} - \boldsymbol{m}_{\mathcal{Z}}\|_2^2 + \|\boldsymbol{\Sigma}_{\mathcal{H}}^{\frac{1}{2}} - \boldsymbol{\Sigma}_{\mathcal{Z}}^{\frac{1}{2}}\|_F^{\frac{1}{2}}. \tag{31}$$

Note that data points from different sources are not necessarily to be paired in computing HSIC, MMD, or 2-Wasserstein distance.

Gaussian kernel with bandwidth $k(\boldsymbol{h}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{h}-\boldsymbol{z}\|_2^2}{\delta}\right)$ is commonly used when computing HSIC and MMD. The bandwidth δ takes the average of Euclidean distances between data points.

These alignment criteria assume exact alignment between data sources, whether in terms of representation or distribution. In contrast, given the substantial differences between whole-slide images and molecular signals, our approach acknowledges that only partial alignment of information is possible.

Supporting information

S1 Appendix. Supplementary Materials. This supporting document contains all supplementary tables and figures cited in the main text. It includes the following sections:

- Enrichment analysis in STAD-DX
- Analysis of attention scores
- Foundation models as feature extractors
- Distributions of mean squared errors in GBM-DX
- Memory cost of the dot-product attention
- Descriptions of datasets
- Architecture of patch fusion network
- Architecture of masked transcriptomic autoencoder
- Training settings of TEMI with foundation-model features

(PDF)

Author contributions

Conceptualization: Weiwen Wang, Yuanyan Xiong.

Formal analysis: Weiwen Wang, Xiwen Zhang.

Funding acquisition: Weiwen Wang, Yuanyan Xiong.

Methodology: Weiwen Wang.

Visualization: Weiwen Wang.

Writing – original draft: Weiwen Wang, Xiwen Zhang.

Writing – review & editing: Weiwen Wang, Xiwen Zhang.

References

1. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021;27(5):775–84. https://doi.org/10.1038/s41591-021-01343-4 PMID: 33990804
2. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70. https://doi.org/10.1038/s41551-020-00682-w PMID: 33649564
3. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9. https://doi.org/10.1038/s41591-019-0508-1 PMID: 31308507
4. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis Oncol. 2020;4:14. https://doi.org/10.1038/s41698-020-0120-3 PMID: 32550270
5. Zhao Y, Xiong S, Ren Q, Wang J, Li M, Yang L, et al. Deep learning using histological images for gene mutation prediction in lung cancer: a multicentre retrospective study. Lancet Oncol. 2025;26(1):136–46. https://doi.org/10.1016/S1470-2045(24)00599-0 PMID: 39653054
6. Shao W, Han Z, Cheng J, Cheng L, Wang T, Sun L, et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. IEEE Trans Med Imaging. 2020;39(1):99–110. https://doi.org/10.1109/TMI.2019.2920608 PMID: 31170067
7. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018;115(13):E2970–9. https://doi.org/10.1073/pnas.1717139115 PMID: 29531073
8. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054–6. https://doi.org/10.1038/s41591-019-0462-y PMID: 31160815
9. Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, et al. Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. Nat Commun. 2023;14(1):6695. https://doi.org/10.1038/s41467-023-42453-6 PMID: 37932267
10. Chang X, Wang J, Zhang G, Yang M, Xi Y, Xi C, et al. Predicting colorectal cancer microsatellite instability with a self-attention-enabled convolutional neural network. Cell Rep Med. 2023;4(2):100914. https://doi.org/10.1016/j.xcrm.2022.100914 PMID: 36720223
11. Jin X, Zhou Y-F, Ma D, Zhao S, Lin C-J, Xiao Y, et al. Molecular classification of hormone receptor-positive HER2-negative breast cancer. Nat Genet. 2023;55(10):1696–708. https://doi.org/10.1038/s41588-023-01507-7 PMID: 37770634
12. Zhu X-Z, Zhang H-Y-L, Zhou Y-F, Chen Y-Y, Fu T, Jin M-L, et al. Subtyping-directed precision treatment refines traditional one-size-fits-all therapy for HR+/HER2- breast cancer. Cancer Res. 2025;85(20):3983–98. https://doi.org/10.1158/0008-5472.CAN-24-5002 PMID: 40824544
13. Shi X, Su H, Xing F, Liang Y, Qu G, Yang L. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med Image Anal. 2020;60:101624. https://doi.org/10.1016/j.media.2019.101624 PMID: 31841948
14. Marini N, Otálora S, Müller H, Atzori M. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Med Image Anal. 2021;73:102165. https://doi.org/10.1016/j.media.2021.102165 PMID: 34303169
15. Pati P, Foncubierta-Rodríguez A, Goksel O, Gabrani M. Reducing annotation effort in digital pathology: a co-representation learning framework for classification tasks. Med Image Anal. 2021;67:101859. https://doi.org/10.1016/j.media.2020.101859 PMID: 33129150
16. Zhu C, Chen W, Peng T, Wang Y, Jin M. Hard sample aware noise robust learning for histopathology image classification. IEEE Trans Med Imaging. 2022;41(4):881–94. https://doi.org/10.1109/TMI.2021.3125459 PMID: 34735341
17. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019;54:280–96. https://doi.org/10.1016/j.media.2019.03.009 PMID: 30959445
18. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. Conf Comput Vis Pattern Recognit Workshops. 2021;2021:14318–28. https://doi.org/10.1109/CVPR46437.2021.01409 PMID: 35047230
19. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021). 2021. p. 2136–47.
20. Lin T, Yu Z, Hu H, Xu Y, Chen CW. Interventional bag multi-instance learning on whole-slide pathological images. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. p. 19830–9. https://doi.org/10.1109/cvpr52729.2023.01899

