*Bioinformatics*, 2026, **42**(2), btag046
https://doi.org/10.1093/bioinformatics/btag046
Published: 27 January 2026
**Original Paper | Systems biology**

OXFORD

# X-intNMF: a cross- and intra-omics regularized NMF framework for multi-omics integration

**Tien-Thanh Bui**[1,2], **Rui Xie**[3,4], **Wei Zhang**[1,2,*]

[1]Department of Computer Science, University of Central Florida, Orlando, FL 32816, United States
[2]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, United States
[3]Department of Statistics and Data Science, University of Central Florida, Orlando, FL 32816, United States
[4]College of Nursing, University of Central Florida, Orlando, FL 32816, United States
*Corresponding author. Department of Computer Science, University of Central Florida, Orlando, FL 32816, United States. E-mail: wzhang.cs@ucf.edu
Associate Editor: Anthony Mathelier

## Abstract

***Motivation:*** **The rapid accumulation of multi-omics data presents a valuable opportunity to advance our understanding of complex diseases and biological systems, driving the development of integrative computational methods. However, the complexity of biological processes, spanning multiple molecular layers and involving intricate regulatory interactions, requires models that can capture both intra- and cross-omics relationships. Most existing integration methods primarily focus on sample-level similarities or intra-omics feature interactions, often neglecting the interactions across different omics layers. This limitation can result in the loss of critical biological information and suboptimal performance. To address this gap, we propose X-intNMF, a network-regularized non-negative matrix factorization (NMF) framework that simultaneously integrates intra- and cross-omics feature interactions into a shared low-dimensional representation (see Fig. 1). By modeling these multi-layered relationships, X-intNMF enhances the representation of biological interactions and improves integration quality and prediction accuracy.**

***Results:*** **For evaluation, we applied X-intNMF to predict breast cancer phenotypes and classify clinical outcomes in lung and ovarian cancers using mRNA expression, microRNA expression, and DNA methylation data from TCGA. The results show that X-intNMF consistently outperforms state-of-the-art methods. Ablation studies confirm that incorporating both cross-omics and intra-omics interactions contributes significantly to the model's improved performance. Additionally, survival analysis on 25 TCGA cancer datasets demonstrates that the integrated multi-omics representation offers strong prognostic value for both overall survival and disease-free status. These findings highlight X-intNMF's ability to effectively model multi-layered molecular interactions while maintaining interpretability, robustness, and scalability within the NMF framework.**

***Availability and implementation:*** **The source code and datasets supporting this study are publicly available at GitHub (https://github.com/compbiolabucf/X-intNMF) and archived on Zenodo (https://doi.org/10.5281/zenodo.18238385).**

## 1 Introduction

Advancements in high-throughput sequencing technologies have enabled the large-scale collection of omics data across diverse molecular layers, with each modality capturing a distinct aspect of biological regulation. However, the inherently multi-layered and complex nature of biological systems means that analyzing any single omics type in isolation often leads to an incomplete understanding. To gain a more comprehensive and accurate view of biological processes, it is essential to integrate data from multiple omics layers. As a result, multi-omics integration has emerged as a critical area in biomedical research, aiming to combine heterogeneous molecular data to better characterize biological systems (Kang *et al.* 2022). This integrative strategy not only deepens our insight into biological complexity but also enables a wide range of biomedical applications, such as precision oncology, personalized medicine, and microbiome research (Abdelaziz *et al.* 2024, Acharya and Mukhopadhyay 2024).

Despite the advantages of multi-omics integration, several key challenges remain, including the high dimensionality within individual omics layers, heterogeneity across data distributions, limited interpretability, and scalability issues when processing multimodal data. To address these challenges, a wide range of computational approaches have been developed, including kernel-based methods, probabilistic models (Argelaguet *et al.* 2020), matrix

**Received:** 2 July 2025. **Revised:** 15 January 2026. **Accepted:** 16 January 2026

factorization techniques (Cai *et al.* 2011, Pierre-Jean *et al.* 2022), and deep learning frameworks (Ahmed *et al.* 2021, Wang *et al.* 2021, Chen *et al.* 2024). Matrix factorization and probabilistic approaches are commonly used for their simplicity and interpretability, but they often suffer from high computational costs and limited scalability. In contrast, deep learning methods have demonstrated strong performance in multi-omics integration tasks but typically require large training datasets and often lack transparency and interpretability. Furthermore, many existing methods are designed for specific omics types, limiting their generalizability and flexibility when applied to heterogeneous datasets. Notably, biological interaction networks, particularly cross-omics interactions such as mRNA–miRNA regulatory relationships, are frequently neglected. Overlooking these interactions can lead to substantial information loss and diminished model performance, despite their critical role in understanding complex regulatory mechanisms (Yang *et al.* 2024).

