BioLM-NET: an interpretable deep learning model combining prior biological knowledge and contextual LLM gene embeddings on multi-omics data to predict disease

Jubair Ibn Malik Rifat, Thasina Tabashum, Md Marufi Rahman, Md Farhad Mokter, Sarthak Engala

Department of computer Science & Engineering, University of North Texas, Denton, Texas, 76203, USA

Email: jubairibnmalikrifat@my.unt.edu; thasinatabashum@my.unt.edu; mdmarufirahman@my.unt.edu; mdfarhadmokter@my.unt.edu; sarthakengala@my.unt.edu

Serdar Bozdag

Department of Computer Science & Engineering, Department of Mathematics, Center for Computational Life Sciences, University of North Texas, Denton, Texas, 76203, USA

Email: serdar.bozdag@.unt.edu

Biologically informed deep neural networks, which connect input layer to hidden layers based on gene-pathway relationship have gained popularity in recent years. However, most existing methods do not incorporate protein-protein interactions (PPI) and protein-DNA interactions (PDI) in their designs. In this study, we introduce BioLM-NET, a deep learning-based framework that fuses single cell or bulk gene expression data and DNA methylation data with prior biological knowledge including Protein-Protein Interactions (PPI), Protein-DNA Interactions (PDI). BioLM-NET also aggregates latent representation of omics signals at pathway-level through an attention-based pathway layer where a pre-trained large language model (LLM) was incorporated to generate context-specific gene embeddings. We evaluated BioLM-NET on single cell colorectal cancer data from scTrioseq2 platform to predict primary and metastatic cancer cells, on TCGA-BRCA, TCGA-GBM, TCGA-COAD to predict cancer subtypes and ROSMAP data to predict Alzheimer's disease patient. Our results showed that BioLM-NET outperformed baseline and state-of-the-art (SOTA) methods, P-NET and PASNet with statistical significance on scTrioseq2 data, TCGA-COAD and ROSMAP data and ties with SVM and Dense neural network on TCGA-BRCA data. Our ablation studies demonstrated the importance of incorporating PPI, PDI data and attention-based pathway layer. We also interpret our models and found out that our important input features are significantly enriched in GO terms and KEGG pathways and can serve as potential biomarkers or therapeutic targets for the corresponding disease.

Keywords: Biologically Informed Neural Network, Large Language Model, scTrioseq2, Alzheimer's disease.

1. Introduction

The integration of multi-omics data—including genomics, transcriptomics, epigenomics, and proteomics—holds significant promise for enhancing the accuracy of disease prediction and classification. Several machine learning (ML) and deep learning (DL) models have been developed to integrate multi-omics datasets. However, most existing approaches lack interpretability reducing their practical usability.

To address this issue, recently biologically informed deep neural networks have been developed. For instance, Elmarakeby et al. introduced P-NET, a biologically informed deep neural network initially developed to classify prostate cancer patients into treatment-resistant or primary groups[1]. P-NET demonstrated not only superior prediction performance compared to conventional methods but also provided critical insights into novel therapeutic targets through its interpretable architecture. Although the model is easy to interpret, it relies on existing pathway databases, thereby might miss new or unknown interactions not yet captured in those resources. Other biologically informed neural network architectures such as DrugCell[2] and DCell[3] have demonstrated capability in modeling biological processes and predicting therapeutic responses. However, because DCell was validated in yeast, its direct translation to human diseases may fail to capture the complexity of human biology.

Fortelny and Bock et al. proposed Knowledge-Primed Neural Networks (KPNNs) by incorporating biological network priors into models analyzing single-cell RNA-seq data[4]. While KPNNs showed good interpretability and accuracy, they did not utilize multi-omics data and depended heavily on the quality of the input biological networks. Wang et al. developed BioNet for analyzing tumor heterogeneity from imaging data, but its accuracy may vary with image quality and clinical settings[5]. Another interpretable model, PASNet incorporated sparse coding and pathway-level representations to predict prognosis from high-throughput gene expression data[6]. While PASNet demonstrated robust performance and clear interpretability through pathway hierarchies, it did not utilize multi-omics inputs and was tested only on bulk transcriptomic data from glioblastoma.

Most of the current models use either pathway or Protein-Protein Interaction (PPI) and Protein-DNA Interaction (PDI) networks but none combine all three. These models may be interpretable but are not context-specific, and most of them work on single omics type from bulk tissue rather than single-cell and multi-omics data.

To address these gaps, we introduce BioLM-NET, a novel PPI, PDI, and pathway aware neural network framework that fuses single-cell gene expression, DNA methylation, and pretrained Large Language Model (LLM) based gene embeddings. We conducted comprehensive benchmarking across five datasets, namely scTrio-seq2 data for colorectal cancer[7], TCGA-BRCA[8] for breast cancer, TCGA-GBM[9] for glioblastoma, TCGA-COAD[10] for colon cancer, and ROSMAP[11] for Alzheimer's disease (AD) and achieved state-of-the-art (SOTA) performance in most cases. Our results demonstrated enhanced biological interpretability showing how pathway-level attention and LLM-derived embeddings uncover key molecular features of cancer progression and metastasis, cancer subtyping, and enable stratification of AD patients. The source code and datasets for BioLM-NET is available at

https://github.com/bozdaglab/BioLM-NET under Creative Commons Attribution-Noncommercial 4.0 International Public License.

