Bioinformatics, 2026, **42**(6), btag253
https://doi.org/10.1093/bioinformatics/btag253
Published: 9 May 2026
Original Paper | Data and text mining

OXFORD

PEARL: integrative multi-omics classification and omics feature discovery via deep graph learning

Quan Zhao[1,‡], Jiawen Du[2,‡], Muqing Zhou[3], Xu-Wen Wang[4,*], Quan Sun[5,6,*], Can Chen[1,2,7,8,*]

[1]Carolina Health Informatics Program, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, United States
[5]Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Phildaelphia, PA 19104, United States
[6]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States
[7]School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[8]Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
*Corresponding authors. Xu-Wen Wang, Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, United States. E-mail: spxuw@channing.harvard.edu; Quan Sun, Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Phildaelphia, PA 19104, United States. E-mail: sunq@chop.edu; Can Chen, School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States. E-mail: canc@unc.edu.
[‡]Equal contribution.
Associate Editor: Jonathan Wren

Abstract

Motivation: **Integrating multi-omics data provides valuable insights into biological processes by capturing information across multiple molecular layers, enabling a comprehensive understanding of complex diseases and driving advancements in precision medicine. However, existing computational methods for multi-omics integration face significant challenges, such as low reliability and poor generalizability, due to the high dimensionality and low sample size nature of omics data.**

Results: **To address these challenges, we present PEARL (Pearson-Enhanced spectrAl gRaph convoLutional networks), a novel deep graph learning method for biomedical classification and functional important omics features identification. PEARL leverages a simple yet effective learning architecture to achieve superior and robust performance in high-dimensional, low-sample-size multi-omics settings. Our results demonstrate that PEARL significantly outperforms existing state-of-the-art methods on both synthetic and real biomedical datasets. Furthermore, applied to Alzheimer's disease (AD) brain multi-omics data, features prioritized by PEARL lead to functionally important genes that demonstrate significant enrichment in AD-related pathways. These findings highlight PEARL's practical utility in biomedical research and its potential to enhance biological interpretability in multi-omics studies.**

Availability and implementation: **The source code of our computational framework is available at https://github.com/zqq121017/PEARL.**

1 Introduction

Advances in high-throughput sequencing technologies have revolutionized biomedical research by enabling the generation of diverse and extensive omics data, including genomic, transcriptomic, and epigenetic profiles, which are crucial for understanding complex diseases (Huang *et al.* 2017, Nicora *et al.* 2020, Subramanian *et al.* 2020, Wörheide *et al.* 2021, Shahrajabian and Sun 2023, Du *et al.* 2026). Integrating these multi-omics datasets provides a powerful approach to unraveling disease heterogeneity, uncovering functional important genes, and revealing dynamic molecular interactions underlying disease progression (Hasin *et al.* 2017, Karczewski and Snyder 2018, Kreitmaier *et al.* 2023, Chen *et al.* 2023a, Yan *et al.* 2024). Moreover, as personalized medicine advances, multi-omics integration has become indispensable for refining disease diagnosis, tailoring individualized therapies, and discovering new therapeutic targets (Reska *et al.* 2021, Mohr *et al.* 2024, Molla and Bitew 2024). Therefore, developing robust multi-omics integration methods is salient to fully harness the potential of these diverse data types, leading to more accurate disease biology models, deeper understanding of molecular-level disease mechanisms, and more effective therapeutic strategies.

Received: 19 October 2025. **Revised:** 30 March 2026. **Accepted:** 28 April 2026

Numerous computational approaches have been developed for patient stratification, disease subtype classification, and drug target discovery through the integration of multi-omics data (Bersanelli *et al.* 2016, Tini *et al.* 2019, Wang *et al.* 2023, Chen *et al.* 2023a). In general, integration strategies can be categorized into three types: early integration, which directly concatenates features across omics; intermediate integration, which first learns omics-specific representations before combining them; and late integration, which fuses predictions from separate models. Among these, early integration methods typically combine features from multiple omics datasets into a single representation, while ensemble approaches aggregate predictions from separate classifiers, each trained individually on distinct omics data types (Xu *et al.* 2021, Vahabi and Michailidis 2022, Spooner *et al.* 2025). However, these methods struggle to accommodate the intrinsic properties of omics data, such as low sample sizes and high dimensionality, and fail to capture the complex correlations across different types of omics data. Most recent approaches therefore fall into intermediate and late integration categories, as they provide more flexible and effective ways to capture cross-omics relationships. The classical statistical method DIABLO (Singh *et al.* 2019) employs canonical correlation analysis for multi-view integration but still faces difficulties in effectively modeling the nonlinear relationships across omics platforms with low sample sizes. With advanced computational capabilities, deep learning methods have emerged as powerful tools for multi-omics analysis. For instance, DeepKEGG (Lan *et al.* 2024) introduces biologically-informed hierarchical frameworks that leverage pathway–gene relationships for cancer recurrence prediction and functional genes discovery. DeepMo (Lin *et al.* 2020) employs unsupervised learning approaches to stratify disease subtypes, while MOMA (Moon and Lee 2022) utilizes attention mechanisms to identify prognostic subtypes from multi-omics profiles. More recently, DeepCentroid (Xie *et al.* 2024) was introduced as a deep cascaded ensemble classifier that utilizes feature scanning and centroid-based learning to maintain robust performance in high-dimensional, small-sample-size regimes. Despite their improved performance over traditional methods, these deep learning approaches often require substantial training data and computational resources, presenting barriers to clinical adoption.

Deep graph learning methodologies, such as graph convolutional networks (GCNs), have proven to be promising solutions for enhancing multi-omics integration. These methods excel at capturing both intra- and inter-omics relationships by representing samples as nodes in similarity networks. For example, MOGONET (Wang *et al.* 2021) is a patient classification method that employs omics-specific GCNs constructed from sample similarity networks, integrating their predictions through a view correlation discovery network. Similarly, MOGDx (Ryan *et al.* 2024) utilizes a similarity network fusion to construct integrated patient networks, enabling robust disease classification even with heterogeneous data sources. MVGNN (Ren *et al.* 2024) and MOGAT (Tanvir *et al.* 2024) further advance these approaches by incorporating attention mechanisms to fuze multi-omics features and weigh the significance of molecular interactions. Recent surveys highlight the growing use of GCNs in multi-omics cancer research (Zohari and Chehreghani 2025). Large-scale comparative studies across 31 cancer types have benchmarked GCN, GAT, and GTN architectures, underscoring both their potential and the tradeoffs in scalability and robustness (Alharbi *et al.* 2025b). Other GCN-based approaches integrate embedded feature selection, such as differential expression and LASSO-based strategies, to enhance interpretability and biomarker discovery (Alharbi *et al.* 2025a). However, despite their potential, these methods often rely on highly complex architectures (e.g. graph attention networks), making them prone to overfitting and resulting in poor predictive performance when sample size is small.

To address these challenges, we propose PEARL (Pearson-Enhanced spectrAl gRaph convoLutional networks), a novel deep graph learning method for biomedical classification and functional important genes identification. PEARL leverages a simple but effective learning architecture, including weighted Pearson correlation, simple spectral GCNs, and a multi-layer perceptron (MLP), offering robust performance against sample size variations and noise commonly present in omics data. We systematically evaluate PEARL and other existing state-of-the-art multi-omics integration methods using both simulated and real biomedical datasets. Our results demonstrate that PEARL significantly outperforms existing methods, particularly under conditions of small sample sizes, while effectively identifying critical disease-associated functional genes, highlighting its potential for advancing precision medicine.

2 Materials and methods

2.1 Overview of PEARL

PEARL is a supervised intermediate multi-omics integration approach designed for biomedical classification and functional important omics features identification, effectively tackling the challenges of high dimensionality and limited sample sizes inherent in omics data. The framework consists of three key components: similarity network construction, feature refinement with GCNs, and feature integration (Fig. 1).