21. Castro-Macías FM, Morales-Álvarezm P, Wu Y, Molina R, Katsaggelos AK. Sm: enhanced localization in multiple instance learning for medical imaging classification. In: 38th Conference on Neural Information Processing Systems (NeurIPS). 2024.
22. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;630(8015):181–8. https://doi.org/10.1038/s41586-024-07441-w PMID: 38778098
23. Ding J, Ma S, Dong L, Zhang X, Huang S, Wang W, et al. Longnet: scaling transformers to 1,000,000,000 tokens. arXiv preprint 2023. https://doi.org/10.48550/arXiv.2307.02486
24. Wang X, Zhao J, Marostica E, Yuan W, Jin J, Zhang J, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature. 2024;634(8035):970–8. https://doi.org/10.1038/s41586-024-07894-z PMID: 39232164
25. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850–62. https://doi.org/10.1038/s41591-024-02857-3 PMID: 38504018
26. Bioptimus. H-optimus-1. 2025. https://huggingface.co/bioptimus/H-optimus-1
27. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med. 2024;30(10):2924–35. https://doi.org/10.1038/s41591-024-03141-0 PMID: 39039250
28. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: learning robust visual features without supervision. arXiv preprint 2023. https://arxiv.org/abs/2304.07193
29. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: The 9th International Conference on Learning Representations (ICLR). 2021.
30. Liu Y, Wang W, Ren CX, Dai DQ. MetaCon: meta contrastive learning for microsatellite instability detection. In: 24th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021). 2021. p. 267–76.
31. Srinidhi CL, Kim SW, Chen F-D, Martel AL. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med Image Anal. 2022;75:102256. https://doi.org/10.1016/j.media.2021.102256 PMID: 34717189
32. Gong D, Arbesfeld-Qiu JM, Perrault E, Bae JW, Hwang WL. Spatial oncology: Translating contextual biology to the clinic. Cancer Cell. 2024;42(10):1653–75. https://doi.org/10.1016/j.ccell.2024.09.001 PMID: 39366372
33. Zhang D, Schroeder A, Yan H, Yang H, Hu J, Lee MYY, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. Nat Biotechnol. 2024;42(9):1372–7. https://doi.org/10.1038/s41587-023-02019-9 PMID: 38168986
34. Bao R, Hutson A, Madabhushi A, Jonsson VD, Rosario SR, Barnholtz-Sloan JS, et al. Ten challenges and opportunities in computational immuno-oncology. J Immunother Cancer. 2024;12(10):e009721. https://doi.org/10.1136/jitc-2024-009721 PMID: 39461879
35. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans Med Imaging. 2022;41(4):757–70. https://doi.org/10.1109/TMI.2020.3021387 PMID: 32881682
36. Ning Z, Du D, Tu C, Feng Q, Zhang Y. Relation-aware shared representation learning for cancer prognosis analysis with auxiliary clinical variables and incomplete multi-modality data. IEEE Trans Med Imaging. 2022;41(1):186–98. https://doi.org/10.1109/TMI.2021.3108802 PMID: 34460368
37. Shao W, Wang T, Sun L, Dong T, Han Z, Huang Z, et al. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. Med Image Anal. 2020;65:101795. https://doi.org/10.1016/j.media.2020.101795 PMID: 32745975
38. Carrillo-Perez F, Pizurica M, Zheng Y, Nandi TN, Madduri R, Shen J, et al. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. Nat Biomed Eng. 2024.
39. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11(1):3877. https://doi.org/10.1038/s41467-020-17678-4 PMID: 32747659
40. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. p. 15979–88. https://doi.org/10.1109/cvpr52688.2022.01553
41. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020). 2020. p. 1597–607.
42. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020). 2020. p. 18661–73.
43. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018). 2018. p. 2127–36.
44. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med. 2019;25(10):1519–25. https://doi.org/10.1038/s41591-019-0583-3 PMID: 31591589
45. Ratti M, Lampis A, Hahne JC, Passalacqua R, Valeri N. Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches. Cell Mol Life Sci. 2018;75(22):4151–62. https://doi.org/10.1007/s00018-018-2906-9 PMID: 30173350
46. Vargas-Castellanos E, Rincón-Riveros A. Microsatellite instability in the tumor microenvironment: the role of inflammation and the microbiome. Cancer Med. 2025;14(8):e70603. https://doi.org/10.1002/cam4.70603 PMID: 40231893
47. McInnes L, Healy J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint 2018. https://arxiv.org/abs/1802.03426

48. Wang L, Babikir H, Müller S, Yagnik G, Shamardani K, Catalan F, et al. The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. Cancer Discov. 2019;9(12):1708–19. https://doi.org/10.1158/2159-8290.CD-19-0329 PMID: 31554641
49. The Jackson Laboratory. ZarrDataset. 2025. [cited 2026 Jan 27]. https://github.com/TheJacksonLaboratory/zarrdataset
50. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). 2016. p. 2818–26.
51. Baltrusaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43. https://doi.org/10.1109/TPAMI.2018.2798607 PMID: 29994351
52. Kouw WM, Loog M. A review of domain adaptation without target labels. IEEE Trans Pattern Anal Mach Intell. 2021;43(3):766–85. https://doi.org/10.1109/TPAMI.2019.2945942 PMID: 31603771
53. Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B. Kernel mean embedding of distributions: a review and beyond. FNT in Machine Learning. 2017;10(1–2):1–141. https://doi.org/10.1561/2200000060
54. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. J Mach Learn Res. 2012;13:723–73.
55. He R, Wu X, Sun Z, Tan T. Wasserstein CNN: learning invariant features for NIR-VIS face recognition. IEEE Trans Pattern Anal Mach Intell. 2019;41(7):1761–73. https://doi.org/10.1109/TPAMI.2018.2842770 PMID: 29993534