2. Materials and Methods

2.1 *Dataset*

In this study, we utilized gene expression and DNA methylation data from five datasets: Single Cell Trio-seq2 (scTrio-seq2), TCGA-BRCA, TCGA-GBM, TCGA-COAD, and ROSMAP. We downloaded scTrio-seq2 data from the Gene Expression Omnibus (GEO) under accession number GSE97693[7]. Processed TCGA-BRCA and ROSMAP datasets were obtained from Wang et al.[12], while processed TCGA-GBM and TCGA-COAD datasets were obtained from Yang et al.[13]. We utilized the top 500 features from both TCGA-GBM and TCGA-COAD datasets. Biological prior information was obtained from three sources: protein-protein interactions (PPI) from STRING[14] database, protein-DNA interactions (PDI) from the DoRothEA[15] database, and pathway information from KEGG database[16,17]. We also utilized LLM embedding for genes from GenePT [18]. A summary of the five datasets is provided in Table 1.

Table 1: Overview of the five datasets:. GE: Gene expression, DM: DNA methylation, PPI: Protein-protein interaction, PDI: Protein-DNA interaction

Dataset	# samples	# features (GE)	# features (DM)	# PDI (GE)	# PDI (DM)	# PPI (GE)	# PPI (DM)	# Pathways (GE)	# Pathways (DM)
scTrioseq2	609	1,673	1,712	2,459	2,293	21,968	6,690	161	173
BRCA	875	999	999	1,922	1,704	6,280	3,088	133	142
GBM	244	500	500	682	602	2,276	2,358	140	102
COAD	256	500	500	866	874	2,796	2,182	103	97
ROSMAP	351	200	200	497	380	432	620	134	134

2.2 *scTrio-seq2 gene expression data preprocessing*

The gene expression data had two units: FPKM (510 samples, 25,373 genes) and TPM (688 samples, 23,457 genes). We converted all values to TPM to standardize the data. We only kept genes that are common in both units. The final dataset contained 1,198 samples across five classes: Primary Tumor (PT, n=616), Lymph Node Metastasis (LN, n=224), Liver Metastasis (ML, n=226), Post-Treatment Liver Metastasis (MP, n=102), and Normal Colon (NC, n=20). We excluded NC from further analysis to mitigate class imbalance due to its small sample size. We performed differential expression (DE) with MAST[19] (adj. $p < 0.05$, log2FC > 2) and identified 1,673 unique genes across four comparisons. DE analysis details and figures are provided in Supplemental Note 1 and Supplemental Figures 1-2.

2.3 *scTrio-seq2 DNA methylation data preprocessing*

Single-base DNA methylation data were aggregated to the gene-level by averaging promoter signals. From 1,295 samples and 22,910 genes, we retained 17,123 after filtering and selected the top 10% highly variable genes. The details of DNA methylation preprocessing are provided in Supplemental Note 1.

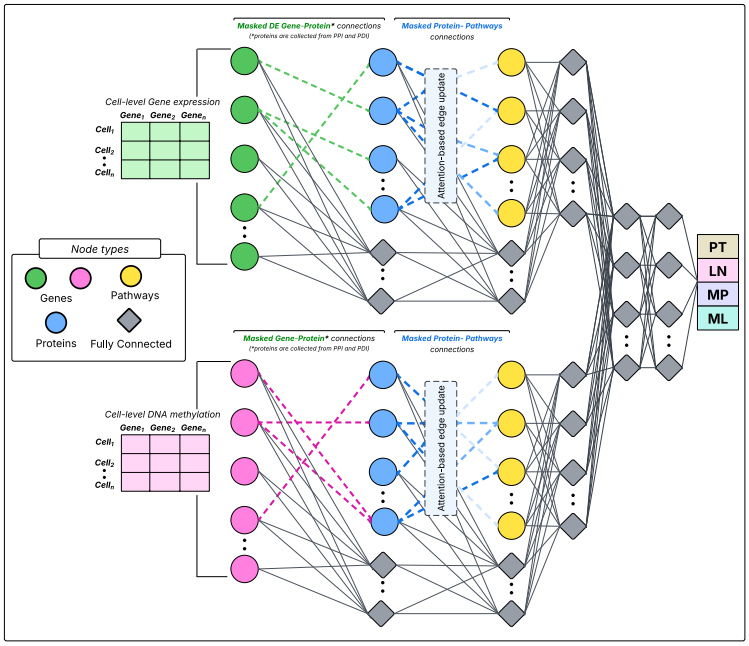

Fig. 1. The architecture of BioLM-NET. (Here, model outputs are four classes: PT = Primary Tumor, LN = Lymph Node Metastasis, MP= Post-Treatment Liver Metastasis, ML= Liver Metastasis)

2.4 *Protein–DNA interaction (PDI) data*

PDI data were obtained from the DoRothEA[15] database. It categorizes TF–target interactions into five confidence levels: A (highest) to E (lowest). For this study, we used only levels A, B, and C. The raw dataset included 32,443 interactions between 430 unique TFs and 9,290 unique target genes. We filtered out target genes that were not DE (gene expression) or not highly variable (DNA methylation)..

2.5 *Protein–protein interaction (PPI) data*

Human PPI data were downloaded from the STRING[14] database. Each PPI interaction has a *combined score* derived from multiple evidence sources. We selected interactions with a combined score > 0.7 to represent high-confidence associations. This filtering resulted in 504,062 high-confidence PPIs. Next, we retained the top 10% of interactions by score to preserve the highest-confidence associations. We further filtered interactions so that at least one of the proteins corresponded to a differentially expressed gene (gene expression data branch) or highly variable gene (DNA methylation data branch).

2.6 *Pathway data*

To compute pathways associated with the genes in PDI and PPI, we utilized the R package clusterProfiler[20] to perform pathway enrichment and kept only the significantly enriched pathways (Benjamini–Hochberg adjusted p-value < 0.05).