First, PEARL constructs sample similarity networks for different omics data using weighted Pearson correlation (Langfelder and Horvath 2008), where features with higher mean values have greater influence. This weighting strategy is based on the assumption that features with lower mean values are more susceptible to noise, thus offering greater robustness than other network construction techniques, such as cosine similarity. Importantly, the similarity networks are built solely from omics feature values and do not incorporate any class label information. This ensures that no information from the testing data labels leaks into the training process, preventing overfitting. Next, the similarity networks are represented as graphs for feature refinement using simple spectral graph convolutional networks (SSGConvs) (Zhu and Koniusz 2021). SSGConv is a simple yet effective graph learning architecture that enhances feature smoothing, reduces overfitting, and improves computational efficiency compared to traditional GCNs, making it particularly well-suited for high-dimensional, low-sample-size omics data (Zhu and Koniusz 2021). The refinement process involves an initial SSGConv layer followed by ReLU activation and dropout for regularization, as well as a second SSGConv layer with activation. This dual-layer design allows the model to capture both local and global patterns in the data while minimizing overfitting. Finally, the refined features from each omics dataset are integrated using combined pooling and/or concatenation, and fed into a MLP

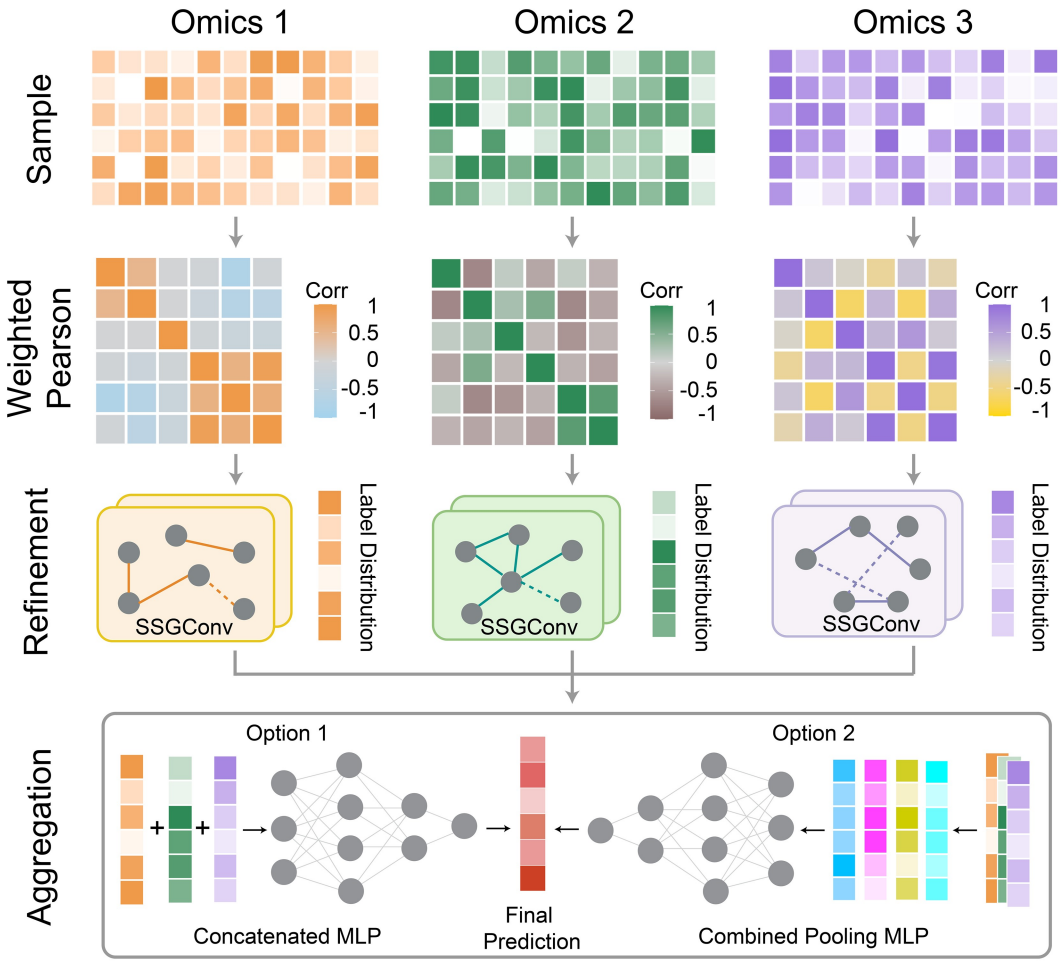

Figure 1 Overview of PEARL. PEARL is a novel multi-omics integration method based on deep graph learning for biomedical classification and functional important genes identification. Leveraging a simple yet effective architecture, PEARL first defines the graph structure for omics data using a weighted Pearson correlation similarity matrix. Omics-specific simple spectral graph convolutional networks are then employed for robust feature extraction from each view. For multi-view feature integration, PEARL offers two alternative strategies: a concatenated MLP or a combined pooling MLP, allowing for flexible fusion before the final prediction. The efficacy of PEARL has been validated across both simulated and real biomedical datasets.

for the final prediction. The combined pooling method unifies various operations such as maximum, minimum, mean, and max–min pooling, which enables the model to capture inter-omics patterns, such as dominant signals from particular modalities and variability across omics layers, without requiring the construction of a complex heterogeneous graph. Meanwhile, concatenation effectively integrates multiple feature sets by preserving all available information from diverse data sources. Detailed descriptions of each component can be found in the following.

2.2 Network construction

PEARL employs weighted Pearson correlation (Langfelder and Horvath 2008) to assign feature-specific weights and compute a correlation network for each view. Unlike the traditional Pearson correlation coefficient, the weighted version incorporates a tailored weighting scheme for each feature, effectively addressing a key challenge in multi-omics data analysis, namely, that not all features contribute equally to underlying biological processes. Our approach leverages feature means as biologically-informed weights, based on the observation that features with higher average expression or activity levels often reflect greater biological relevance. Specifically, given two samples with feature vectors $\mathbf{x}$ and $\mathbf{y}$, and a feature weight vector $\mathbf{w}$ derived from the mean value of each feature across all samples, we first compute the weighted feature representations as $\mathbf{x}_{\text{weighted}} = \mathbf{w} \odot \mathbf{x}$ and $\mathbf{y}_{\text{weighted}} = \mathbf{w} \odot \mathbf{y}$, respectively, where $\odot$ denotes element-wise multiplication. We then center the weighted features by subtracting their respective means, i.e. $\mathbf{x}_{\text{centered}} = \mathbf{x}_{\text{weighted}} - \text{mean}(\mathbf{x}_{\text{weighted}})$ and $\mathbf{y}_{\text{centered}} = \mathbf{y}_{\text{weighted}} - \text{mean}(\mathbf{y}_{\text{weighted}})$. Finally, the weighted Pearson correlation between the two samples is computed as:

$$\text{corr} = \frac{\mathbf{x}_{\text{centered}}^{\top} \cdot \mathbf{y}_{\text{centered}}}{\|\mathbf{x}_{\text{centered}}\| \cdot \|\mathbf{y}_{\text{centered}}\|}. \quad (1)$$

Using this simple yet biologically meaningful formulation, we effectively construct a sample similarity network for each modality that captures latent relationships between samples.

2.3 Feature refinement

PEARL utilizes SSGConv layers (Zhu and Koniusz 2021) to refine sample features over the constructed similarity networks. SSGConv is a simple but sophisticated spectral-based deep learning architecture that is particularly well-suited for multi-omics data analysis scenarios characterized by high dimensionality and limited sample sizes. For each omics-specific similarity network, the SSGConv layer is defined as

$$H^{(l+1)} = (\alpha \cdot \widehat{A} H^{(l)} + (1-\alpha) \cdot H^{(l)}) \Theta^{(l)}, \tag{2}$$

where $\widehat{A}$ is the normalized adjacency matrix of the network with added self-loops, $H^{(l)}$ is the feature matrix at layer l ($H^{(0)}$ represents the original omics data), $\Theta^{(l)}$ is the learned parameter matrix at layer l, and α is a weighting factor that balances the smoothness of the neighborhood aggregation with the retention of the previous features. SSGConv can be implemented efficiently using sparse matrix operations, crucial for handling high-dimensional omics data. In PEARL, we employ a two-layer SSGConv architecture to refine each omics feature representation over its corresponding similarity network, incorporating ReLU as the nonlinear activation function and dropout for regularization. The use of SSGConv equips PEARL with a robust framework for handling high-dimensional, low-sample-size multi-omics data, enabling effective feature learning while ensuring computational efficiency.

