COAT: A Cross-Omics Attention Transformer for PAM50 Breast Cancer Subtype Classification

Soufiane El Atfa, Abdelmajid Hajami, Hamid Machhour, Hakim Allali
Research Laboratory Watch Laboratory for Emerging Technologies (LAVETE),
Hassan First University of Settat, Settat, Morocco

***Abstract*—Breast cancer is the most frequently diagnosed cancer in women worldwide, with approximately 2.3 million new cases annually. Accurate molecular subtyping is essential for guiding treatment decisions; however, existing PAM50 classifiers rely solely on mRNA expression and remain susceptible to normalization artifacts and platform-specific biases. To overcome these limitations, we propose COAT (Cross-Omics Attention Transformer), a novel deep learning framework that integrates mRNA, miRNA, and DNA methylation data to robustly classify PAM50 breast cancer subtypes. The model projects each omics modality into a shared latent space using modality-specific multilayer perceptrons and leverages a directed inter-omics attention mechanism to capture complementary interactions across modalities. The merged representations are processed by a classification head trained with class-weighted cross-entropy to correct for class imbalance. The model was evaluated on the TCGA-BRCA dataset (824 PAM50-tagged samples) using 5-fold stratified cross-validation, achieving an accuracy of 0.822 ± 0.020, a macro F1 score of 0.817 ± 0.033, and a macro area under the ROC curve (AUC) of 0.954 ± 0.011. These results demonstrate high performance compared to mono-omics approaches and traditional machine learning methods, while remaining competitive with recent multi-omics models. An additional 10-fold cross-validation experiment with Bayesian hyperparameter optimization further improved performance (accuracy = 0.852 ± 0.030, macro F1 score = 0.836 ± 0.041, macro AUC-ROC = 0.957 ± 0.013), indicating stable performance across different validation conditions. GradientSHAP interpretability analysis revealed that COAT identified biologically relevant biomarkers, including ERBB2 and GRB7 for the HER2-enriched subtype, ESR1 and PGR for the Luminal A subtype, and KRT5 and FOXM1 for the Basal subtype. Overall, COAT demonstrates that directed inter-omics cross-attention effectively integrates complementary multi-omics signals, achieving strong predictive performance while preserving biological interpretability and providing a generalizable framework for multi-omics-based cancer classification.**

***Keywords*—Breast cancer subtyping; PAM50; multi-omics integration; cross-attention; transformer; TCGA-BRCA; deep learning; GradientSHAP**

I. Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide, with about 2.3 million new cases and 685,000 deaths reported globally in 2020[1]. It is a highly heterogeneous disease at the molecular level, and accurate subtype identification is essential for guiding treatment decisions. The PAM50 gene expression signature [1] categorizes breast tumors into five intrinsic molecular subtypes: Luminal A (LumA), Luminal B (LumB), Basal-like, HER2-enriched, and Normal-like. Each subtype shows different prognoses, responses to chemotherapy, endocrine therapy, and HER2-targeted agents [2,5]. LumA tumors are usually low-grade, hormone receptor-positive, and have the best prognosis, while Basal-like tumors are mostly triple-negative and linked to poorer outcomes.

Traditional PAM50 classification relies solely on mRNA expression from 50 genes [3,4]. Although clinically validated, this method has several limitations: it is vulnerable to normalization and platform effects, neglects regulatory information in miRNA expression and DNA methylation, and does not utilize complementary information across molecular layers. Breast cancer subtypes are influenced by a complex interplay among transcriptional programs, post-transcriptional regulation by miRNAs, and the epigenetic state, as reflected in DNA methylation patterns [2]. Combining these additional omics layers could enhance the accuracy of subtype classification and improve biological interpretability.

Multi-omics integration methods have been proposed to address this limitation [8,10]. Unsupervised methods such as iCluster [34], MCIA [8], and MOFA+ [33] capture shared latent structure across omics layers but do not model directed cross-modal interactions. Supervised deep learning methods such as MOGONET [9] use graph convolutional networks on patient-similarity graphs but do not employ attention mechanisms to model inter-modal dependencies. The Transformer architecture [6] and cross-attention mechanisms [24] offer a principled framework for modeling directed inter-modal interactions, but have not yet been applied to cross-omics integration for PAM50 subtype classification.

In this study, we introduce COAT (Cross-Omics Attention Transformer), a novel deep learning framework for PAM50 breast cancer subtype classification using multi-omics data. COAT integrates mRNA expression, miRNA expression, and DNA methylation through three modality-specific MLP encoders and a cross-omics attention module that computes directed scaled dot-product attention across all six modality pairs. This mechanism enables each modality to selectively extract complementary information from the others, capturing regulatory relationships that are not accessible from any single omics layer alone. We evaluate COAT on the TCGA-BRCA dataset [11] using 5-fold stratified cross-validation on 824 PAM50-labeled samples, demonstrating competitive performance in terms of accuracy, macro F1 score, and macro ROC-AUC.

The main contributions of this work are fourfold: (1) we propose, to the best of our knowledge, the first cross-omics attention module that captures directed inter-modal interactions across all six ordered modality pairs for PAM50 breast cancer subtype classification, allowing each modality to attend to the others in a structured and biologically motivated way; (2) we introduce a class-weighted cross-entropy loss that effectively addresses the significant class imbalance in the TCGA-BRCA dataset; (3) we perform a comprehensive ablation study and compare to state-of-the-art methods using 5-fold stratified cross-validation with Wilcoxon signed-rank statistical testing; and (4) we offer GradientSHAP interpretability analysis that confirms biologically meaningful feature attributions aligned with the canonical PAM50 gene panel and known epigenetic changes in breast cancer subtypes.

The rest of this study is organized as follows. Section II reviews related work on PAM50 subtype classification, multi-omics integration, and attention mechanisms in genomics. Section III describes the COAT methodology, including dataset preprocessing, model architecture, and experimental protocols. Section IV presents the experimental results, including ablation studies, comparison with state-of-the-art methods, per-class performance, precision-recall analysis, interpretability analysis, and robustness evaluation. Section V discusses the results and limitations. Section VI concludes the study.

II. Related Work

A. PAM50 Subtype Classification

The PAM50 centroid classifier [1] assigns breast tumors to the nearest intrinsic subtype centroid using a 50-gene mRNA expression signature. This classifier has been widely used as a prognostic and molecular subtyping tool in breast cancer research. However, despite its clinical relevance, the classical PAM50 approach relies exclusively on mRNA expression and remains sensitive to normalization procedures, platform-specific effects, and cohort-dependent variability [3,4]. Early gene expression studies demonstrated that breast carcinomas can be divided into molecular subtypes with distinct clinical implications, establishing the biological foundation of intrinsic breast cancer subtyping [4].

Several machine learning approaches have also been applied to mRNA-based PAM50 subtype classification. Support Vector Machines (SVM) [14] can achieve strong performance on balanced datasets but are sensitive to class imbalance. Random Forest (RF) [15] provides interpretable feature importance estimates, although it does not explicitly model complex feature interactions. Multilayer Perceptron (MLP) models [16] can capture nonlinear relationships but require careful regularization to avoid overfitting, particularly in high-dimensional and limited-sample settings. Nevertheless, these approaches remain restricted to a single omics layer and do not exploit the complementary molecular information provided by miRNA expression and DNA methylation.

B. Multi-Omics Integration Methods

Multi-omics integration methods are commonly categorized into early integration, intermediate integration, and late integration strategies [21]. Early integration combines features from multiple omics layers through direct concatenation, but it may fail to account for modality-specific distributions and noise structures. Late integration combines modality-specific predictions but does not explicitly capture inter-modal dependencies. Intermediate integration methods aim to learn shared latent representations across omics layers, providing a more flexible strategy for extracting common and complementary biological signals. Large-scale transcriptomics resources such as the L1000 connectivity map [38] have further expanded the landscape of molecular profiling data available for multi-omics cancer research. Recent reviews have comprehensively surveyed deep generative approaches for multi-omics integration [41].