2.7 *LLM embedding*

GenePT[18] is an LLM framework that uses OpenAI's text embedding APIs to convert gene-centric descriptions into embeddings. These embeddings capture gene specific context such as aging, drug interactions, and pathway membership. We extracted GenePT embeddings from Hugging Face and kept only embeddings for genes with KEGG pathways identified in our analysis. These gene embeddings were then used to implement an attention mechanism between genes and pathways that allows the model to assign greater importance to genes with higher attention scores so that it enhances model performance and interpretability.

2.8 *Framework architecture*

Our framework uses a dual branch design that processed gene expression and DNA methylation data in parallel. In each branch, PDI and PPI priors are connected via a custom masked dense layer, and gene level activations are aggregated into pathway level embeddings by a custom attention pathway layer. The two pathway embeddings are then concatenated and passed through an MLP to generate the final prediction. The framework architecture is shown in Figure 1.

2.8.1. *Biological mask matrices*

We constructed two biologically informed mask matrices and then merged them. First, the PDI mask matrix, $M_{PDI} \in \{0,1\}^{n \times m}$ (Eq. 1) captures transcriptional regulation, where n is the number of DE genes or highly variable genes and m is the number of unique target genes.

$$M_{PDI}[i,j] = \begin{cases} 1 & if\ gene\ i\ is\ a\ TF\ regulating\ gene\ j \\ 0 & otherwise, \end{cases} \quad (1)$$

Second, the PPI Mask matrix, $M_{PPI} \in \mathbb{R}^{p_1 \times p_2}$ (Eq. 2) encodes interaction strengths between proteins, where p_1 is the number of DE genes or highly variable genes and p_2 is the number of unique proteins.

$$M_{PPI}[i,j] = \begin{cases} \text{combined score}\ between\ protein\ i\ and\ j \\ 0 & otherwise, \end{cases} \quad (2)$$

Finally, we horizontally concatenated those matrices to get a PDI-PPI mask matrix (Eq. 3).

$$M_{PDI-PPI} = [M_{PDI} | M_{PPI}] \quad (3)$$

2.8.2. *Pathway mask matrix*

The pathway mask matrix, $M_{Path} \in \{0,1\}^{r \times s}$ (Eq. 4) is a binary matrix, similar to the M_{PDI}. Here, r is size of the union between target genes from PDI and proteins from PPI data and s is the number of unique KEGG pathways.

$$M_{Path}[i,q] = \begin{cases} 1 & if\ gene\ i\ is\ annotated\ to\ pathway\ q \\ 0 & otherwise, \end{cases} \quad (4)$$

2.8.3. *Masked layer*

A custom masked layer was implemented to pass information from genes to their corresponding genes in the hidden layer based on PDI and PPI. This layer operates by performing element-wise multiplication of the mask matrix with the weight matrix, $W \in \mathbb{R}^{d_{in} \times d_{out}}$ of the dense layer. (Eq. 5)

$$W_{masked} = W \odot M \quad (5)$$

2.8.4. *Attention-based pathway layer*

From PDI-PPI genes, we extracted KEGG pathways and organized them into a pathway layer. Each node i in this layer represents one pathway and is associated with a trainable query vector $w_q^{(i)} \in \mathbb{R}^d$. We stack these k vectors into a matrix $W_q \in \mathbb{R}^{k \times d}$. To pass information from genes to pathways, we employ an attention mechanism. To calculate attention, we begin with fixed, pretrained GenePT embeddings $G \in \mathbb{R}^{r \times d}$ and the previous layer's output $X^{PDI-PPI} \in \mathbb{R}^{b \times r}$, where b is batch size. Rather than linking every gene equally to its pathways, we compute raw attention scores using Eq. 6. After masking out gene-pathway links (Eq. 7), we compute attention weights using Eq. 8. Finally, we aggregate gene signals per pathway using Eq. 9.

$$S = W_q G^T \in \mathbb{R}^{k \times r} \quad (6)$$

$$S_{mask} = S^T \odot M_{path} \quad (7)$$

$$\alpha_{i,k} = \frac{\exp(S_{mask}, i, k)}{\sum_{i'=1}^{r} \exp(S_{mask}, i', k)} \in \mathbb{R}^{r \times k} \quad (8)$$

$$Z_{path} = X^{PDI-PPI} \alpha \in \mathbb{R}^{b \times k} \quad (9)$$

This step gives us pathway-level representations where each pathway combines information from related genes based on learned attention scores. Then we pass pathway-level representation to a dense layer named projection layer to reduce the high dimensional pathway representations. We apply this for both gene expression and DNA methylation branches, obtaining Z_{GE} and Z_{DNA}, respectively.

2.8.5. *Fusion layer*

We fuse the two pathway outputs and pass them through two fully connected layers before the final output (Eq. 10-13).

$$Z_{fused} = [Z_{GE} \,||\, Z_{DNA}] \in \mathbb{R}^{b\times 2k} \tag{10}$$

$$H^1 = \sigma\left(Z_{fused}W^1 + \beta^1\right) \in \mathbb{R}^{b\times d_1} \tag{11}$$

$$H^2 = \sigma(H^1W^2 + \beta^2) \in \mathbb{R}^{b\times d_2} \tag{12}$$

$$\hat{y} = Softmax(H^2W^{out} + \beta^{out}) \in \mathbb{R}^{b\times c} \tag{13}$$

In Eq. 10-13, σ is the activation function, d_1, d_2 are hidden layer dimensions, W^1, W^2, W^{out} are weights, $\beta^1, \beta^2, \beta^{out}$ are biases, and c is the number of classes.