2.4 Feature integration

PEARL offers two feature integration strategies: combined pooling and concatenation. The combined pooling approach integrates features from multiple omics views using a variety of pooling operations, such as maximum, minimum, average, and difference pooling, followed by a fully connected MLP for classification. Given the refined features from M views $\{h^{(1)}, h^{(2)}, \ldots, h^{(M)}\}$ for each sample, the pooling operations can be mathematically defined as:

$$\begin{aligned}
\text{MaxPool}\left(\{h^{(1)}, h^{(2)}, \ldots, h^{(M)}\}\right)_i &= \max_{m=1,2,\ldots,M} h_i^{(m)}, \\
\text{MinPool}\left(\{h^{(1)}, h^{(2)}, \ldots, h^{(M)}\}\right)_i &= \min_{m=1,2,\ldots,M} h_i^{(m)}, \\
\text{AvePool}\left(\{h^{(1)}, h^{(2)}, \ldots, h^{(M)}\}\right)_i &= \frac{1}{M}\sum_{m=1}^{M} h_i^{(m)}, \\
\text{DiffPool}\left(\{h^{(1)}, h^{(2)}, \ldots, h^{(M)}\}\right)_i &= \max_{m=1,2,\ldots,M} h_i^{(m)} \\
&\quad - \min_{m=1,2,\ldots,M} h_i^{(m)}
\end{aligned}$$

The pooled features are concatenated to form a unified feature vector for each sample, denoted by h. This unified feature vector is then passed through an MLP with a final output dimension of two for binary classification, i.e.

$$h^{(l+1)} = \sigma\left(W^{(l)} h^{(l)} + b^{(l)}\right), \tag{3}$$

where $W^{(l)}$ and $b^{(l)}$ represent the weights and biases of the lth layer, respectively, and σ denotes the activation function (e.g. Softmax). The second pooling approach, i.e. concatenation, is relatively simpler, where it directly concatenates the refined features from different omics views and feeds them into an MLP for final classification. In our experiment, we used combined pooling for multi-class datasets and concatenated MLP for binary classification tasks, while providing both integration strategies as user-definable input options in our published implementation. These two pooling strategies in PEARL provide a comprehensive yet flexible framework for integrating features across diverse omics types.

2.5 Feature identification

Beyond its prediction power, PEARL also has the ability to identify the important features of omics data that contribute the most to prediction. Specifically, for a given feature, we first set its value vector to zero to simulate the absence of this feature as a perturbed dataset. This dataset is then processed as usual and is used to train a perturbed prediction model. We define the importance score of this feature by calculating the difference in model performance metrics ($F1$ score or $F1$ macro) between the perturbed model and the baseline model that includes this feature. We note that each omics data may contribute to the prediction unequally. Thus, we normalize the importance scores across all the features within each omics layer. In our experiments, we performed five repeats of each dataset and calculated the average of feature importance scores to account for randomness of training process.

2.6 GO and KEGG enrichment analysis

To explore the biological functions and pathways associated with the functional important genes identified by PEARL, we first used *MethylationEPIC* v2.0 annotation from Illumina and *TargetScan* to map the corresponding genes from epigenetic-markers and miRNA. After obtaining the gene set, GO and KEGG pathway enrichment analysis were conducted with the *enrichGO* and *enrichKEGG* functions, respectively, from the *clusterProfiler* R package (Yu *et al.* 2012). To account for multiple testing, FDR was calculated by adjusting P-values using the Benjamini–Hochberg method. GO terms and pathways with adjusted P-values $<.05$ are considered to be significant.

3 Results

3.1 PEARL demonstrates consistent superiority over existing methods across simulated scenarios

We first evaluated PEARL on synthetic data generated by InterSIM (Chalise *et al.* 2016), a method for generating synthetic multi-omics datasets that simulate real biological data. We considered three types of omics data, including methylation, gene expression, and protein data, and examined various cluster mean shifts and noise levels (i.e. difficulty levels) across sample sizes of 300, 500, and 800, resulting in a total of nine synthetic datasets (Note 1, available as supplementary data at *Bioinformatics* online). PEARL was compared with state-of-the-art multi-omics

integration methods MOGONET (Wang *et al.* 2021) (late integration), MOMA (Moon and Lee 2022) (late integration), MOGDx (Ryan *et al.* 2024) (intermediate integration) (Note 2, available as supplementary data at *Bioinformatics* online), as well as baseline classifiers, including *K*-nearest neighbor (KNN, early integration) and random forest (RF, early integration). We used accuracy, weighted average *F*1 score (*F*1 weighted), and macro-averaged *F*1 score (*F*1 macro) as evaluation metrics and reported the mean and SD of those metrics from 30 random train-test splits with train-test ratio 8:2.

As illustrated in Fig. 2, PEARL consistently outperforms all competing methods across varying sample sizes and difficulty levels, excelling in all evaluated metrics (accuracy, *F*1 weighted, and *F*1 macro). Notably, PEARL demonstrates exceptional robustness in low-sample-size settings ($n = 300$). For example, at the intermediate difficulty level (Fig. 2b), PEARL achieves an