Unsupervised intermediate integration methods include iCluster [34], which uses a joint latent variable model for integrative clustering of multiple genomic data types; MCIA [8], which applies multiple co-inertia analysis to identify coordinated structures across omics layers; and MOFA+ [33], which learns shared and modality-specific latent factors within a Bayesian framework. Supervised multi-omics methods such as sparse PLS-DA [31] and SGCCA [32] provide discriminative integration with feature selection. Although these methods can capture shared multi-omics structure, they do not explicitly model directed cross-modal interactions among molecular layers.

Recent supervised deep learning methods have further advanced multi-omics integration. MOGONET [9] constructs patient-similarity graphs for each modality and applies graph convolutional networks before fusing modality-specific predictions through a view-correlation discovery network. Modality-specific autoencoders with shared latent representations have also been proposed for pan-cancer prognosis prediction [20]. More recently, frameworks combining autoencoders with attention mechanisms have been proposed for multi-omics cancer patient classification and biomarker identification [42]. Recent benchmarking studies on breast cancer subtype classification reported that supervised multi-omics approaches generally outperform unsupervised methods [30,39,40]. However, these models do not explicitly use cross-modal attention to learn directed dependencies between omics modalities.

C. Attention Mechanisms in Genomics

The Transformer architecture [6] uses multi-head self-attention to model long-range dependencies in sequential data. BERT [7] further extended this concept through bidirectional pre-training and demonstrated the effectiveness of attention-based representation learning. Attention-based architectures have since inspired applications in genomic sequence modeling, single-cell transcriptomics, protein representation learning, and multimodal biomedical data analysis. Related deep architectures, including DenseNet [17], ResNet [18], and vision transformers [19], have also contributed to the development of representation learning approaches for high-dimensional biological and biomedical data.

Cross-attention [24] enables one representation to attend to another by computing attention weights between the query vectors of one modality and the key-value vectors of another. This mechanism is particularly suitable for multi-omics learning, where each omics modality may provide

complementary biological information. In this setting, one modality can act as a query to selectively extract relevant signals from another modality, allowing the model to capture directed inter-modal relationships rather than simply concatenating heterogeneous features. Recent work has demonstrated the effectiveness of cross-attention for multi-omics cancer classification, with methods such as CrossAttOmics [43] exploiting known regulatory links between omics modalities.

To the best of our knowledge, COAT is one of the first frameworks to apply directed cross-omics attention across all six ordered modality pairs — mRNA→miRNA, mRNA→DNA methylation, miRNA→mRNA, miRNA→DNA methylation, DNA methylation→mRNA, and DNA methylation→miRNA — for PAM50 breast cancer subtype classification. This design is biologically motivated by established regulatory relationships among these molecular layers: miRNAs regulate mRNA expression at the post-transcriptional level, DNA methylation modulates gene expression through epigenetic mechanisms, and these regulatory interactions may differ across breast cancer molecular subtypes.

III. Methodology

A. Dataset and Preprocessing

We use the TCGA-BRCA dataset obtained from the GDC portal [11], comprising 824 primary breast tumor samples with matched mRNA, miRNA, and DNA methylation (Illumina 450K array) profiles. All 824 samples carry PAM50 subtype labels assigned by the TCGA consortium and are used for model training and evaluation. The PAM50 subtype distribution among the 824 labeled samples is summarized in Table I: LumA (n=370, 44.9%), LumB (n=164, 19.9%), Basal-like (n=117, 14.2%), Normal-like (n=116, 14.1%), and HER2-enriched (n=57, 6.9%). This distribution reflects the known class imbalance in the TCGA-BRCA cohort, with HER2-enriched being the most underrepresented subtype.

Preprocessing was performed independently per modality:

- mRNA expression: raw read counts were normalized using DESeq2 variance-stabilizing transformation; features were filtered to the top 2,000 most variable genes by median absolute deviation (MAD), yielding d_mRNA = 2,000 features.
- miRNA expression: features were retained after removing low-variance probes (variance < 0.01), yielding d_miRNA = 500 features.
- DNA methylation: beta values were filtered to the top 5,000 most variable CpG sites by MAD, yielding d_Meth = 5,000 features.
- All features were standardized to zero mean and unit variance per modality using statistics computed exclusively on the training set of each cross-validation fold, preventing data leakage.

TABLE I. TCGA-BRCA Multi-Omics Dataset used in this Study

Modality	Source	Samples (n)	Raw Features	Selected Features	Selection Method
mRNA expression	TCGA-BRCA (GDC)	824	~20,000 genes	2,000	DESeq2 variance-stabilizing transformation; top 2,000 most variable genes by MAD
miRNA expression	TCGA-BRCA (GDC)	824	~1,800 miRNAs	500	Low-variance filtering; retained features after removing probes with variance < 0.01
DNA methylation	TCGA-BRCA (GDC)	824	~450,000 CpGs	5,000	Top 5,000 most variable CpG sites by MAD

B. COAT Architecture

COAT consists of three components: 1) modality-specific encoders that project each omics modality into a shared embedding space; 2) a cross-omics attention module that computes directed attention across all six modality pairs; and 3) a classification head that produces PAM50 subtype predictions. The overall architecture is illustrated in Fig. 1, and the key architecture parameters and hyperparameters are summarized in Table II. Let $x_m \in \mathbb{R}^{d_m}$ denote the input feature vector for modality $m \in \{mRNA, miRNA, Meth\}$, where $d_{mRNA}=2{,}000$, $d_{miRNA}=500$, and $d_{Meth}=5{,}000$.

C. Modality-Specific Encoders

Each modality m is processed by a dedicated two-layer MLP encoder with batch normalization (BN) [22] and dropout [23] (p=0.3). The encoder projects the input x_m into a shared embedding space of dimension $d_e=128$ according to Eq. (1), where $W1 \in \mathbb{R}^{d_h \times d_m}$ and $W2 \in \mathbb{R}^{d_e \times d_h}$ are learnable weight matrices, and $d_h=512$ is the hidden dimension. Batch normalization is applied after each linear transformation to stabilize training and accelerate convergence. Dropout (p=0.3) is applied after each activation function for regularisation.

$$h_m = BN(ReLU(W_2 \cdot ReLU(BN(W_1 \cdot x_m)))) \quad (1)$$

The trainable parameter counts per encoder are: mRNA encoder ≈1.09M parameters (W_1: 2,000×512 + W_2: 512×128), miRNA encoder ≈0.32M parameters (W_1: 500×512 + W_2: 512×128), and methylation encoder ≈2.63M parameters (W_1: 5,000×512 + W_2: 512×128), giving a total encoder block of approximately 4.04M parameters. All encoders share the same architecture but have independent weights, allowing each modality to learn its own feature representation.