2.8.6. *Customized loss function*

To mitigate class imbalance, we used a customized sparse categorical cross-entropy loss by assigning higher weights to underrepresented classes. The loss for a sample i with true label y_i and predicted probability $\hat{y}_i$ is defined in Eq. 14

$$L_i = -w_i.\log(\hat{y}_i) \tag{14}$$

where w_i is the class weight. The total loss over a batch of N samples is defined in Eq. 15

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{15}$$

Class weights are computed as $w_c = \frac{n}{C.n_c}$, where n_c is the number of samples in class c, C is the number of classes, and n is the total number of samples.

3. Results

To evaluate BioLM-NET, we tested it using five datasets: scTrioseq2, TCGA-BRCA, TCGA-GBM, TCGA-COAD, and ROSMAP and compared with other tools. The dataset splitting and model evaluation procedures are described in Supplemental Note 2. We also interpreted the model outputs and performed an ablation study to identify the key components of BioLM-NET.

3.1. *Comparing BioLM-NET with baseline and SOTA models*

We compared BioLM-NET to three traditional machine learning classifiers (i.e., Support Vector Machine (SVM), Random Forest (RF), and XGBoost), a fully dense neural network, and two SOTA methods (i.e., P-NET and PASNet). Since SVM, RF, and XGBoost cannot directly incorporate prior

pathway knowledge or LLM derived embeddings, we concatenated the gene expression and DNA methylation data to train these models. For the Dense NN, we preserved BioLM-NET's overall architecture but replaced its sparse connections with dense layers and excluded the attention-pathway module. For both P-NET and PASNet, we concatenated gene expression and DNA methylation data into a single input. P-NET dynamically selects relevant pathways during training, while PASNet relies on user-provided pathways.

Figure 2a summarizes F1 score performance for seven methods across five datasets. BioLM-NET attained the highest F1 score on scTrioSeq2, COAD and ROSMAP, and was on par with Dense NN and SVM on BRCA (0.82). Wilcoxon rank-sum tests confirmed that these improvements over both baseline and state-of-the-art models were statistically significant. Results for precision, recall, and accuracy are provided in Supplemental Figures 3–5. BioLM-NET achieved the highest precision on scTrioSeq2 (0.95), and ROSMAP (0.83) datasets and second highest precision on BRCA dataset. For recall, BioLM-NET attained the top values on scTrioSeq2 (0.96), COAD (0.80) and ROSMAP (0.83) and second highest recall on BRCA dataset. BioLM-NET consistently achieved the highest accuracy on scTrioSeq2 (0.95), COAD (0.87) and ROSMAP (0.83). BioLM-NET, along with other deep learning methods performed comparatively lower on GBM dataset due to significant class imbalance (30.3%, 25.0%, 18.8%, 17.6%, and 8.3%) and limited sample size (195 training and 49 testing). In these settings, traditional machine learning models often outperform deep learning models.

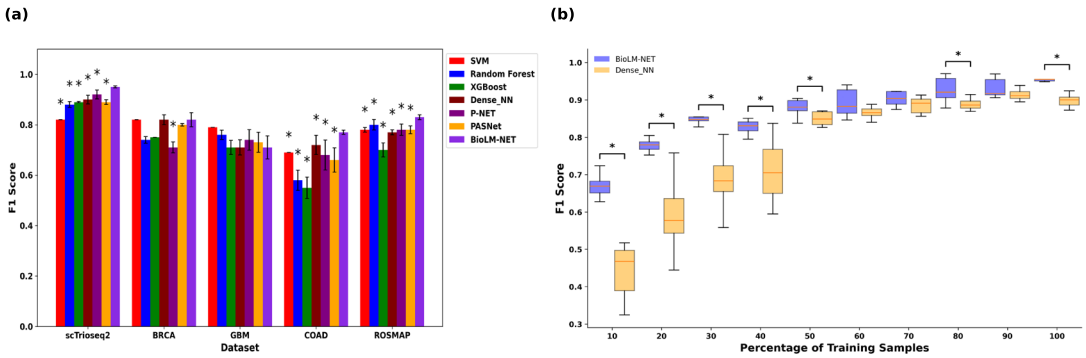

Fig. 2. (a) F1 score comparison of BioLM-NET with baseline and SOTA models. (b) Boxplots of F1 score for BioLM-NET (blue) and Dense_NN (orange) across 10 independent runs at each percentage of training samples. Wilcoxon rank sum test was performed by comparing BioLM-NET with other models (p-value < 0.05 is denoted by *)

3.2. *Predictive performance of BioLM-NET under limited data*

To assess robustness with limited training data, we compared BioLM-NET against a fully dense neural network across smaller training sets. We first split our data into training (80%) and test (20%). We kept the test set constant while subsampling the training data at 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of its original size. We tested each model trained using a subsample training data on the same test set. We observed that the performance of models improved with more training data. BioLM-

NET outperformed the Dense NN at every fraction (Figure 2b), which shows that it remains robust even when trained on substantially fewer samples. We performed a similar analysis using ROSMAP data (Supplemental Figure 6). In this case, Dense NN outperformed BioLM-NET when using 10–50% of the training data; however, BioLM-NET surpassed Dense NN beyond 50%. This result is possibly due to the limited sample size of the ROSMAP dataset, where BioLM-NET requires more data to learn effectively.

3.3. *Ablation study*

To assess the performance of each data modality with or without sparse connections, we conducted an ablation study using scTrioseq2 dataset. We trained a dense neural network using only gene expression data, which had a F1 score of 0.89 (Table 2). When we incorporated sparse connections based on PDI, the F1 score increased by 3%, whereas adding PPI caused a 1% drop (Table 2). When both PDI and PPI were utilized for sparse connections, the F1 score fell by 5%, which could be due to the extra sparsity introduced by combining PPI with PDI connections.