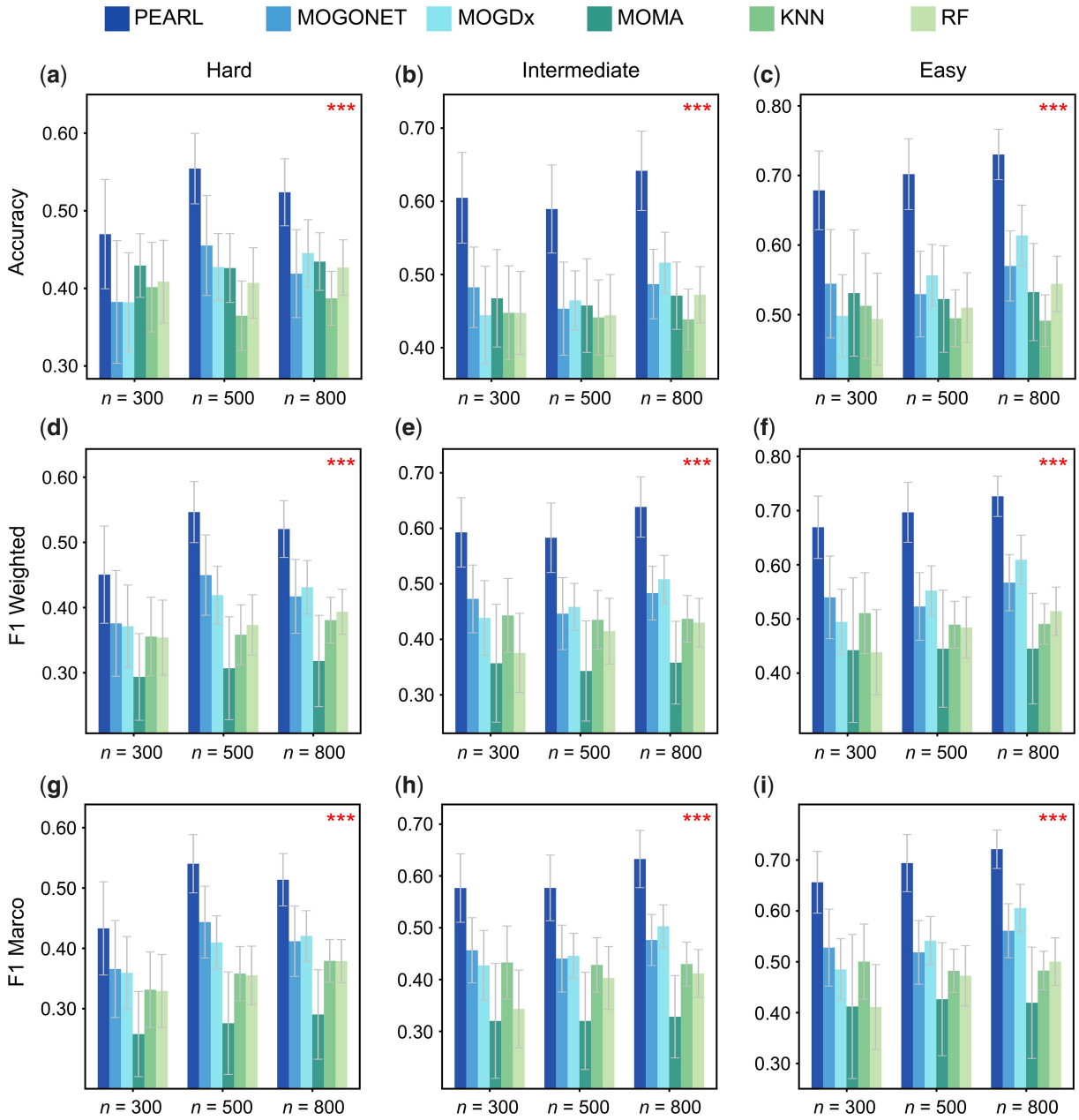

Figure 2 Performance comparison on synthetic datasets. Three sets of parameters were used with InterSIM to generate synthetic datasets. Hard refers to the cases when the cluster mean shift (delta) is equal to 0.1 and the noise level is equal to 0.4; intermediate refers to the cases when the cluster mean shift is equal to 0.15 and the noise level is equal to 0.45; and then easy refers to the cases when the cluster mean shift is equal to 0.2, and the noise level is equal to 0.5. For each case, we used three sample sizes (*n*) to generate synthetic datasets, which are 300, 500, and 800. We utilized accuracy, *F*1 weighted, and *F*1 macro to measure the performance of PEARL, MOGONET, MOGDx, MOMA, KNN, and RF on the synthetic datasets. To assess statistical significance, we conducted paired *t*-tests comparing PEARL with the second-best method in each case. The resulting *P*-values are indicated with asterisks in the plots, where three asterisks denote a *P*-value less than .01 across all sample sizes.

average of 25% higher relative accuracy compared to the second-best method MOGONET (*P*-value = 4.69E−10, one-sided paired Wilcoxon signed-rank). This performance advantage persists as sample sizes increase, with 30% and 32% higher accuracy compared to MOGONET in $n = 500$ and 800 scenarios (*P*-values = 4.18E−11 and 9.97E−14) under the same intermediate difficulty level. Additionally, PEARL effectively adapts to tasks of varying complexity, handling both noise-heavy scenarios with subtle patterns and simpler cases with clear class separations. Across all conditions, PEARL significantly surpasses alternatives like MOGONET, MOMA, and MOGDx with an average of 27%, 28%, and 27% higher accuracy. These results on synthetic data highlight the superiority of PEARL in various real-world data scenarios, especially given that multi-omics datasets are often limited by small sample sizes and high noise.

3.2 PEARL outperforms existing methods in various real datasets

We applied our method and other competing algorithms to three real datasets, including the Religious Orders Study/Memory and Aging Project (ROSMAP) for Alzheimer's disease (AD) patients versus normal control (NC) classification (Bennett *et al.* 2012, De Jager *et al.* 2018), BReast CAncer (BRCA) (Colaprico *et al.* 2016) for breast invasive carcinoma PAM50 subtype classification, and Lung Adenocarcinoma Collection (LUAD) from The Cancer Genome Atlas (TCGA) (Albertina *et al.* 2016) for lung cancer pathological stage classification. ROSMAP and BRCA contain three types of omics data: mRNA expression data (mRNA), DNA methylation data (meth), and miRNA expression data (miRNA), and LUAD only contains mRNA and methylation data. The numbers of features used for training are included in Table 1.

To ensure fairness and comparability, we followed the exact same preprocessing and feature selection strategy as used in MOGONET. Specifically, features were pre-filtered by applying variance thresholds followed by ANOVA *F*-values to retain those with the greatest discriminative potential. This procedure is a standard practice in multi-omics studies to reduce dimensionality while preserving biologically relevant signals (St and World 1989, Golugula *et al.* 2011, Chen *et al.* 2023b). Each dataset was randomly split into 80% training and 20% testing sets, repeated 30 times to ensure robust performance evaluation. For the binary ROSMAP dataset, we used accuracy (ACC), *F*1 score (*F*1), and area under the receiver operating characteristic curve (AUC) as evaluation metrics. Due to the multi-class characteristic of the BRCA and LUAD, we calculated ACC, average *F*1 score weighted by support (F1_weighted), and macro-averaged *F*1 score (F1_macro). We reported mean and SD of each evaluation metrics across the 30 realizations. To further ensure fairness and to address integration strategies across categories, we additionally included XGBoost (Chen and Guestrin 2016), two most recent late integration methods [DeepCentroid (Xie *et al.* 2024), 2024; fuseMLR (Fouodo *et al.* 2025), 2025], and one intermediate integration method [iDeepViewLearn (Wang *et al.* 2024), 2024] in the real data experiment.

Our results show that PEARL consistently outperforms all other methods in all datasets (Fig. 3), with notable improvements in all metrics. In the BRCA dataset with a challenging situation of classifying five different subtypes of breast cancer (Fig. 3a–c), PEARL outperforms the second best by 2% in accuracy, 3% in *F*1 weighted, and 1% in *F*1 macro (*P*-value = .14, .018, .05, respectively). In LUAD dataset (Fig. 3d–f) with situation of classifying three pathological states, PEARL demonstrates remarkable improvements compared to the second best by 5% in accuracy, 12% in *F*1 weighted and 13% in *F*1 macro (*P*-value = 1E−3, 1.6E−7, 1E−5).

In the ROSMAP dataset (Fig. 3g–i), which presents the challenging task of distinguishing AD patients from NCs, PEARL demonstrates improvements compared to the second-best model, XGBoost, achieving gains of 2% in accuracy and AUC, and 3% in *F*1 score (*P*-value = .013, .05, .018, respectively). These improvements are particularly meaningful in the context of AD diagnosis, where even small increases in accuracy can potentially translate to earlier and more reliable disease detection, potentially enabling earlier intervention strategies.

All methods were run with their default parameter configurations to ensure consistency in comparative evaluation in above analysis. To evaluate the impact of parameter tuning on existing methods, we adopted a grid search strategy, and the results confirm the superiority of PEARL (Fig. 1, available as

Table 1 Summary of datasets.[a]

Dataset	Categories	Number of features mRNA, meth, miRNA	Number of features training mRNA, meth, miRNA
ROSMAP	NC: 169, AD: 182	55 889, 23 788, 309	200, 200, 200
BRCA	Normal-like: 115, Basal-like: 131, HER2-enriched: 46, Luminal A: 436, Luminal B: 147	20 531, 20 106, 503	1000, 1000, 503
LUAD	Stage I: 278, Stage II: 124, Stage III/IV: 110	20 531, 22 601, –	1000, 1000, –
METABRIC	Basal: 209, Her2: 224, Normal: 148, LumA: 700, LumB: 475	20 385, 13 184, –	782, 468, –

[a] mRNA refers to mRNA expression data. meth refers to DNA methylation data. miRNA refers to miRNA expression data. The ROSMAP dataset is for the classification of Alzheimer's disease (AD) patients versus normal control (NC). The BRCA dataset is for breast invasive carcinoma (BRCA) PAM50 subtype classification with normal-like, basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal A, and Luminal B subtypes. The LUAD dataset is for lung adenocarcinoma pathological stage classification with Stage I, Stage II, and Stage III/IV. METABRIC is an independent breast cancer cohort from the Molecular Taxonomy of Breast Cancer International Consortium, used for external validation of models trained on BRCA, with the same PAM50 subtype classification task using mRNA and DNA methylation data.