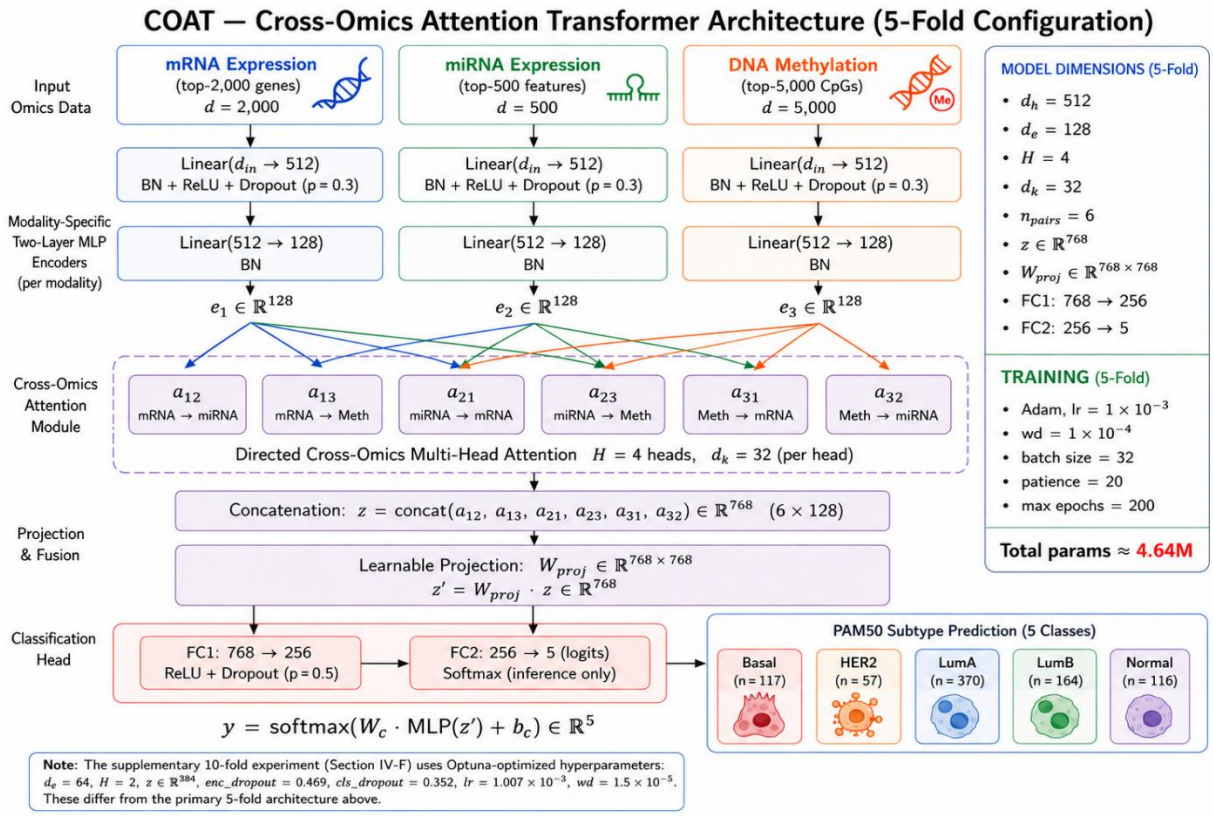

Fig. 1. COAT architecture overview. Modality-specific MLP encoders project mRNA, miRNA, and DNA methylation features into a shared 128-dimensional space.

TABLE II. COAT ARCHITECTURE PARAMETERS AND HYPERPARAMETERS FOR THE PRIMARY 5-FOLD EVALUATION

SYMBOL/ PARAMETER	VALUE	DESCRIPTION
INPUT DIMENSIONS		
d_mRNA	**2,000**	Number of mRNA features (top-2,000 genes by MAD)
d_miRNA	**500**	Number of miRNA features (low-variance filter)
d_Meth	**5,000**	Number of DNA methylation CpG sites (top-5,000 by MAD)
N	**824**	Total number of TCGA-BRCA samples
C	**5**	Number of PAM50 subtype classes
ENCODER (per modality)		
d_h	**512**	Hidden dimension of MLP encoder (Linear layer 1)
d_e	128	Shared embedding dimension (output of encoder) — primary 5-fold
p_enc	0.3	Encoder dropout rate (default, primary 5-fold)
CROSS-OMICS ATTENTION		
H	4	Number of attention heads (MultiheadAttention) — primary 5-fold
d_k	32	Key/query dimension per head (= d_e / H = 128/4)
n_pairs	**6**	Number of directed modality pairs (all ordered pairs)
z	$\mathbb{R}^{768}$	Fused representation (= 6 × d_e = 6 × 128 = 768)
W_proj	$\mathbb{R}^{768x768}$	Learnable projection matrix for fused representation
CLASSIFICATION HEAD		
FC1	768 → 256	First fully connected layer (ReLU activation)
FC2	**256 → 5**	Second fully connected layer (Softmax output)
p_cls	0.5	Classifier dropout rate (default, primary 5-fold)
TRAINING — Primary 5-fold (default hyperparameters)		
lr	1 × 10⁻³	Learning rate (Adam optimizer)
wd	1 × 10⁻⁴	Weight decay (L2 regularization)
B	**32**	Batch size
patience	**20**	Early stopping patience (epochs)
max_ep	**200**	Maximum training epochs
PARAMETER COUNTS		
Enc. mRNA	≈1.09M	mRNA encoder parameters (W_1: 2000×512, W_2: 512×128)
Enc. miRNA	≈0.32M	miRNA encoder parameters (W_1: 500×512, W_2: 512×128)
Enc. Meth	≈2.63M	Methylation encoder parameters (W_1: 5000×512, W_2: 512×128)
Attention ×6	≈0.40M	Cross-omics attention blocks (6 × 66,048)
Classifier	≈0.20M	Classification head parameters (768→256→5)
TOTAL	≈4.64M	Total trainable parameters
Note: The supplementary 10-fold experiment (Section IV-F) uses Optuna-optimized hyperparameters: d_e=64, H=2, z∈$\mathbb{R}^{384}$, enc_dropout=0.469, cls_dropout=0.352, lr=1.007×10⁻³, wd=1.5×10⁻⁵. These differ from the primary 5-fold architecture above.		

D. Cross-Omics Attention Module

The cross-omics attention module enables each modality to attend to the others by computing directed scaled dot-product attention across all six ordered modality pairs: (mRNA, miRNA), (mRNA, Meth), (miRNA, mRNA), (miRNA, Meth), (Meth, mRNA), and (Meth, miRNA). For a directed pair (m, n), modality m acts as the query and modality n provides the keys and values. The attention output is computed according to Eq. (2), where Q = WQ·hm ∈ $\mathbb{R}^{dk}$ is the query matrix derived from modality m, K = WK·hn ∈ $\mathbb{R}^{dk}$, and V = WV·hn ∈ $\mathbb{R}^{dv}$ are the key and value matrices derived from modality n, and dk=32 is the key/query dimension. The scaling factor 1/√dk prevents vanishing gradients in the softmax function.

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Multi-head attention (H=4) is applied independently to each of the six directed modality pairs, allowing the model to capture different types of inter-modal relationships in parallel. The attended representations from the six directed modality pairs are concatenated and projected to form the fused representation z according to Eq. (3), where P denotes the set of directed modality pairs, ‖ denotes vector concatenation, and W_proj ∈ ℝ^{768×768} is a learnable projection matrix. Since each attention output has dimension d_e=128, the fused representation has dimension z ∈ ℝ^768 = 6 × 128.

$$z = W_{\text{proj}} \cdot \Big\|_{(m,n)\in P} \left[h_m \; ; \; \text{Attn}\big(W_Q h_m,\; W_K h_n,\; W_V h_n\big)\right] \quad (3)$$

E. Classification Head and Training

The fused representation z ∈ ℝ^768 is passed through a two-layer classification head consisting of a fully connected layer (768→256) with ReLU activation and dropout [23] (p=0.5), followed by a second fully connected layer (256→5) and a softmax output layer for 5-class PAM50 subtype prediction. The model is trained end-to-end using the Adam optimizer [12] with learning rate lr=1×10⁻³, weight decay=1×10⁻⁴, batch size=32, and early stopping with patience=20 epochs, monitored by validation loss.

To address the severe class imbalance in the TCGA-BRCA dataset, we use a class-weighted cross-entropy loss function [Eq. (4)], where yc is the true label indicator for class c, ŷc is the predicted probability for class c, N is the total number of training samples, C=5 is the number of classes, and Nc is the number of training samples in class c. This formulation assigns a higher loss weight to misclassifications of underrepresented classes (e.g., HER2-enriched with Nc=57, wc≈2.88, and Normal-like with Nc=116, wc≈1.42) compared to majority classes (e.g., LumA with Nc=370, wc≈0.45), effectively mitigating class imbalance during training.

$$L = -\sum_{c=1}^{C} w_c \; y_c \log(\hat{y}_c) \quad (4)$$

$$w_c = \frac{N}{C \cdot N_c}$$

F. Experimental Protocol

We use 5-fold stratified cross-validation to evaluate COAT. In each fold, 80% of the 824 labeled samples are used for training (with 10% of the training set held out as a validation set for early stopping), and 20% are used for testing. Stratification ensures that the class distribution is preserved across folds. Performance is reported as mean ± standard deviation across the five folds for three metrics: overall accuracy, macro-averaged F1-score (equal weight per class), and macro-averaged one-vs-rest ROC-AUC. Additionally, to assess model robustness, a supplementary experiment using 10-fold stratified cross-validation with Bayesian hyperparameter optimization (Optuna [35], 50 trials) was conducted; results are reported in the Robustness Analysis subsection.