Table 2. Effect of adding Prior biological knowledge and multi omics data in scTrioseq2 data. GE: Gene Expression data, DM: DNA Methylation data, PPI: Protein-Protein Interaction data, PDI: Protein-DNA Interaction data. The result values are shown as mean ± standard deviation of 10 individual runs.

Model	Precision	Recall	F1 score	Accuracy
GE_Dense	0.90 ± 0.0135	0.90 ± 0.0116	0.89 ± 0.0140	0.89 ± 0.0136
GE_PDI	0.93 ± 0.0113	0.92 ± 0.0104	0.92 ± 0.0105	0.92 ± 0.0096
GE_PPI	0.89 ± 0.0658	0.89 ± 0.0600	0.88 ± 0.0792	0.88 ± 0.0647
GE_PDI_PPI	0.86 ± 0.0896	0.86 ± 0.0605	0.84 ± 0.0783	0.84 ± 0.0580
DM_Dense	0.92 ± 0.0043	0.92 ± 0.002	0.92 ± 0.0035	0.92 ± 0.0040
DM_PDI	0.93 ± 0.0087	0.92 ± 0.0098	0.92 ± 0.0095	0.93 ± 0.0075
DM_PPI	0.91 ± 0.0059	0.91 ± 0.0089	0.91 ± 0.0079	0.92 ± 0.0074
DM_PDI_PPI	0.93 ± 0.0161	0.93 ± 0.0243	0.92 ± 0.0206	0.94 ± 0.0151
GE_DM_Dense	0.92 ± 0.0428	0.89 ± 0.0997	0.88 ± 0.0997	0.89 ± 0.0761
GE_DM_PDI	0.91 ± 0.0134	0.91 ± 0.0096	0.90 ± 0.0124	0.90 ± 0.0120
GE_DM_PPI	0.92 ± 0.0240	0.91 ± 0.0316	0.91 ± 0.0319	0.92 ± 0.0267
GE_DM_PDI_PPI	0.94 ± 0.0153	0.94 ± 0.0113	0.94 ± 0.0126	0.94 ± 0.0111
BioLM-NET	**0.95 ± 0.0062**	**0.96 ± 0.0022**	**0.95 ± 0.0039**	**0.95 ± 0.0039**

We observed that a dense neural network model trained with only DNA methylation data achieved a higher performance (0.92) than the model trained using only gene expression data (Table 2) indicating that DNA methylation data plays a stronger role for distinguishing primary tumors from metastatic sites. Adding sparse connections based on PPI to the DNA methylation-based model reduced the F1 score by

1%, whereas we did not observe any change in the performance when PDI data was utilized with or without PPI.

A dense neural network trained using both gene expression and DNA methylation achieved a slightly lower performance (0.88) than either single-modality model. However, adding sparse connections boosted its performance. Incorporating PDI and PPI individually increased the F1 score by 2% and 3%, respectively. When both PDI and PPI were integrated, the F1 score increased by 6%. The best performance was achieved with BioLM-NET, which adds a pathway attention layer to this framework.

We also evaluated the impact of the attention layer in BioLM-NET. When attention scores were removed or replaced with scores from random embeddings, the F1 score dropped by 2% (Table 3) indicating the importance of attention scores. These results collectively confirm that genuine pathway structure and LLM-derived embeddings both contribute to classifying primary tumors and metastatic regions.

Table 3. Effect of adding Pathway Attention Layer in scTrioseq2 data

Model	Precision	Recall	F1 score	Accuracy
BioLM-NET (Without LLM Embedding)	0.94 ± 0.0079	0.93 ± 0.0103	0.93 ± 0.0082	0.94 ± 0.0079
BioLM-NET (Random LLM embedding)	0.93 ± 0.0168	0.94 ± 0.0250	0.93 ± 0.0215	0.93 ± 0.0204
BioLM-NET	**0.95 ± 0.0062**	**0.96 ± 0.0022**	**0.95 ± 0.0039**	**0.95 ± 0.0039**

To assess the biologically informed connections through PDI, PPI, and pathway associations, for each model with sparse connection, we trained a model with randomized connections while preserving the number of sparse connections (Supplemental Table 1).We observed that GE_PDI, DM_PDI, DM_PPI, GE_DM_PPI models outperformed their random counterparts by 5%, 2%, 6%, and 15% in F1 score, respectively which shows the impact of authentic PDI and PPI connections. In GE_PPI model, performance was unchanged compared to randomized version, suggesting that PPI alone adds no additional signal in this context. However, the random version of GE_DM_PDI model outperformed the true model by 1% underscoring the need of more curated TF-target interactions. When all PDI and PPI connections were randomized while leaving other component of BioLM-NET unchanged, performance dropped by 40% in F1 score relative to BioLM-NET. Randomizing pathway connections and LLM embeddings (while preserving shape) further reduced performance by 7%. Altogether, this ablation study demonstrates that incorporating authentic PPI, PDI, and pathway priors, and LLM embedding is essential for BioLM-NET to achieve superior predictive performance.

3.4. *Interpretability of BioLM-NET*

To understand how multimodal features, genes, and pathways contribute BioLM-NET's performance, we first applied SHAP analysis to rank the top 50 most important input features from gene expression and DNA methylation data. We then performed Gene Ontology (GO) and KEGG pathway enrichment to evaluate their biological relevance.