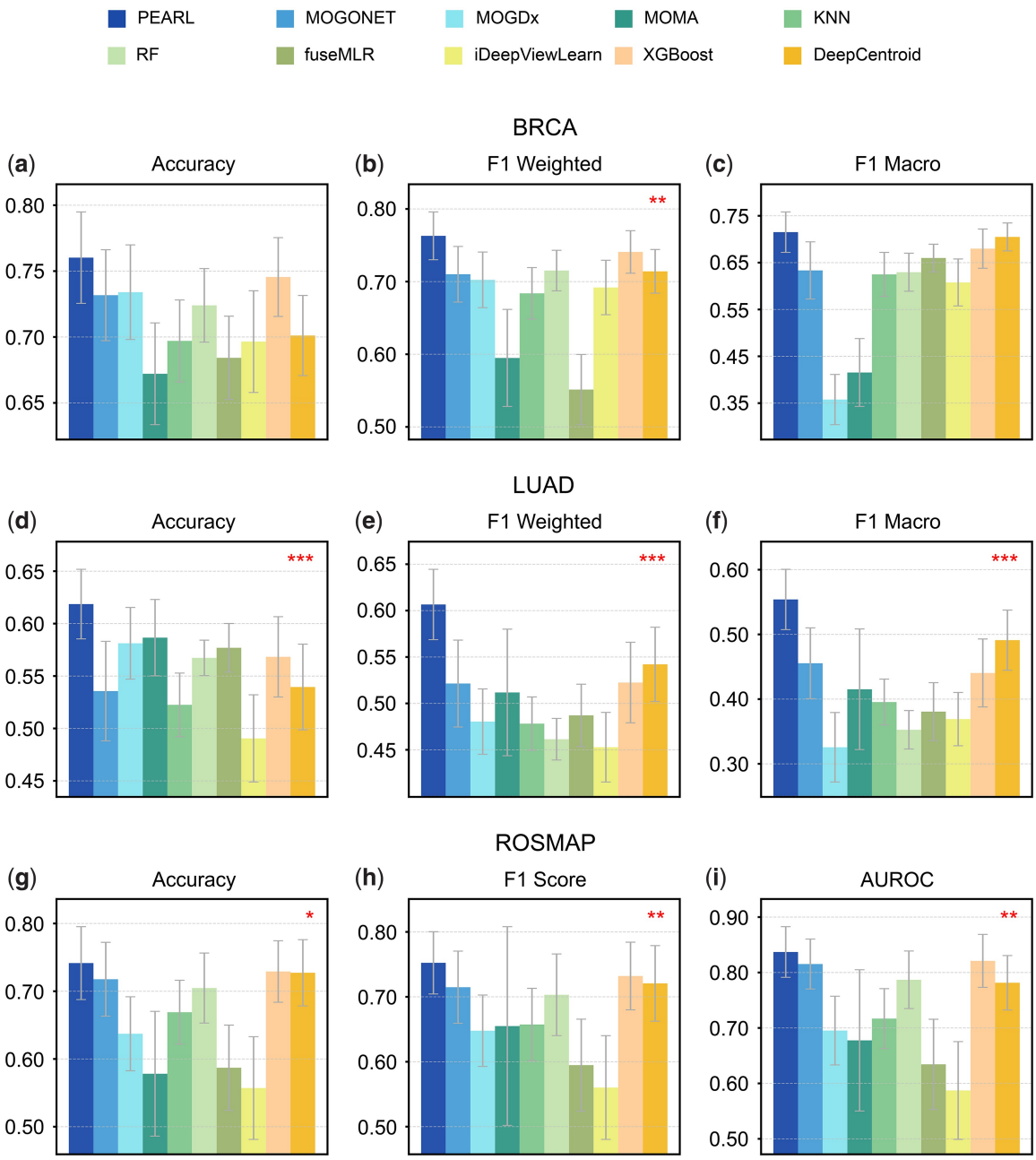

Figure 3 Performance comparison on real datasets. We compared the performance of PEARL with nine other methods, MOGONET, MOGDx, MOMA, KNN, RF, fuseMLR, iDeepViewLearn, XGBoost, and DeepCentroid on the BRCA, LUAD, and ROSMAP datasets. We used accuracy, *F*1 weighted, and *F*1 macro as metrics on the BRCA and LUAD multi-view datasets. For the ROSMAP binary dataset, we calculated accuracy, *F*1 score, and AUROC as evaluation metrics. To assess statistical significance, we conducted Wilcoxon signed-rank tests comparing PEARL with the second-best method in each case. The resulting *P*-values are indicated with asterisks in the plots [*P*-value $<$.1(*)/.05(**)/.01(***)].

supplementary data at *Bioinformatics* online). We further performed an ablation study to examine the impact of our key innovations in PEARL (Note 3, available as supplementary data at *Bioinformatics* online). The results demonstrate that removing any one of our key components, including weighted Pearson correlation, SSGConv, or feature integration, causes a stepwise decline in all performance metrics, confirming that each innovation is essential for PEARL's best results (Fig. 2, available as supplementary data at *Bioinformatics* online).

To further demonstrate the real-world applicability of our framework, we conducted an independent cross-study validation. We trained PEARL on the BRCA dataset and evaluated its performance on METABRIC dataset (Curtis *et al.* 2012, Pereira *et al.* 2016) (Table 1), which serves as an independent resource for breast cancer multi-omics. In this independent validation, PEARL outperformed the second-best model (i.e. MOGONET) by 42.4% in accuracy, 97.3% in *F*1 weighted and 33.3% in *F*1 macro (Fig. 4). This independent validation underscores PEARL's ability to

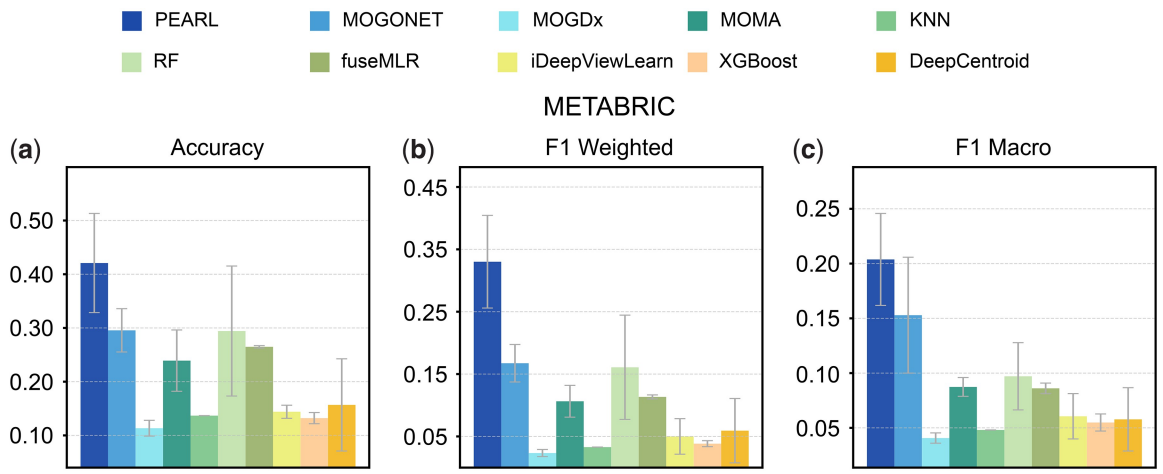

Figure 4 Performance on independent dataset. We trained PEARL on BRCA and tested on METABRIC cohort. We compared the performance of PEARL with nine other methods, MOGONET, MOGDx, MOMA, KNN, RF, fuseMLR, iDeepViewLearn, XGBoost, and DeepCentroid. Accuracy, *F*1 weighted, and F1 macro were used as evaluation metrics.

Table 2 Feature identification results from ROSMAP by PEARL.

Omics data type	Features
mRNA expression	ENSG00000117266.11, ENSG00000155980.6, ENSG00000105643.3, ENSG00000204219.5, ENSG00000182310.7, ENSG00000168743.8, ENSG00000071054.10, ENSG00000102032.8, ENSG00000166682.6
DNA methylation	cg07232688, cg22518733, cg27091787, cg25736482, cg14387505, cg12447832, cg08418111, cg02008416, cg06946880, cg10157098, cg24765079, cg04983977, cg14847483
miRNA expression	hsa-miR-517c_hsa-miR-519a, hsa-miR-424, hsa-miR-216a, hsa-miR-651, hsa-miR-127-3p, hsa-miR-640, hsa-miR-199b-5p, hsa-miR-199a-5p

Table 3 Feature identification results from BRCA by PEARL.

Omics data type	Features
mRNA expression	SOX11|6664, HPDL|84842, CNGA1|1259, KIF2C|11004, AR|367, CAPN13|92291, CDC6|990, KIF14|9928, PTCHD1|139411, C1orf64|149563, PHYHD1|254295
DNA methylation	C14orf72, OR1J4, TM4SF18, LEMD1, ROPN1, ATP8B1, C21orf82, GPR37L1, KDM2B, ATP10B, LGALS2
miRNA expression	hsa-mir-187, hsa-mir-9–1, hsa-mir-9–2, hsa-mir-203, hsa-mir-21, hsa-mir-184, hsa-mir-944, hsa-mir-551b

Table 4 Feature identification results from LUAD by PEARL.

Omics data type	Features
mRNA expression	TMEM8B, VCL, MRPL19, PHF1, PLK1, NAPB, SERPINI1, RPS6KA5, CPT1B, C2CD4D-AS1, ZC3H15, TRIM13, PGM1, NPHP3, GPR55, MAT2A, PGM5, ORMDL3, CLK4, KLHDC1, ZC3HAV1L, PYGL, FOXM1, SRGAP3, AGPS, LRFN4
DNA methylation	cg21770145, cg07015079, cg05322019, cg08554114

transcend study specific batch effects, confirming the reliability for clinical applications across different data sources.