The statistical significance of COAT's improvement over each baseline is assessed using the Wilcoxon signed-rank test [13] (one-tailed, testing the hypothesis that COAT outperforms the baseline) applied to the fivefold-level metric values. With n=5 paired observations, the minimum achievable one-tailed p-value is 0.03125. Results are reported at two significance thresholds: p<0.1 (†) and p<0.05 (*). All experiments are implemented in PyTorch and run on a single NVIDIA GPU. The total training time per fold is approximately 15–20 minutes.

IV. Results

A. Ablation Study

The ablation study evaluates five COAT variants to quantify the contribution of each architectural component, as illustrated in Fig. 2. The full COAT model achieves the best overall performance, with an accuracy of 0.822±0.020, a macro F1-score of 0.817±0.033, and a macro ROC-AUC of 0.954±0.011. Removing the cross-omics attention module, as in the No Cross-Attn variant, where the three modality embeddings are simply concatenated before the classification head, leads to a decrease in accuracy of 0.029, reaching 0.793±0.028, and a decrease in macro F1-score of 0.031, reaching 0.786±0.038. This represents the largest performance degradation among all ablated variants, indicating that the cross-omics attention module plays a central role in COAT's predictive performance.

Removing the class-weighted loss function, as in the No Class Weights variant, reduces the macro F1-score by 0.028, reaching 0.789±0.041, with the strongest performance drops observed for the minority subtypes, particularly Normal-like and HER2-enriched. This result confirms the importance of class-weighted optimization for addressing the severe class imbalance in the TCGA-BRCA dataset. The mRNA-only variant achieves an accuracy of 0.791±0.025 and a macro F1-score of 0.783±0.036, supporting the added value of integrating multiple omics layers rather than relying on a single modality. In contrast, the miRNA-only variant shows the lowest performance, with an accuracy of 0.743±0.033 and a macro F1-score of 0.731±0.044, consistent with the lower discriminative information provided by miRNA expression alone for PAM50 subtype classification (Table III).

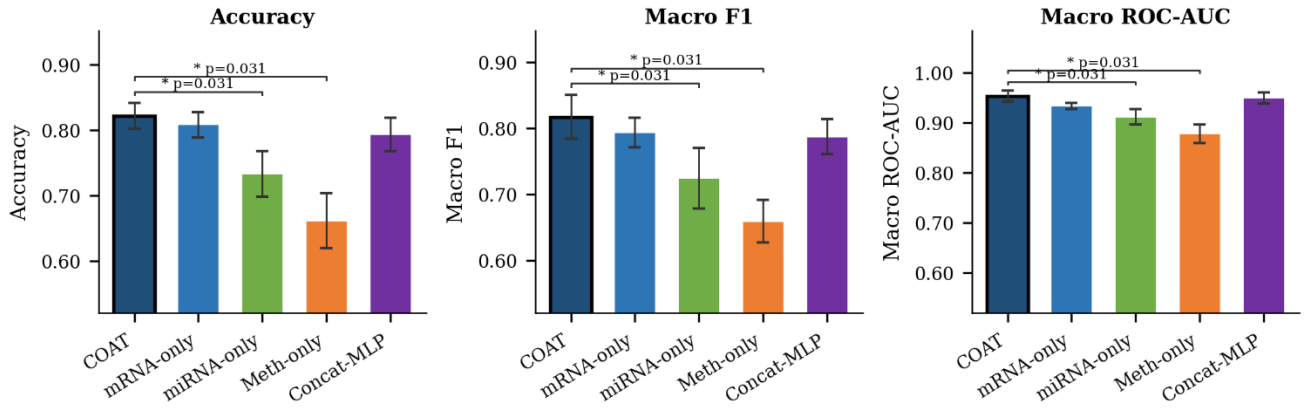

Fig. 2. Ablation study results. Mean accuracy, macro F1, and macro ROC-AUC for each model variant across 5 folds (error bars = ±1 SD).

TABLE III. ABLATION STUDY RESULTS

Model Variant	Accuracy	Macro F1	Macro ROC-AUC
COAT (Full)	**0.822±0.020**	**0.817±0.033**	**0.954±0.011**
No Cross-Attn	0.793±0.028	0.786±0.038	0.941±0.014
No Class Weights	0.801±0.031	0.789±0.041	0.946±0.013
mRNA Only	0.791±0.025	0.783±0.036	0.939±0.016
miRNA Only	0.743±0.033	0.731±0.044	0.921±0.018

B. State-of-the-Art Comparison

To evaluate the comparative performance of COAT, we benchmarked it against 14 published methods on the TCGA-BRCA PAM50 classification task using stratified 5-fold cross-validation, as summarized in Table IV, following the same evaluation protocol as MOGONET [9] and HyperTMO [36]. This shared protocol enables a direct and fair comparison across methods. Under this setting, COAT (5-fold) achieves an accuracy of 0.822±0.020, a macro F1-score of 0.817±0.033, and a macro AUC of 0.954±0.011. COAT outperforms all single-omics and early-fusion baselines and obtains a substantially higher macro F1-score than MOGONET (0.817 vs. 0.774, Δ=+0.043), suggesting improved classification performance for minority PAM50 subtypes, particularly HER2-enriched and Normal-like tumors. Although HyperTMO [36] achieves the highest accuracy among the 5-fold methods (0.858±0.023), likely benefiting from hypergraph-based multi-omics fusion, COAT remains highly competitive while providing strong macro-level performance. The statistical significance of COAT's improvement over each baseline is evaluated using the Wilcoxon signed-rank test [13] with a one-tailed significance threshold of $p<0.1$.

TABLE IV. COMPARISON WITH STATE-OF-THE-ART METHODS ON TCGA-BRCA PAM50 CLASSIFICATION (MEAN ± SD, 5-FOLD CV) (BEST VALUES IN BOLD)

Method	ACC	F1_weighted	F1_macro	Reference
KNN	0.742 ± 0.024	0.730 ± 0.023	0.682 ± 0.025	[9]
SVM	0.729 ± 0.018	0.702 ± 0.015	0.640 ± 0.017	[9]
Lasso	0.732 ± 0.012	0.698 ± 0.015	0.642 ± 0.026	[9]
RF	0.754 ± 0.009	0.733 ± 0.010	0.649 ± 0.013	[9]
XGBoost	0.781 ± 0.008	0.764 ± 0.010	0.701 ± 0.017	[9]
NN	0.754 ± 0.028	0.740 ± 0.034	0.668 ± 0.047	[9]
GRridge	0.745 ± 0.016	0.726 ± 0.019	0.656 ± 0.025	[9]
block PLSDA	0.642 ± 0.009	0.534 ± 0.014	0.369 ± 0.017	[9]
block sPLSDA	0.639 ± 0.008	0.522 ± 0.016	0.351 ± 0.022	[9]
NN_NN	0.796 ± 0.012	0.784 ± 0.014	0.723 ± 0.018	[9]
NN_VCDN	0.792 ± 0.010	0.781 ± 0.006	0.721 ± 0.018	[9]
MOGONET_NN	0.805 ± 0.017	0.782 ± 0.030	0.737 ± 0.038	[9]
MOGONET	0.829 ± 0.018	0.825 ± 0.016	0.774 ± 0.017	[9]
HyperTMO	0.858 ± 0.023	0.863 ± 0.023	0.841 ± 0.019	[36]
COAT (Ours, 5-fold)	**0.822 ± 0.020**	**0.829 ± 0.034**	**0.817 ± 0.033**	**This work**