In the ROSMAP dataset, we identified 32 significantly enriched GO terms from the top 50 SHAP-ranked genes in the gene expression data (Supplemental Table 2). Focusing on the ten most significant terms and taking their associated genes from our gene list, we highlight eight candidates: *KIF1C, NPNT, CCDC69, GPER1, KIF5A, KIF5B, FOXO4, MAP4K4.* Genetic fine-mapping implicates *KIF1C* as an AD risk locus and highlights a role for faulty kinesin mediated vesicular transport in disease susceptibility across ancestries[21]. *NPNT* is consistently upregulated in AD brain transcriptomes across diverse populations[22]. *CCDC69* has emerged as a novel genetic locus associated with AD, pointing to unexpected roles for spindle-assembly proteins in neurodegeneration[23]. Elevated *GPER1* expression correlates with increased tau tangle burden, implicating estrogen-receptor signaling in the modulation of tau pathology[24]. In AD, amyloid-β damages *KIF5A* so mitochondria cannot travel down axons, causing energy shortages that lead to synapse loss[25]. *KIF5B* regulates tau protein dynamics, and its dysregulation may drive tau pathology in AD[26]. As a stress response transcription factor, *FOXO4* governs oxidative-stress and insulin-signaling pathways that are disrupted in AD[27]; and upregulation of *MAP4K4* in Alzheimer's microglia drives neuroinflammatory signaling, marking it as a promising therapeutic target to modulate inflammation in AD[28].

Table 4. Significant survival related genes from top 50 SHAP features in GBM, BRCA, and COAD

Dataset	Branch of BioLM-NET	Significant survival related genes
GBM	Gene Expression	*CBX7*[32], *CENPJ*[33], *DACH1*[34], *NUDT22, PM20D2, RALGAPA1P1, TBX3, WDPCP*
	DNA Methylation	CPS1[35], *C16orf89, CANT1, CDKN2D, ELSPBP1, GON4L, LDLRAP1*
BRCA	Gene Expression	*BCL2*[36], *CXCL1*[37], *EIF2S2*[38], *FGD3*[39], *KDM4B*[40], *NXNL2*[41], *SOX11*[42], *SPARCL1*[43], *SUSD3*[44]
	DNA Methylation	*IGFALS, IGFBP4*, *SCUBE2*[45], *SERPINA12*
COAD	Gene Expression	*ASPHD2, COX11, NMNAT1*[46]
	DNA Methylation	*ARRDC1, CAMTA1, CDC14A, DNAJC17, FABP4*[47], *FGF22*

In the ROSMAP dataset, the top 50 DNA methylation probe-associated genes were enriched in the KEGG pathway hsa00010 (Glycolysis/Gluconeogenesis). Recent evidence indicates that dysregulated glycolytic metabolism plays a critical role in AD development[29]. Qiu et al. further demonstrated that the glycolysis index is markedly lower in AD patients than in controls across four distinct brain regions[30]. Among our top DNA methylation features were three hsa00010 genes: *GAPDHS, LDHC*, and *ALDH3B1*. Consistent with these findings, Wang et al. identified *GAPDHS* via SHAP-based feature

selection and reported its significant differential methylation in an AD prediction model, and also highlighted *LDHC* among eight glycolysis genes with altered methylation profiles in AD versus control subjects[31]. In our DNA methylation dataset, two-sample t-tests comparing AD and control groups yielded significant differences for *GAPDHS* ($t = -2.995$, $p = 2.945 \times 10^{-3}$), *LDHC* ($t = 3.889$, $p = 1.212 \times 10^{-4}$), and *ALDH3B1* ($t = 3.122$, $p = 1.948 \times 10^{-3}$). Density plots of these distributions are presented in Supplemental Figure 7.

From our top 50 SHAP features derived from DNA methylation data in the GBM dataset, we identified three significantly enriched GO terms: GO:0009112 (nucleobase metabolic process), GO:0006206 (pyrimidine nucleobase metabolic process), and GO:0046112 (nucleobase biosynthetic process). Jiang et al. reported that GBM relies on nucleotide metabolism to fuel rapid growth and invasion that leads to poor clinical outcomes[48]. Wang et al. demonstrated that pyrimidine metabolic pathways are consistently upregulated in TCGA-GBM tumors relative to normal brain tissue[49]. Zhou et al. showed that elevated activity of nucleobase biosynthetic enzymes correlates with poor prognosis in GBM[50]. We performed Kaplan–Meier survival analysis in GBM, BRCA, and COAD datasets using the top 50 SHAP-ranked genes from both gene expression and DNA methylation branches. Several genes showed significant associations with patient survival, and their relevance is supported by multiple studies. The gene lists are provided in Table 4, with survival plots shown in Supplemental Figures 8–13. These findings highlight BioLM-NET's ability to uncover biologically meaningful features with potential as biomarkers and therapeutic targets in cancer and AD.

To interpret BioLM-NET using an alternative approach, we applied LIME[51], a model agnostic explainability framework, to identify the most important features contributing to our predictions and asses potential feature biases in comparison with SHAP. We applied LIME on the ROSMAP and GBM datasets and selected the top 50 important features from both the gene expression and DNA methylation branches. We then examined the overlap between LIME and SHAP derived features using a one-sided hypergeometric test. In ROSMAP, we identified 47 common genes ($p = 2.384 \times 10^{-38}$) and 43 common DNA methylation features ($p = 9.565 \times 10^{-30}$). In TCGA-GBM, we found 24 common genes ($p = 6.409 \times 10^{-14}$) and 21 common DNA methylation features ($p = 1.261 \times 10^{-10}$). GO and KEGG analysis of these features further confirmed their disease relevance (Supplemental Note 3).