3.3 Identification of disease-associated genes and pathways

Finally, we investigated the key omics features that contributed the most to the prediction (Tables 2–4). Details of the identification process are provided in Section 2.

In the ROSMAP dataset, we chose the top 50 features with highest importance scores, across mRNA, DNA methylation, and miRNA data. We annotated the cpg sites to their nearest gene using Illumina MethylationEPIC v2.0 manifest file. For miRNA, they were filtered to those present in the TargetScan human database (McGeary *et al.* 2019) and further mapped to corresponding miRNA families. Further, the target genes were extracted from the TargetScan's target predictions. All gene symbols were subsequently converted to ENSEMBL gene IDs using the *org. Hs.eg.db* (Carlson *et al.* 2019) annotation database. Overall, these 50 features correspond to 2861 genes. Then, we

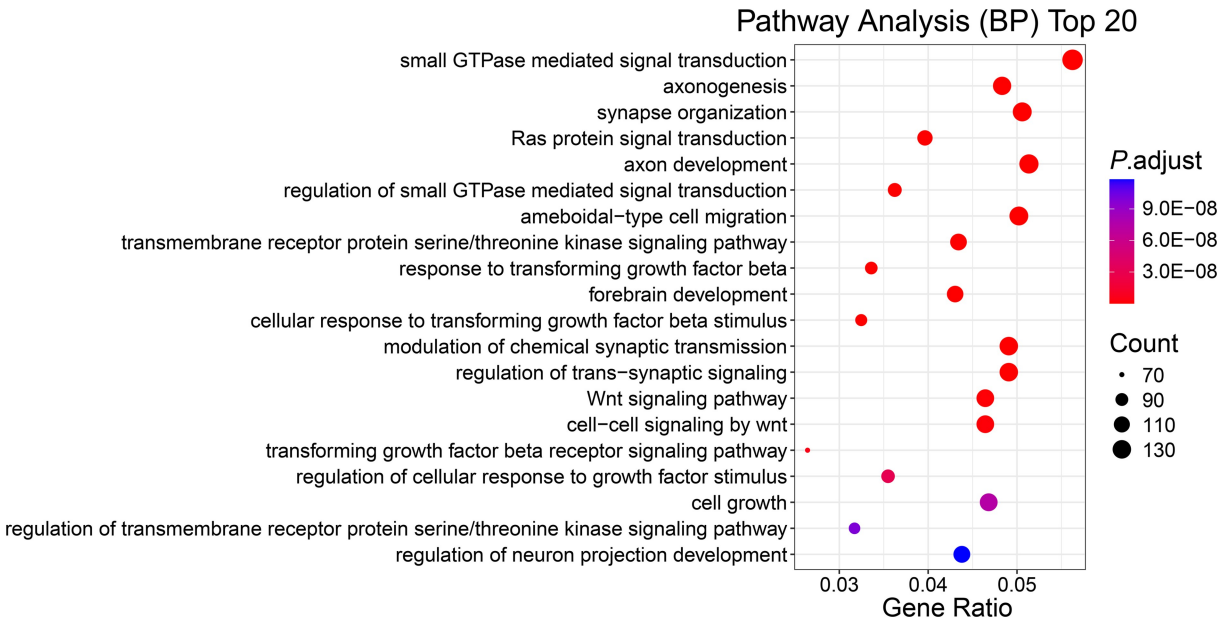

Figure 5 Top 20 enriched biological processes from pathway analysis ordered by adjusted *P*-value. Each dot represents a biological process, with the *x*-axis showing the gene ratio (the ratio of genes associated with a given term to the total number of genes analyzed), and the *y*-axis listing the description of each biological process. Color corresponds to the adjusted *P*-value, with more intense red indicating higher significance. Dot size reflects the gene count. Several notable significantly enriched pathways include processes related to GTPase-mediated signaling, axongenesis, and axon development.

performed enrichment analysis for the functional genes identified in ROSMAP, using both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Kanehisa and Goto 2000). GO terms within three ontology categories are considered: biological process (BP), cellular components (CCs), and molecular function (MF). To account for multiple testing, false discovery rate (FDR) correction was applied and the adjusted *P*-values were reported. We found these genes were enriched for several GO terms that have been reported in other studies to be associated with AD. Notable examples include small GTPase-mediated signal transduction (biological process, GO:0007264, adjusted *P*-value = 1.07E−16, Fig. 5) (Hu *et al.* 2015), axongenesis (biological process, GO:0007409, adjusted *P*-value = 5.11E−13, Fig. 5) (Huang *et al.* 2018), axon development (biological process, GO:0061564, adjusted *P*-value = 2.56E−12, Fig. 5) (Santa-Maria *et al.* 2015), glutamatergic synapse (cellular composition, GO:0098978, adjusted *P*-value = 1.36E−23, Fig. 3, available as supplementary data at *Bioinformatics* online) (Ma *et al.* 2024), and protein serine/threonine kinase activity (molecular function, GO:0004674, adjusted *P*-value = 2.98E−9, Fig. 4, available as supplementary data at *Bioinformatics* online) (Zhang *et al.* 2022). The full GO enrichment results are provided in Tables 3–5, available as supplementary data at *Bioinformatics* online.

The KEGG pathway enrichment analysis also identified several important pathways that are known to be critically related to AD (Table 6and Worldsupplementary data at *Bioinformatics* online). The two most notable pathways are axon guidance (adjusted *P*-value = 2.06E−13, Fig. 5, available as supplementary data at *Bioinformatics* online) and MAPK signaling pathway (adjusted *P*-value = 1.40E−10, Fig. 5, available as supplementary data at *Bioinformatics* online). Axon-guidance molecules are playing import roles in occurrence and development of AD by influencing in connectivity and repair mechanism (Lin *et al.* 2009), and the activate MAPK signaling pathway has been considered to contribute to AD via multiple mechanism like neuronal apoptosis and transcriptional and enzymatic activation of β- and γ-secretases (Kim and Choi 2010). These findings demonstrate that PEARL effectively identifies trait-associated functional genes from multi-omics data, supporting its utility in uncovering biologically meaningful insights.

4 Discussion

Multi-omics integration plays crucial roles in understanding disease mechanisms, providing pivotal perspectives in disease diagnosis and prognosis. In this study, we present PEARL, a novel framework for multi-omics data integration and analysis that advances the field through several key innovations. Our comprehensive evaluation demonstrates PEARL's superior performance across both synthetic and real-world datasets, while also providing biologically meaningful insights through its feature identification capabilities.

PEARL's performance can be attributed to several technical design choices. First, the weighted Pearson correlation approach in the preprocessing stage provides an alternative to traditional similarity measures for constructing patient similarity network. By incorporating a position-based weighting scheme, PEARL captures the varying importance of different molecular features, reflecting the biological reality that not all features contribute equally to disease mechanisms. This builds upon insights from previous studies showing that representing data as patient similarity networks can retain information and have

better performance compared to standard Euclidean methods (Li *et al.* 2015, Gliozzo *et al.* 2020). Our sensitivity analysis further validates this choice, as the weighted Pearson approach consistently outperforms standard Pearson correlation, Spearman correlation, and cosine similarity (Fig. 6, available as supplementary data at *Bioinformatics* online). Second, the use of simple but effective SSGConv layers in PEARL's GCN architecture provides another crucial advantage. Unlike traditional approaches that rely on simple feature concatenation (Sun *et al.* 2007) or majority voting for network integration (Gliozzo *et al.* 2022), SSGConv's ability to balance between feature preservation and neighborhood aggregation is especially valuable in handling the noise and heterogeneity inherent in multi-omics data. This is reflected in the consistently smaller SDs observed in our results, particularly in the ROSMAP dataset. Third, the dual integration strategies, combined pooling and concatenation MLP, demonstrate PEARL's robustness to different classification scenarios. This flexibility addresses limitations identified in existing methods (Kim *et al.* 2015) where integrative analysis often struggles with different types of classification tasks. Each component of PEARL contributes to its superiority over competing methods, validated through our results with enhanced performance for both synthetic and real datasets.