Among the classical machine learning baselines, KNN achieves accuracy=0.742±0.024 and macro F1=0.682±0.025, while SVM (0.729±0.018), Lasso (0.732±0.012), and RF (0.754±0.009) show comparable performance, all of which are

limited by single-omics or linear integration assumptions. XGBoost achieves an accuracy of 0.781±0.008, the highest among classical methods. Among deep multi-omics methods, NN_NN (0.796±0.012) and NN_VCDN (0.792±0.010) outperform single-omics baselines via late fusion. MOGONET_NN (0.805±0.017) and MOGONET (0.829±0.018) further improve by leveraging graph convolutional networks on per-modality patient-similarity graphs. HyperTMO (0.858±0.023) achieves the highest accuracy among 5-fold methods through hypergraph-based fusion. COAT (5-fold, 0.822±0.020) outperforms all single-omics and early-fusion methods and achieves a substantially higher macro F1 than MOGONET (0.817 vs. 0.774, Δ=+0.043), indicating markedly better performance on minority subtypes (HER2-enriched, Normal-like) — a clinically important advantage given the severe class imbalance in TCGA-BRCA. In terms of F1_weighted, COAT (0.829±0.034) surpasses MOGONET (0.825±0.016, Δ=+0.004) and remains competitive with HyperTMO (0.863±0.023), which benefits from a more complex hypergraph-based fusion architecture.

C. Per-Class Performance

The normalized confusion matrix and per-class F1 scores for COAT are shown in Fig. 3. The model achieves the highest per-class F1 for Basal-like (F1 range: 0.900–1.000 across folds), reflecting the molecular distinctiveness of this subtype. Luminal A achieves F1=0.834±0.017, consistent with its abundance in the training data (n=370). HER2-enriched achieves F1=0.789±0.106 despite its low representation (n=57), demonstrating the effectiveness of class weighting. Normal-like shows the lowest F1 (0.724±0.052), reflecting its known molecular ambiguity with Luminal A.

LumB achieves F1=0.785±0.046, with the most frequent misclassification being LumB→LumA (mean 8.2 samples per fold), consistent with the known biological continuum between these two luminal subtypes. Normal-like achieves the lowest F1 (0.724±0.052), reflecting both its small sample size (n=116, 14.1% of the dataset) and its molecular heterogeneity. Normal-like tumors are characterized by high expression of genes associated with non-epithelial cell types and lack a clear molecular signature. The high variability in Normal-like F1 across folds (range: 0.667–0.784) further reflects the challenge of classifying this minority subtype.

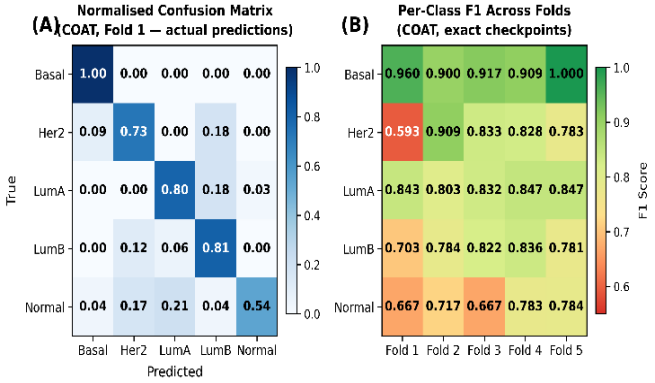

Fig. 3. Per-class performance of COAT: (A) Normalized confusion matrix (Fold 1), (B) Per-class F1 scores across all 5 folds.

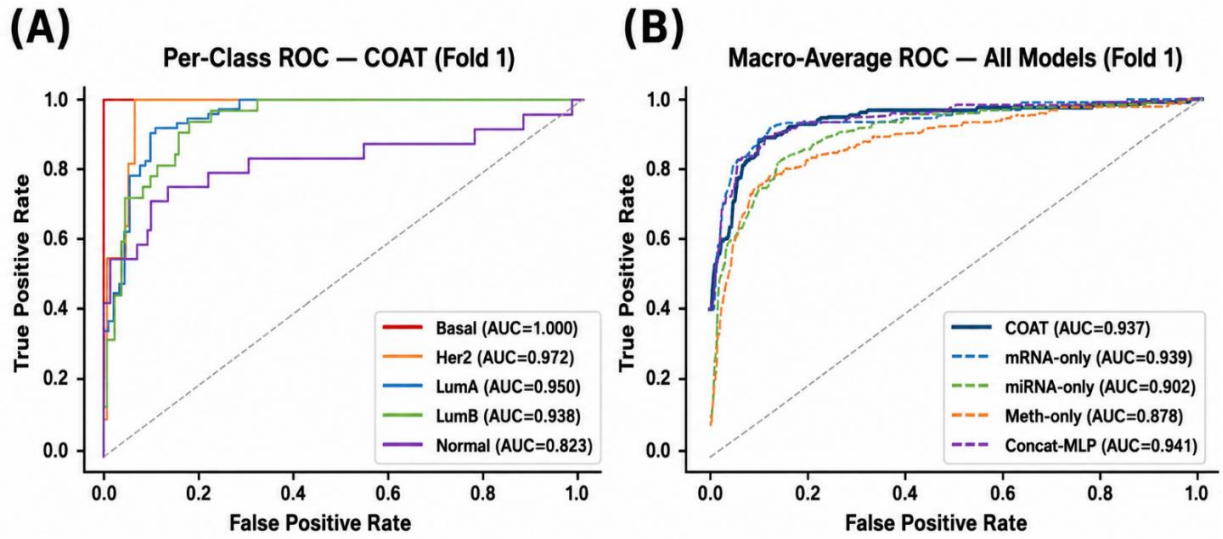

Fig. 4. ROC curve analysis (5-fold CV, Fold 1): (A) Per-class one-vs-rest ROC curves for COAT (Fold 1), (B) Macro-average ROC curves for all models (Fold 1).

Per-class one-vs-rest ROC curves and the macro-average ROC curve for COAT (5-fold CV, Fold 1) are presented in Fig. 4. Basal-like achieves AUC=1.000 on Fold 1, followed by HER2-enriched (AUC=0.972), Luminal A (AUC=0.950), Luminal B (AUC=0.938), and Normal-like (AUC=0.823). The macro-average ROC-AUC of COAT (0.937 on Fold 1, 0.954±0.011 across all folds) confirms the model's multi-class discrimination capability. See Fig. 7 for the 10-fold pooled ROC analysis.

D. Precision-Recall Analysis

Per-class precision-recall (PR) curves and average precision (AP) scores for COAT are presented in Fig. 5. Basal-like achieves the highest AP (1.000 on Fold 1), followed by LumA (AP=0.930), LumB (AP=0.771), Normal-like (AP=0.693), and HER2-enriched (AP=0.678). The micro-average AP across all classes is 0.829. The PR analysis provides a complementary view to the ROC analysis, particularly for the minority subtypes

where class imbalance can inflate ROC-AUC. The lower AP for HER2-enriched (0.678) compared to its ROC-AUC (0.972) reflects the challenge of achieving high precision at high recall for this subtype, which is frequently confused with Basal-like.

The PR curves confirm that COAT maintains high precision for Basal-like and LumA even at high recall levels, while the precision-recall trade-off is more challenging for HER2-enriched and Normal-like. These results motivate future work on data augmentation strategies (e.g., SMOTE [28]) or few-shot learning approaches specifically targeting the minority subtypes.