To assess the key nodes in BioLM-NET layers, we examined per sample activation scores at every layer of BioLM-NET trained on the scTrio-seq2 data for colorectal cancer and selected the ten nodes with the highest mean activation score (Supplemental Figure 14). Early in the network, the class specific density curves overlap almost completely, indicating minimal separability. As data propagate through the fusion layer, the distributions begin to separate, and by the final (output) layer, each class's activation density is distinctly isolated. This illustrates how the model transforms mixed features into clear, class-specific patterns.

Based on the activation scores of the attention-based pathway layer, we selected the top 20 pathways and evaluated their relevance to colorectal cancer and their metastasis. In the gene-expression branch, five KEGG pathways were significant (adjusted p-value <0.05): hsa04151 (PI3K–Akt signaling

pathway), hsa04140 (Autophagy), hsa04144 (Endocytosis), hsa03050 (Proteasome), and hsa00190 (Oxidative phosphorylation). In the DNA methylation branch, three pathways reached significance (adjusted p-value <0.05): hsa05206 (MicroRNAs in cancer), hsa03050 (Proteasome), and hsa00190 (Oxidative phosphorylation). Duan et al. showed that PIK3CA mutations activate the PI3K/Akt cascade to drive colorectal tumor growth and metastasis[52]. Manzoor et al. described how autophagy supports metastatic colorectal cancer cell survival under stress[53]. Qin et al. demonstrated that dysregulated endocytosis facilitates immune escape and metastatic dissemination in colon cancer[54]. Li et al. reported that proteasome-mediated degradation of HIF-1α impairs angiogenesis and migration, thereby restraining CRC metastatic progression[55]. Liu et al. revealed that enhanced oxidative phosphorylation in colorectal cancer stem cells increases ATP production to fuel invasion and metastatic colonization[56]. Together, these findings confirm the relevance of these pathways to colorectal cancer metastasis and demonstrate that the pathway attention layer can identify potential therapeutic targets.

4. Conclusion

In this study, we developed BioLM-NET, a deep learning–based model that integrates prior biological knowledge and LLM embeddings with multi-omics data to improve predictive performance and interpretability. We evaluated BioLM-NET on five datasets spanning various prediction tasks, including cancer subtype classification, AD patient detection, and identification of metastatic regions in colorectal cancer. BioLM-NET achieved the highest performance on three datasets outperforming state-of-the-art models such as P-NET, PASNet, and traditional machine learning and deep learning models and tied on one dataset. Furthermore, model interpretation revealed that the most important input features identified by BioLM-NET were significantly enriched in Gene Ontology (GO) terms and KEGG pathways, suggesting potential as biomarkers or therapeutic targets for the respective diseases. However, BioLM-NET is currently limited to bulk and single-cell omics data and cannot yet handle spatial omics or longitudinal data. Its dual-branch design also restricts integration to only two modalities. Future work will focus on extending the framework to incorporate additional omics types, spatial information, and temporal dynamics to broaden its applicability and biological insight. Supplemental materials are available at https://github.com/bozdaglab/BioLM-NET/blob/main/Supplemental_Material.pdf .

References

1. Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).

2. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* **38**, 672-684.e6 (2020).
3. Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* **15**, 290–298 (2018).
4. Fortelny, N. & Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biology* **21**, 190 (2020).
5. Wang, H. *et al.* Biologically-informed deep neural networks provide quantitative assessment of intratumoral heterogeneity in post-treatment glioblastoma. *Res Sq* rs.3.rs-3891425 (2024) doi:10.21203/rs.3.rs-3891425/v1.
6. Hao, J., Kim, Y., Kim, T.-K. & Kang, M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics* **19**, 510 (2018).
7. Bian, S. *et al.* Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).
8. Comprehensive molecular portraits of human breast tumors. *Nature* **490**, 61–70 (2012).
9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
10. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

11. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* **64**, S161–S189 (2018).
12. Wang, T. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* **12**, 3445 (2021).
13. Yang, Z. *et al.* MLOmics: Cancer Multi-Omics Database for Machine Learning. *Sci Data* **12**, 913 (2025).
14. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).
15. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363–1375 (2019).
16. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
17. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53**, D672–D677 (2025).
18. Chen, Y. & Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv* 2023.10.16.562533 (2024) doi:10.1101/2023.10.16.562533.

19. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278 (2015).
20. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **16**, 284–287 (2012).
21. Rajabli, F. *et al.* Admixture mapping identifies novel Alzheimer disease risk regions in African Americans. *Alzheimers Dement* **19**, 2538–2548 (2023).
22. Felsky, D. *et al.* The Caribbean-Hispanic Alzheimer's disease brain transcriptome reveals ancestry-specific disease mechanisms. *Neurobiol Dis* **176**, 105938 (2023).
23. Wainberg, M., Andrews, S. J. & Tripathy, S. J. Shared genetic risk loci between Alzheimer's disease and related dementias, Parkinson's disease, and amyotrophic lateral sclerosis. *Alzheimers Res Ther* **15**, 113 (2023).
24. Oveisgharan, S. *et al.* G-protein coupled estrogen receptor 1, amyloid-β, and tau tangles in older adults. *Commun Biol* **7**, 569 (2024).
25. Wang, Q., Tian, J., Chen, H., Du, H. & Guo, L. Amyloid beta-mediated KIF5A deficiency disrupts anterograde axonal mitochondrial movement. *Neurobiol Dis* **127**, 410–418 (2019).
26. Selvarasu, K. *et al.* Reduction of kinesin I heavy chain decreases tau hyperphosphorylation, aggregation, and memory impairment in Alzheimer's disease and tauopathy models. *Front Mol Biosci* **9**, 1050768 (2022).