PEARL addresses several limitations of existing approaches. Unlike MOGONET, which relies on a fixed integration framework, PEARL offers more flexible handling of different data types and classification scenarios. Compared to MOMA, which faces challenges with complex optimization in multi-task learning settings that can create negative transfer effects, PEARL's architecture avoids these conflicts while maintaining computational efficiency. Similarly, while MOGDx effectively handles missing data through network fusion, it lacks the sophisticated feature weighting mechanisms present in PEARL. The weighted Pearson correlation and SSGConv architecture in PEARL provide a more robust framework for high-dimensional, low-sample-size data compared to these methods. Furthermore, PEARL's adaptive integration approach delivers superior performance across varying classification scenarios without extensive parameter tuning, offering better scalability and robustness as demonstrated in our synthetic data experiments across different sample sizes.

Computationally, PEARL is designed to balance predictive performance with resource efficiency. By utilizing a SSGConv, the framework avoids the expensive matrix operations often required in attention-based architectures. The overall time complexity is approximately $O(M \cdot N^2 \cdot d + K \cdot |E| \cdot d)$, where M is the number of omics modalities, N is the sample size, d is the number of omics features, K is the number of propagation steps, and $|E|$ is the number of edges in the similarity graph. Since feature propagation is handled as a one-time preprocessing step, PEARL requires fewer trainable parameters. Empirically, using a standard workstation (13th Gen Intel Core i7-13700K, NVIDIA GeForce RTX 3080 Ti with 12 GB VRAM), PEARL trained on ROSMAP dataset in 928 s, with a peak memory utilization of 2.48 GB.

PEARL not only achieves higher accuracy in classification tasks, but also has the ability to identify biologically meaningful important features from various omics data, providing valuable insights into disease mechanisms. Like recent successful methods (Serra *et al.* 2015), PEARL effectively integrates multiple omics data types while maintaining interpretability. In the ROSMAP data analysis, the underlying genes of prioritized omics features by PEARL show significant enrichment in AD-related pathways, particularly axon guidance and MAPK signaling, consistent with current understanding of AD pathophysiology. Of particular note is PEARL's ability to identify functionally important pathway associations. Like recent network-based approaches (Wang *et al.* 2014), PEARL excels at capturing strong latent relationships between samples. However, PEARL goes beyond simple network fusion by incorporating weighted correlations and advanced graph convolution techniques, allowing for more nuanced feature importance estimation.

To further enhance model parsimony, we embedded a LASSO (L1) penalty into the PEARL loss function. The primary advantage of LASSO is to produce a sparser, more interpretable model by forcing the weights of noncontributory features to exactly zero. In addition, it is more computational efficient for feature selection than the original method, e.g. setting a feature to zero and re-training the entire model. Our analysis shows that the penalized model yielded comparable classification performance to the original framework (Tables 3 and 4, available as supplementary data at *Bioinformatics* online). Furthermore, the feature importance scores between two models were highly correlated ($\rho = 0.95$ and 0.99 for the BRCA and ROSMAP datasets, respectively), underscoring the robustness of PEARL's native structure. However, the use of a LASSO penalty has some limitations, particularly in biological data (O'Brien 2016). It can arbitrarily select one feature from a group of highly correlated variables, such as genes within the same pathway, while discarding the rest. This can lead to the loss of potentially important co-linear signals. Therefore, a tradeoff exists that while the original PEARL framework may be preferable for retaining a comprehensive feature set, the LASSO option offers a valuable alternative for studies where model simplicity and the identification of a minimal set of biomarkers are the primary goals.

While PEARL demonstrates significant advantages, several areas warrant further investigation. First, the current implementation focuses on classification tasks; extending the framework to handle regression problems could broaden its clinical applications. This could potentially build upon approaches used in previous studies for survival prediction (Chaudhary *et al.* 2018). Second, while PEARL's complexity is highly optimized for current high-dimensional, small-sample-size datasets, the $O(N^2)$ space complexity required for the dense similarity matrix may present memory challenges with ultra-large-scale multi-omics studies (e.g. biobank-scale cohorts). This reflects a common problem in the field (Poirion *et al.* 2018), and future work utilizing subnetwork extraction or alternative sparse-graph methods could improve scalability. Third, although PEARL effectively uses combined pooling MLP to capture inter-omics information, explicitly modeling these interactions through cross-view graph connections remains an important objective. Developing robust methods to construct unified heterogeneous networks without exacerbating overfitting in low-sample-size cohorts could be the future direction. Fourth, we recognize that the success of feature identification may be influenced by the use of well-characterized datasets such as BRCA and ROSMAP. While population-scale biobanks could offer an alternative for validation, we could not find external datasets that exactly match the features used in our current study. Additionally, the reliance on

ICD codes for case identification in large biobanks may also introduce diagnostic biases, which may yield features that reflect clinical practice patterns rather than the molecular driven signals captured in curated research cohorts. Last, another promising direction would be to integrate PEARL with emerging approaches in multi-omics integration, such as canonical correlation analysis methods (Singh *et al.* 2019) or advanced group-regularized techniques (Van De Wiel *et al.* 2016), which may provide additional insights into cellular heterogeneity and tissue-specific patterns in disease progression.

In summary, PEARL represents a significant advancement in multi-omics data integration and analysis. Its superior performance across various datasets, coupled with its ability to identify biologically meaningful features, positions it as a valuable tool for precision medicine applications. Building on insights from previous unsupervised integration methods (Shen *et al.* 2009), PEARL extends these capabilities to supervised learning while maintaining robust performance in the presence of noise. As multi-omics studies continue to grow in scale and complexity, we anticipate PEARL will become increasingly valuable for understanding disease mechanisms and advancing personalized medicine approaches.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions

Quan Zhao (Data curation [equal], Formal analysis [equal], Methodology [lead], Software [lead], Validation [equal], Visualization [equal], Writing—original draft [equal]), Jiawen Du (Data curation [equal], Formal analysis [equal], Methodology [supporting], Software [supporting], Validation [equal], Visualization [equal], Writing—original draft [equal]), Muqing Zhou (Methodology [supporting], Software [supporting], Validation [equal]), Xu-Wen Wang (Supervision [equal], Writing—review & editing [equal]), Quan Sun (Supervision [equal], Writing—review & editing [equal]), and Can Chen (Conceptualization [lead], Investigation [lead], Project administration [lead], Supervision [equal], Writing—review & editing [equal])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

The authors declare that they have no competing interests.

Funding

None declared.

Data availability

The ROSMAP dataset was obtained from AMP-AD Knowledge Portal (https://adknowledgeportal.synapse.org/). Omics data of BRCA and LUAD was obtained from The Cancer Genome Atlas Program (TCGA) through Broad GDAC Firehose (https://gdac.broadinstitute.org/). PAM50 breast cancer subtypes of TCGA BRCA patients were obtained through the TCGAbiolinks R package v2.12.6. The source data and code of our computational framework is available at https://github.com/zqq121017/PEARL. The archival version of the code is preserved on Zenodo at https://doi.org/10.5281/zenodo.19875866 (zqq121017 2026).

References

Albertina B, Watson M, Holback C *et al.* The cancer genome atlas lung adenocarcinoma collection (TCGA-LUAD). *Cancer Imaging Arch* 2016.

Alharbi F, Budhiraja N, Vakanski A *et al.* Interpretable graph kolmogorov-arnold networks for multi-cancer classification and biomarker identification using multi-omics data. *Sci Rep* 2025a;**15**:27607. https://doi.org/10.1038/s41598-025-13337-0

Alharbi F, Vakanski A, Zhang B *et al.* Comparative analysis of multi-omics integration using graph neural networks for cancer classification. *IEEE Access* 2025b;**13**:37724–36. https://doi.org/10.1109/ACCESS.2025.3540769

Bennett DA, Schneider JA, Arvanitakis Z *et al.* Overview and findings from the religious orders study. *Curr Alzheimer Res* 2012;**9**:628–45.

Bersanelli M, Mosca E, Remondini D *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17 Suppl 2**:15–177.

Carlson M, Falcon S, Pages H *et al.* org. hs. eg. db: genome wide annotation for human. *R Package Version* 2019;**3**:3.

Chalise P, Raghavan R, Fridley BL. Intersim: simulation tool for multiple integrative 'omic datasets'. *Comput Methods Programs Biomed* 2016;**128**:69–74.

Chaudhary K, Poirion OB, Lu L *et al.* Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59.