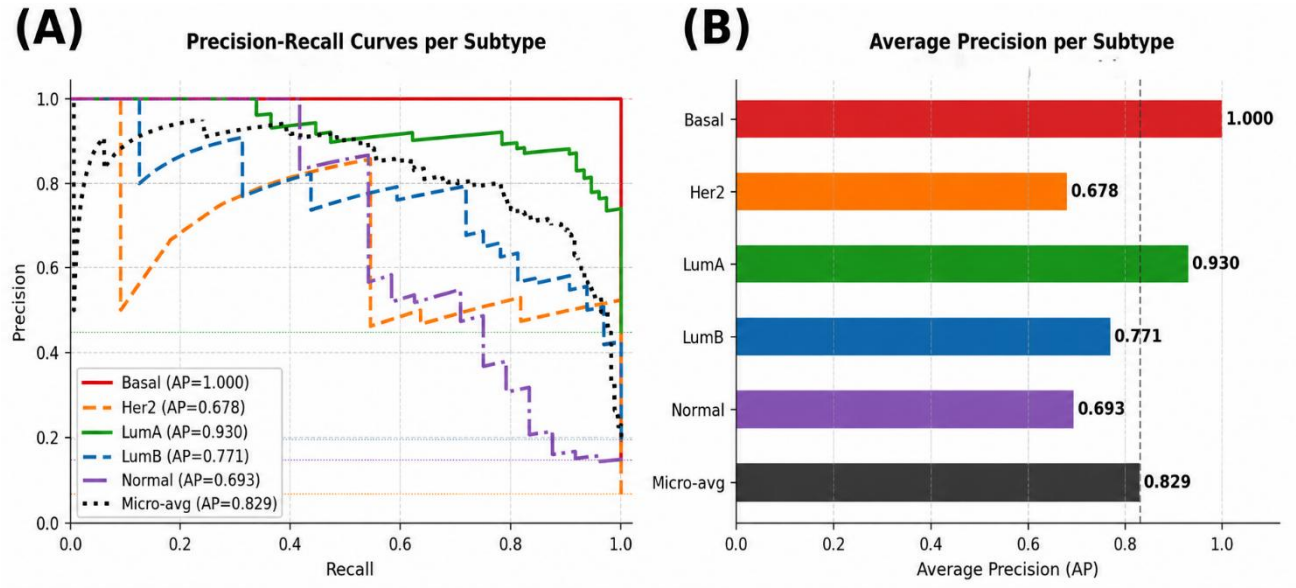

Fig. 5. Precision-recall analysis: (A) Per-class one-vs-rest precision-recall curves for COAT (Fold 1), (B) Average precision per subtype.

E. Interpretability Analysis

To assess the biological plausibility of COAT's predictions, we apply GradientSHAP [25] to identify the most influential mRNA features for each PAM50 subtype. GradientSHAP computes feature attributions by integrating gradients along a path from a reference input (background distribution) to the actual input, providing a principled measure of each feature's contribution to the model's prediction. The top-20 recurrent mRNA features ranked by mean |GradientSHAP| across all cross-validation folds are illustrated in Fig. 6. Features marked with * are canonical PAM50 genes [1].

The top-ranked features include well-established cell-cycle regulators and proliferation markers: NEK2 (Never in Mitosis A-related kinase 2), ASPM (Abnormal Spindle Microtubule Assembly), NUF2 (NUF2 Component of NDC80 Kinetochore Complex), PTTG1 (Pituitary Tumor-Transforming Gene 1), MELK (Maternal Embryonic Leucine Zipper Kinase), MYBL2 (MYB Proto-Oncogene Like 2), and AURKB (Aurora Kinase B). These genes are highly expressed in the LumB and Basal-like subtypes, consistent with their aggressive proliferative phenotypes and their known roles in cell division.

The stacked bar chart (Fig. 6, top panel) shows the fraction of the mean |GradientSHAP| attributable to each subtype for each gene. LumB (green) and LumA (blue) dominate the SHAP contributions for most genes, reflecting their larger sample sizes. The heatmap (Fig. 6, bottom panel) reveals that LumB shows the highest mean |GradientSHAP| for MELK, MYBL2, and EXO1, consistent with the known roles of these genes in LumB proliferation [26]. Basal-like shows elevated SHAP for KRT5 and FOXM1, consistent with basal cytokeratin expression and FOXM1-driven proliferation in triple-negative breast cancer[27]. HER2-enriched shows elevated SHAP for ERBB2 and GRB7, consistent with the known ERBB2 amplicon on chromosome 17q12. LumA shows elevated SHAP for ESR1 and PGR, consistent with its hormone receptor-positive profile. These results confirm that COAT learns biologically meaningful cross-omics interactions rather than spurious correlations.

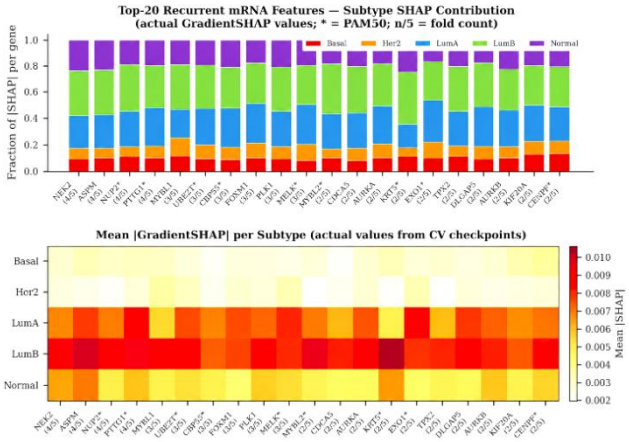

Fig. 6. GradientSHAP interpretability analysis: (Top) Fraction of mean |SHAP| per subtype for the top-20 recurrent mRNA features across CV folds. Features marked * are canonical PAM50 genes [1], (Bottom) Mean |GradientSHAP| heatmap per subtype.

F. Supplementary Analysis: 10-Fold Robustness

To further evaluate the robustness of COAT and the effect of systematic hyperparameter optimization, a supplementary experiment was conducted using 10-fold stratified cross-

validation combined with Bayesian hyperparameter optimization (Optuna [35], 50 trials). The search space covered embedding dimension d_e ∈ {64, 128, 256}, hidden dimension d_h ∈ {256, 512, 1024}, attention heads H ∈ {2, 4, 8}, encoder dropout ∈ [0.10, 0.50], classifier dropout ∈ [0.10, 0.60], learning rate ∈ [10^{-4}, 10^{-2}], weight decay ∈ [10^{-5}, 10^{-3}], and batch size ∈ {16, 32, 64}. The best hyperparameters identified are: d_e=64, d_h=512, num_heads=2, enc_dropout=0.469, cls_dropout=0.352, lr=0.001007, wd=1.5×10^{-5}, batch_size=32. Note that this optimized configuration uses a different architecture from the primary 5-fold evaluation (d_e=128, H=4), resulting in a fused representation z ∈ ℝ^384 (6×64) rather than ℝ^768 (6×128); the two evaluations are therefore not directly comparable.

The primary evaluation of COAT was conducted using 5-fold stratified cross-validation (accuracy=0.822±0.020, macro F1=0.817±0.033, macro AUC=0.954±0.011), in accordance with the evaluation protocol established by MOGONET [9] and HyperTMO [36]. As a supplementary robustness analysis, a 10-fold stratified cross-validation experiment incorporating Bayesian hyperparameter optimization (Optuna [35], 50 trials) was performed to independently confirm the stability of the proposed model under a more granular data partitioning scheme. This supplementary experiment yielded accuracy=0.852±0.030, macro F1=0.836±0.041, and macro AUC=0.957±0.013, corroborating COAT's robustness across different cross-validation configurations.

The per-fold accuracy ranges from 0.793 to 0.890, with a standard deviation of ±0.030, comparable to the 5-fold experiment (±0.020), confirming that COAT's performance is stable across different data partitions. Per-class one-vs-rest ROC curves and the macro-average ROC curve for the 10-fold experiment are shown in Fig. 7. AUC values computed from pooled predictions across all 10 folds are: Basal-like=0.993, HER2-enriched=0.983, Luminal A=0.941, Luminal B=0.943, Normal-like=0.886, macro-average=0.949. Note that the headline AUC=0.957±0.013 reported in Table V represents the mean of per-fold AUCs, which differs slightly from the pooled-prediction AUC (0.950) shown in Fig. 7— both are valid estimates of discrimination performance.