27. Maiese, K. Targeting the core of neurodegeneration: FoxO, mTOR, and SIRT1. *Neural Regeneration Research* **16**, 448 (2021).
28. Jian, C. *et al.* Microglia Mediate the Occurrence and Development of Alzheimer's Disease Through Ligand-Receptor Axis Communication. *Front. Aging Neurosci.* **13**, (2021).
29. Zhang, X., Alshakhshir, N. & Zhao, L. Glycolytic Metabolism, Brain Resilience, and Alzheimer's Disease. *Front Neurosci* **15**, 662242 (2021).
30. Qiu, Z. *et al.* The significance of glycolysis index and its correlations with immune infiltrates in Alzheimer's disease. *Front Immunol* **13**, 960906 (2022).
31. Wang, F. *et al.* A Glycolysis Gene Methylation Prediction Model Based on Explainable Machine Learning for Alzheimer's Disease. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3797592 (2021).
32. Zheng, Z.-Q. *et al.* Chromobox 7/8 serve as independent indicators for glioblastoma via promoting proliferation and invasion of glioma cells. *Front. Neurol.* **13**, (2022).
33. de Freitas, G. P. A. *et al.* Centromere protein J is overexpressed in human glioblastoma and promotes cell proliferation and migration. *J Neurochem* **162**, 501–513 (2022).
34. Wang, J. *et al.* DACH1 inhibits glioma invasion and tumor growth via the Wnt/catenin pathway. *Onco Targets Ther* **11**, 5853–5863 (2018).
35. Wu, G. *et al.* Expression and clinical significance of CPS1 in glioblastoma multiforme. *Curr Res Transl Med* **67**, 123–128 (2019).

36. Hwang, K.-T. *et al.* Prognostic influences of BCL1 and BCL2 expression on disease-free survival in breast cancer. *Sci Rep* **11**, 11942 (2021).
37. Zou, A. *et al.* Elevated CXCL1 expression in breast cancer stroma predicts poor prognosis and is inversely associated with expression of TGF-β signaling proteins. *BMC Cancer* **14**, 781 (2014).
38. Guo, M. *et al.* Eukaryotic Translation Initiation Factor 2 Subunit β as a Prognostic Biomarker Associates With Immune Cell Infiltration in Breast Cancer. *J Surg Res* **295**, 753–762 (2024).
39. Susini, T. *et al.* Immunohistochemical Evaluation of FGD3 Expression: A New Strong Prognostic Factor in Invasive Breast Cancer. *Cancers (Basel)* **13**, 3824 (2021).
40. Peña-Llopis, S., Wan, Y. & Martinez, E. D. Unique epigenetic gene profiles define human breast cancers with poor prognosis. *Oncotarget* **7**, 85819–85831 (2016).
41. Conte, F. *et al.* In silico recognition of a prognostic signature in basal-like breast cancer patients. *PLoS One* **17**, e0264024 (2022).
42. Liu, D.-T. *et al.* Clinical and prognostic significance of SOX11 in breast cancer. *Asian Pac J Cancer Prev* **15**, 5483–5486 (2014).
43. Xu, X.-Y. *et al.* Profile and clinical significance of SPARCL1 and its prognostic significance in breast cancer. *Oncol Lett* **29**, 196 (2025).
44. Dong, Y.-M. & Bao, G.-Q. Characterization of SUSD3 as a novel prognostic biomarker and therapeutic target for breast cancer. *Clin Transl Oncol* **27**, 935–949 (2025).

45. Lin, Y.-C., Lee, Y.-C., Li, L.-H., Cheng, C.-J. & Yang, R.-B. Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition. *J Cell Sci* **127**, 85–100 (2014).
46. Wen, L. *et al.* Prognostic and Immunological Significance of NMNAT1 in Colorectal and Pan-Cancer Contexts. *Onco Targets Ther* **18**, 389–410 (2025).
47. Wu, D. *et al.* Comprehensive analysis of the immune implication of FABP4 in colon adenocarcinoma. *PLOS ONE* **17**, e0276430 (2022).
48. Jiang, L.-W. *et al.* Investigating the relevance of nucleotide metabolism in the prognosis of glioblastoma through bioinformatics models. *Sci Rep* **15**, 5363 (2025).
49. Wang, X. *et al.* Targeting pyrimidine synthesis accentuates molecular therapy response in glioblastoma stem cells. *Sci Transl Med* **11**, eaau4972 (2019).
50. Zhou, W. *et al.* Purine metabolism regulates DNA repair and therapy resistance in glioblastoma. *Nat Commun* **11**, 3811 (2020).
51. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939778.
52. Wang, H., Tang, R., Jiang, L. & Jia, Y. The role of PIK3CA gene mutations in colorectal cancer and the selection of treatment strategies. *Front. Pharmacol.* **15**, (2024).

53. Manzoor, S., Muhammad, J. S., Maghazachi, A. A. & Hamid, Q. Autophagy: A Versatile Player in the Progression of Colorectal Cancer and Drug Resistance. *Front. Oncol.* **12**, (2022).
54. Qin, Y., Mao, Y., Han, Y., Cheng, K. & Shi, J. Immune Escape and Metabolic Reprogramming in Colon Cancer: Insights from Endocytosis-Related Genes. *BIO Web Conf.* **111**, 01019 (2024).
55. Li, J. *et al.* Identification of COX4I2 as a hypoxia-associated gene acting through FGF1 to promote EMT and angiogenesis in CRC. *Cellular & Molecular Biology Letters* **27**, 76 (2022).
56. Liu, S. *et al.* Lactate promotes metastasis of normoxic colorectal cancer stem cells through PGC-1α-mediated oxidative phosphorylation. *Cell Death Dis* **13**, 651 (2022).