Chen C, Wang J, Pan D *et al.* Applications of multi-omics analysis in human diseases. *MedComm (2020)* 2023a;**4**:e315.

Chen C, Weiss ST, Liu Y-Y. Graph convolutional network-based feature selection for high-dimensional and low-sample size data. *Bioinformatics* 2023b;**39**:btad135.

Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: Association for Computing Machinery, 2016, 785–94.

Colaprico A, Silva TC, Olsen C *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71.

Curtis C, Shah SP, Chin S-F *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012;**486**:346–52.

De Jager PL, Ma Y, McCabe C *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data* 2018;**5**:180142.

Du J, Zhou M, Wang H *et al.* Multi-omics integration predicts the incidence of 17 diseases in the UK Biobank. Nat Comm 2026 (in press).

Fouodo CJ, Bleskina M, Szymczak S. fuseMLR: an R package for integrative prediction modeling of multi-omics data. *BMC Bioinformatics* 2025;**26**:221.

Gliozzo J, Perlasca P, Mesiti M *et al.* Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction. *Sci Rep* 2020;**10**:3612.

Gliozzo J, Mesiti M, Notaro M *et al.* Heterogeneous data integration methods for patient similarity networks. *Brief Bioinform* 2022;**23**:bbac207.

Golugula A, Lee G, Madabhushi A. Evaluating feature selection strategies for high dimensional, small sample size datasets. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New York, USA: IEEE, 2011, 949–52.

Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017;**18**:83.

Hu W, Lin X, Chen K. Integrated analysis of differential gene expression profiles in hippocampi to identify candidate genes involved in Alzheimer's disease. *Mol Med Rep* 2015;**12**:6679–87.

Huang J-L, Xu Z-H, Yang S-M *et al.* Identification of differentially expressed profiles of Alzheimer's disease associated circular RNAs in a Panax notoginseng saponins-treated Alzheimer's disease mouse model. *Comput Struct Biotechnol J* 2018;**16**:523–31.

Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017;**8**:84.

Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.

Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;**19**:299–310.

Kim D, Joung J-G, Sohn K-A *et al.* Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *J Am Med Inform Assoc* 2015;**22**:109–20.

Kim EK, Choi E-J. Pathological roles of MAPK signaling pathways in human diseases. *Biochim Biophys Acta* 2010;**1802**:396–405.

Kreitmaier P, Katsoula G, Zeggini E. Insights from multi-omics integration in complex disease primary tissues. *Trends Genet* 2023;**39**:46–58.

Lan W, Liao H, Chen Q *et al.* DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Brief Bioinform* 2024;**25**:bbae185.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.

Li L, Cheng W-Y, Glicksberg BS *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* 2015;**7**:311ra174.

Lin L, Lesnick TG, Maraganore DM *et al.* Axon guidance and synaptic maintenance: preclinical markers for neurodegenerative disease and therapeutics. *Trends Neurosci* 2009;**32**:142–9.

Lin Y, Zhang W, Cao H *et al.* Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes (Basel)* 2020;**11**:888.

Ma Y, Reyes-Dumeyer D, Piriz A *et al.* Epigenetic and genetic risk of Alzheimer disease from autopsied brains in two ethnic groups. *Acta Neuropathol* 2024;**148**:27.

McGeary SE, Lin KS, Shi CY *et al.* The biochemical basis of microRNA targeting efficacy. *Science* 2019;**366**:eaav1741.

Mohr AE, Ortega-Santos CP, Whisner CM *et al.* Navigating challenges and opportunities in multi-omics integration for personalized healthcare. *Biomedicines* 2024;**12**:1496.

Molla G, Bitew M. Revolutionizing personalized medicine: synergy with multi-omics data generation, main hurdles, and future perspectives. *Biomedicines* 2024;**12**:2750.

Moon S, Lee H. Moma: a multi-task attention learning algorithm for multi-omics data interpretation and classification. *Bioinformatics* 2022;**38**:2287–96.

Nicora G, Vitali F, Dagliati A *et al.* Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Front Oncol* 2020;**10**:1030.

O'brien CM. Statistical learning with sparsity: The lasso and generalizations. *Int Stat Rev* 2016;**84**:156–7.

Pereira B, Chin S-F, Rueda OM *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun* 2016;**7**:11908.

Poirion OB, Chaudhary K, Garmire LX. Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Summits Transl Sci Proc* 2018;**2018**:197.

Ren Y, Gao Y, Du W *et al.* Classifying breast cancer using multi-view graph neural network based on multi-omics data. *Front Genet* 2024;**15**:1363896.

Reska D, Czajkowski M, Jurczuk K *et al.* Integration of solutions and services for multi-omics data analysis towards personalized medicine. *Biocybern Biomed Eng* 2021;**41**:1646–63.

Ryan B, Marioni RE, Simpson TI. Multi-omic graph diagnosis (MOGDx): a data integration tool to perform classification tasks for heterogeneous diseases. *Bioinformatics* 2024;**40**:btae523.

Santa-Maria I, Alaniz ME, Renwick N *et al.* Dysregulation of microrna-219 promotes neurodegeneration through post-transcriptional regulation of tau. *J Clin Invest* 2015;**125**:681–6.

Serra A, Fratello M, Fortino V *et al.* MVDA: a multi-view genomic data integration methodology. *BMC Bioinformatics* 2015;**16**:261.

Shahrajabian MH, Sun W. Survey on multi-omics, and multi-omics data analysis, integration and application. *CPA* 2023;**19**:267–81.

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**:2906–12.

Singh A, Shannon CP, Gautier B *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–62.

Spooner A, Moridani MK, Toplis B *et al.* Benchmarking ensemble machine learning algorithms for multi-class, multi-omics data integration in clinical outcome prediction. *Brief Bioinform* 2025;**26**:bbaf116.

St L, Wold S. Analysis of variance (ANOVA). *Chemom Intell Lab Syst* 1989;**6**:259–72. https://doi.org/10.1016/0169-7439(89)80095-4

Subramanian I, Verma S, Kumar S *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051.

Sun Y, Goodison S, Li J *et al.* Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 2007;**23**:30–7.

Tanvir RB, Islam MM, Sobhan M *et al.* MOGAT: a multi-omics integration framework using graph attention networks for cancer subtype prediction. *Int J Mol Sci* 2024;**25**:2788.

Tini G, Marchetti L, Priami C *et al.* Multi-omics integration—a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2019;**20**:1269–79.

Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet* 2022; **13**:854752.

Van De Wiel MA, Lien TG, Verlaat W *et al.* Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat Med* 2016;**35**:368–81.

Wang B, Mezlini AM, Demir F *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; **11**:333–7.

Wang H, Lu H, Sun J *et al.* Interpretable deep learning methods for multiview learning. *BMC Bioinformatics* 2024;**25**:69.

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**:3445.

Wang X-W, Wang T, Schaub DP *et al.* Benchmarking omics-based prediction of asthma development in children. *Respir Res* 2023; **24**:63.

Wörheide MA, Krumsiek J, Kastenmüller G *et al.* Multi-omics integration in biomedical research—a metabolomics-centric review. *Anal Chim Acta* 2021;**1141**:144–62.

Xie K, Hou Y, Zhou X. Deep centroid: a general deep cascade classifier for biomedical omics data classification. *Bioinformatics* 2024;**40**:btae039.

Xu F, Wang S, Dai X *et al.* Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data. *Methods* 2021;**189**:65–73.

Yan H, Weng D, Li D *et al.* Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Brief Bioinform* 2024; **25**:bbae184.

Yu G, Wang L-G, Han Y *et al.* clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS* 2012; **16**:284–7.

Zhang Q, Yang P, Pang X *et al.* Preliminary exploration of the co-regulation of Alzheimer's disease pathogenic genes by microRNAs and transcription factors. *Front Aging Neurosci* 2022;**14**:1069606.

Zhu H, Koniusz P. Simple spectral graph convolution. In: *International Conference on Learning Representations.* 2021.

Zohari P, Chehreghani MH. Graph neural networks in multi-omics cancer research: a structured survey. arXiv preprint, arXiv:2506.17234, 2025, preprint: not peer reviewed.

zqq121017. zqq121017/pearl: Pearl-1.0, April 2026. https://doi.org/10.5281/zenodo.19875866

Bioinformatics, 2026, 42, 1–13
https://doi.org/10.1093/bioinformatics/btag253
Original Paper