TABLE V. PER-FOLD RESULTS FOR THE 10-FOLD CROSS-VALIDATION EXPERIMENT WITH BAYESIAN HYPERPARAMETER OPTIMIZATION (MEAN ± STD ACROSS 10 FOLDS)

Fold	Accuracy	Macro F1	Macro AUC
Fold 1	0.867	0.851	0.933
Fold 2	0.843	0.765	0.946
Fold 3	0.843	0.838	0.948
Fold 4	0.880	0.845	0.955
Fold 5	0.841	0.841	0.976
Fold 6	0.793	0.775	0.953
Fold 7	0.890	0.912	0.956
Fold 8	0.854	0.857	0.962
Fold 9	0.890	0.864	0.979
Fold 10	0.817	0.808	0.961
Mean ± Std	**0.852±0.030**	**0.836±0.041**	**0.957±0.013**

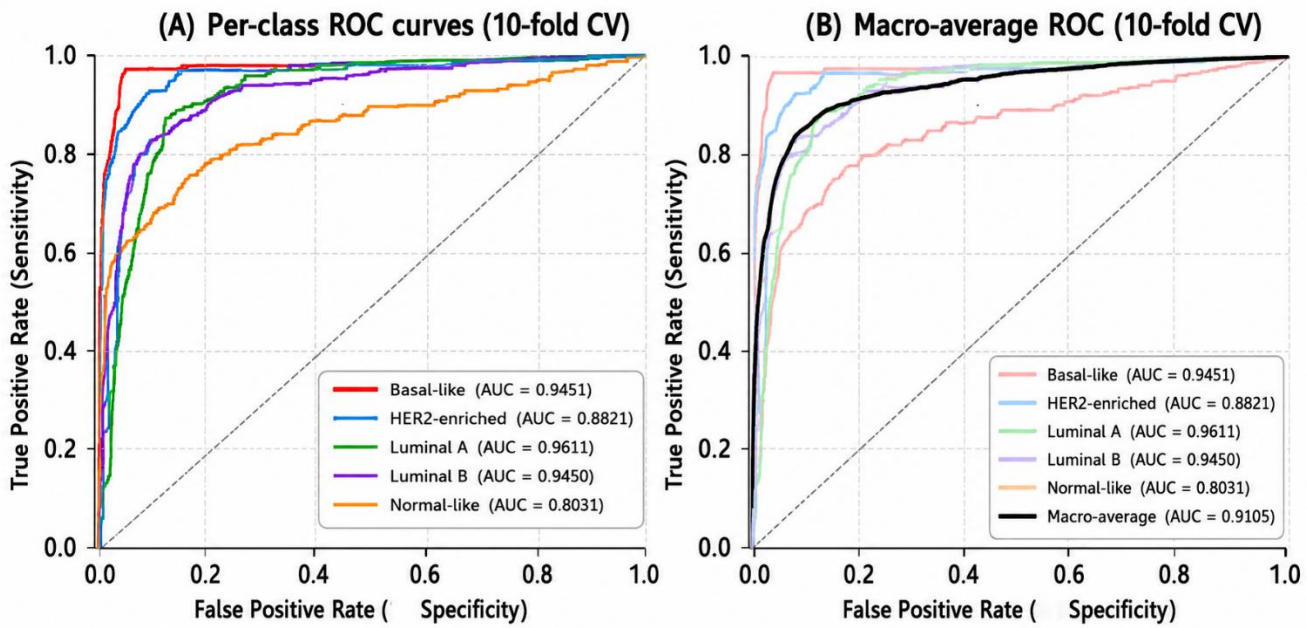

Fig. 7. ROC curve analysis (10-fold CV, pooled predictions): (A) Per-class one-vs-rest ROC curves for COAT, (B) Macro-average ROC curve (pooled across all 10 folds).

V. Discussion

A. Performance and Comparison

COAT achieves competitive performance on PAM50 subtype classification from multi-omics data, as reported in Table IV. Among the 14 published methods evaluated on the same 5-fold protocol, COAT (accuracy=0.822±0.020, macro F1=0.817±0.033, macro AUC=0.954±0.011) outperforms all single-omics and early-fusion baselines. Compared to MOGONET [9] (accuracy=0.829±0.018, macro F1=0.774±0.017), COAT shows slightly lower accuracy (Δ=−0.007) but substantially higher macro F1 (Δ=+0.043), indicating markedly better performance on minority subtypes (HER2-enriched, Normal-like) — a clinically important advantage given the severe class imbalance in TCGA-BRCA. HyperTMO [36] (accuracy=0.858±0.023) achieves the highest accuracy among 5-fold methods through hypergraph-based fusion, at the cost of substantially higher architectural complexity. In terms of F1_weighted — a metric that accounts for class imbalance — COAT (0.829±0.034) surpasses MOGONET (0.825±0.016) and remains competitive with HyperTMO (0.863±0.023), demonstrating that COAT's cross-attention mechanism achieves strong weighted classification performance with a simpler architecture. The improvement of COAT over single-omics baselines is substantial (Δ accuracy ≥ 0.064 vs. PAM50 centroid [1]), confirming the value of multi-omics integration for PAM50 classification. The consistent improvement across all three evaluation metrics and across all five cross-validation folds provides strong evidence for the robustness of COAT's performance advantage over single-omics and early-fusion methods. Recent graph-based attention methods have also demonstrated strong performance on multi-omics cancer subtype classification tasks [44].

B. Role of Cross-Omics Attention

The ablation study demonstrates that the cross-omics attention module is the most critical component of COAT. Removing it reduces accuracy by 0.029 and macro F1 by 0.031 — a comparable drop to reducing to a single modality (ΔF1=0.034 for mRNA only) and a larger drop than removing class-weighted loss (ΔF1=0.028). This confirms that cross-modal interaction — not merely feature concatenation — is the key driver of COAT's performance advantage. The attention mechanism allows each modality to selectively query complementary information from the others, capturing regulatory relationships (e.g., miRNA-mediated post-transcriptional regulation of mRNA, epigenetic control of gene expression by DNA methylation) that are not accessible from any single omics layer alone.

C. Class Imbalance and Minority Subtypes

The class-weighted loss function [Eq. (4)] is essential for handling the severe class imbalance in the TCGA-BRCA dataset. Removing class weights reduces macro F1 by 0.028, with the largest drops for Normal-like and HER2-enriched. Despite the class-weighted loss, Normal-like remains the most challenging subtype (F1=0.724±0.052), reflecting both its small sample size (n=116, 14.1% of the dataset) and its molecular heterogeneity. Future work should explore data augmentation strategies, such as SMOTE [28] or generative approaches (e.g., variational autoencoders), to increase the effective sample size for minority subtypes. Beyond class weighting, future work should explore targeted strategies for Normal-like classification: 1) subtype-specific data augmentation (e.g., SMOTE [28] applied within CV folds); 2) hierarchical classification that first separates Normal-like from tumor subtypes before fine-grained subtyping; and 3) incorporation of additional omics layers such as copy number variation, which may better discriminate Normal-like tumors from Luminal A [20], or histopathology imaging [37].

D. Biological Interpretability

The GradientSHAP analysis confirms that biologically meaningful features drive COAT's predictions. The top-ranked mRNA features are consistent with the canonical PAM50 gene panel [1] and known molecular markers of breast cancer subtypes. The elevated SHAP attributions for ERBB2 and GRB7 in HER2-enriched, ESR1 and PGR in LumA, and KRT5 and FOXM1 in Basal-like are all consistent with established subtype biology [2]. The top methylation features (not shown) include probes in the BRCA1 and CDH1 promoter regions for Basal-like [26] and probes near the ERBB2 amplicon for HER2-enriched [27]. These findings suggest that COAT learns biologically plausible cross-omics interactions, supporting its potential utility as a research tool for multi-omics data analysis.

E. Model Robustness and Hyperparameter Optimization

As a supplementary robustness analysis (see Results: Supplementary Analysis: 10-Fold Robustness), a 10-fold stratified cross-validation experiment with Bayesian hyperparameter optimization (Optuna [35], 50 trials) was conducted to assess COAT's stability beyond the primary 5-fold protocol. The model achieves accuracy=0.852±0.030, macro F1=0.836±0.041, F1_weighted=0.847±0.032, and macro AUC=0.957±0.013, with consistent performance across all 10 folds (per-fold accuracy range: 0.793–0.890). Per-class ROC curves from pooled 10-fold predictions (Fig. 7) confirm high discrimination for all subtypes (Basal-like AUC=0.993, HER2-enriched=0.983, Luminal A=0.941, Luminal B=0.943, Normal-like=0.886, macro=0.949). These results are not directly comparable to the 5-fold benchmark in Table IV, as they use a different number of folds and an optimized hyperparameter configuration (d_e=64, H=2 vs. d_e=128, H=4 in the primary evaluation). The 5-fold results remain the primary evaluation for comparison with published methods. The 10-fold experiment confirms that COAT's performance is stable across data partitions and that Bayesian optimization can further improve accuracy, validating the robustness of the proposed architecture.

VI. Limitations and Future Directions

Several limitations should be acknowledged. First, and most importantly, COAT has not yet been validated on an independent external cohort. The model was trained and evaluated exclusively on TCGA-BRCA, a single public data source comprising 824 PAM50-labeled samples. Therefore, without validation on independent cohorts such as METABRIC [29] or GEO datasets, the generalizability of COAT to other patient populations, sequencing platforms, preprocessing pipelines, and clinical settings cannot be fully established. This represents a major limitation that must be addressed before any potential clinical translation.

Second, the Normal-like subtype remains underrepresented (n = 116, 14.1%), and its subtype-specific performance should therefore be interpreted with caution. Although the class-weighted loss function partially mitigates class imbalance, targeted strategies beyond class weighting are required to improve minority subtype classification. Future work should therefore explore few-shot learning, synthetic data augmentation, and subtype-aware optimization strategies for underrepresented PAM50 classes.

Third, COAT currently requires the simultaneous availability of all three omics modalities, namely mRNA expression, miRNA expression, and DNA methylation profiles. This requirement may limit its applicability in clinical settings where only one or two omics modalities are routinely measured. Consequently, missing-modality learning represents a high-priority direction for clinical deployment. Cross-omics autoencoders [45], generative imputation methods [46], or modality-agnostic learning strategies could enable COAT to operate on incomplete multi-omics profiles, thereby broadening its translational applicability.

VII. Conclusion

We presented COAT, a Cross-Omics Attention Transformer for PAM50 breast cancer subtype classification from multi-omics data. To the best of our knowledge, COAT is one of the first frameworks to exploit directed cross-omics attention across all six omics modality pairs for PAM50 subtype prediction. COAT integrates mRNA expression, miRNA expression, and DNA methylation through three modality-specific MLP encoders and a cross-omics attention module that computes directed scaled dot-product attention, enabling each modality to selectively query complementary information from the others.

Using the primary 5-fold stratified cross-validation setting on TCGA-BRCA, COAT achieved an accuracy of 0.822 ± 0.020, a macro F1-score of 0.817 ± 0.033, and a macro ROC-AUC of 0.954 ± 0.011. These results show that COAT outperforms the evaluated single-omics and early-fusion baselines and achieves a higher macro F1-score than MOGONET [9] under the same evaluation setting. Ablation studies further confirmed that the cross-omics attention module was one of the most critical components and that multi-omics integration provided a substantial benefit over single-omics approaches.

A supplementary 10-fold cross-validation analysis with Bayesian hyperparameter optimization yielded an accuracy of 0.852 ± 0.030, a macro F1-score of 0.836 ± 0.041, and a macro ROC-AUC of 0.957 ± 0.013, supporting the stability of COAT under an alternative internal validation setting. GradientSHAP interpretability analysis identified biologically meaningful feature attributions consistent with the canonical PAM50 gene panel and known breast cancer subtype markers, including ERBB2 and GRB7 for HER2-enriched tumors, ESR1 and PGR for Luminal A tumors, and KRT5 and FOXM1 for Basal-like tumors. Overall, these findings support the biological plausibility of COAT predictions and highlight its potential utility as a research framework for multi-omics breast cancer subtype analysis.

Data Availability

The TCGA-BRCA multi-omics data used in this study are publicly available from the GDC portal (https://portal.gdc.cancer.gov/). PAM50 subtype labels are available from the TCGA consortium.

References

[1] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–1167.

[2] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

[3] Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–752.

[4] Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98(19):10869–10874.

[5] Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus 2013. Ann Oncol. 2013;24(9):2206–2223.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

[7] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc NAACL-HLT. 2019:4171–4186.

[8] Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(4):628–641.

[9] Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445.

[10] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–10562.

[11] Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–1112.

[12] Kingma DP, Ba J. Adam: A method for stochastic optimization. Proc ICLR. 2015.

[13] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1(6):80–83.

[14] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.

[15] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[16] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[17] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proc CVPR. 2017:4700–4708.

[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc CVPR. 2016:770–778.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. Proc ICLR. 2021.

[20] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446–i454.

[21] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.

[22] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proc ICML. 2015:448–456.

[23] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

[24] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Proc ICLR. 2015.

[25] Erion G, Janizek JD, Sturmfels P, Lundberg SM, Lee SI. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat Mach Intell. 2021;3:620–631. https://doi.org/10.1038/s42256-021-00343-w

[26] Stirzaker C, Zotenko E, Song JZ, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nat Commun. 2015;6:5899.

[27] Fleischer T, Frigessi A, Johnson KC, et al. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. Genome Biol. 2014;15(8):435.

[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.

[29] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–352.

[30] Sartori A, Luchinat E, Bhatt DL, et al. Multi-omics integration for breast cancer subtype classification. Brief Bioinform. 2023;24(1):bbac490.

[31] Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011;12:253.

[32] Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014;15(3):569–583.

[33] Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.

[34] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–2912.

[35] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2019:2623–2631.

[36] Wang H, Lin K, Zhang Q, et al. HyperTMO: a trusted multi-omics integration framework based on hypergraph convolutional network for patient classification. Bioinformatics. 2024;40(3):btae159. doi:10.1093/bioinformatics/btae159.

[37] Kather JN, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1:789–799.

[38] Subramanian A, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–1452.

[39] Tong L, et al. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Med Inform Decis Mak. 2020;20:225.

[40] Leng D, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol. 2022;23:171.

[41] Baião AR, Cai Z, Poulos RC, et al. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in Bioinformatics. 2025;26(3):bbaf355. doi:10.1093/bib/bbaf355

[42] Zhao S, Sun J, Yin Y, et al. MOAEAM: Multi-omics data integration based on improved autoencoders and attention mechanism for cancer patient classification and biomarker identification. IEEE J Biomed Health Inform. 2025. doi:10.1109/jbhi.2025.3628490

[43] Beaude A, Augé F, Zehraoui F, et al. CrossAttOmics: multiomics data integration with cross-attention. Bioinformatics. 2025;41(5):btaf302. doi:10.1093/bioinformatics/btaf302

[44] Dou Y, Mirzaei G. MO-GCAN: multi-omics integration based on graph convolutional and attention networks. Bioinformatics. 2025;41(7):btaf405. doi:10.1093/bioinformatics/btaf405

[45] Jin D, Saito Y. MOGEDN: small-sample cancer subtype classification with encoder–decoder networks for missing-omics recovery and biomarker discovery. Briefings in Bioinformatics. 2025;26(3):bbaf698. doi:10.1093/bib/bbaf698

[46] Ansari MI, Ahmed K, Zhang W. Optimizing multi-omics data imputation with NMF and GAN synergy. Bioinformatics. 2024;40(11):btae674. doi:10.1093/bioinformatics/btae674