Matrix factorization, which decomposes an input matrix into the product of two or more low-rank matrices, is a powerful and widely used approach for uncovering interpretable patterns in complex biological data. A well-known variant is non-negative matrix factorization (NMF) (Lee and Seung 1999), which enforces non-negativity constraints on the factorized matrices. These constraints align naturally with biological data, such as gene expression levels, and enhance the interpretability of the learned components. NMF has been successfully extended to multi-omics integration, resulting in models such as iNMF (Yang and Michailidis 2016), IntNMF (Chalise and Fridley 2017), PIntMF (Pierre-Jean *et al.* 2022), GNMF (Cai *et al.* 2011), and iGMFNA (Gao *et al.* 2019). These approaches aim to jointly analyze diverse omics layers and, in some cases, incorporate prior biological knowledge through interaction networks. However, most existing models focus only on intra-omics relationships, such as within-layer feature correlations or sample similarities, while overlooking cross-omics feature interactions. This limitation prevents them from fully capturing the regulatory complexity spanning different molecular layers.

To address this gap, we propose X-intNMF, a network-regularized NMF framework that simultaneously integrates intra- and cross-omics feature interactions into a shared low-dimensional representation. X-intNMF leverages feature–feature interaction networks to model both within-layer and between-layer relationships, enabling the incorporation of known biological interactions such as mRNA–miRNA regulatory links. Sparsity constraints are imposed on omic-specific factor matrices to enhance interpretability, while the model's scalable and computationally efficient design ensures applicability to large, high-dimensional multi-omics datasets, including mRNA, miRNA, and DNA methylation data.

# 2 Materials and methods

This section introduces the X-intNMF framework, including the model formulation, optimization strategy, and the construction of cross- and intra-omics feature interaction networks. It also outlines the evaluation methods, baseline models, performance metrics, and parameter tuning process. All mathematical notations are summarized in Table S1, available as supplementary data at *Bioinformatics* online.

## 2.1 Model description

Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(D)}$ denote the input from $D$ omics layers, where $\mathbf{X}^{(d)} \in \mathbb{R}^{m_d \times N}$, with $m_d$ representing the number of features in the $d$th omics layer and $N$ denoting the number of samples. The proposed model is based on NMF (Lee and Seung 1999), which aims to identify a shared non-negative sample factor matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ that captures sample-level factors, and $D$ non-negative omic-specific factor matrices $\mathbf{W}^{(d)} \in \mathbb{R}_+^{m_d \times K}$ that project each omics layer into a $K$-dimensional latent space. Here, $K$ is the pre-defined number of latent components. NMF seeks to optimize these matrices under the following objective:

$$\mathbf{X}^{(d)} \approx \mathbf{W}^{(d)}\mathbf{H}, \quad \text{s.t.} \quad \mathbf{W}^{(d)} \in \mathbb{R}_+^{m_d \times K}, \mathbf{H} \in \mathbb{R}_+^{K \times N} \tag{1}$$

To ensure the interpretability of the model, sparsity constraints are imposed on both the sample factor matrix and the omic-specific factor matrices. These constraints promote more focused and biologically meaningful latent representations, resulting in the following formulation:

$$\min_{\mathbf{H},\mathbf{W}^{(d)}} \left( \frac{1}{2} \sum_{d=1}^{D} \left|\left| \mathbf{X}^{(d)} - \mathbf{W}^{(d)}\mathbf{H} \right|\right|_F^2 + \sum_{d=1}^{D} \beta_d \left|\left| \mathbf{W}^{(d)} \right|\right|_1 + \sum_{i=1}^{N} \gamma_i \left|\left| \mathbf{H}_{\cdot,i} \right|\right|_1 \right) \tag{2}$$

where $\beta_d$ and $\gamma_i$ are sparsity regularization parameters. The notation $||\cdot||_F$ denotes the Frobenius norm, while $||\cdot||_1$ refers to the $l_1$ norm, and $\mathbf{H}_{\cdot,i}$ represents the $i$th column, corresponding to the $i$th sample, of the sample factor matrix $\mathbf{H}$. However, standard NMF captures only internal structure within each omics layer and does not account for interactions within or between omics layers. To address this limitation, we propose an all omics feature–feature interaction network $\mathbf{A}$, defined as a $D \times D$ block matrix in Equation (3). The detailed procedure for constructing $\mathbf{A}$ matrix is described in Section 2.3

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{(11)} & \mathbf{A}^{(12)} & \ldots & \mathbf{A}^{(1D)} \\ \mathbf{A}^{(21)} & \mathbf{A}^{(22)} & \ldots & \mathbf{A}^{(2D)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}^{(D1)} & \mathbf{A}^{(D2)} & \ldots & \mathbf{A}^{(DD)} \end{bmatrix} \in \{0,1\}^{M \times M} \tag{3}$$

where $\mathbf{A}^{(pq)}$ denotes the adjacency matrix representing the interaction network between the $p$th and $q$th omics layers, and $M = \sum_d m_d$ represents the total number of features across all omics layers. The corresponding normalized adjacency matrix is given by $\tilde{\mathbf{A}} = \mathbf{\Delta}^{-1/2}\mathbf{A}\mathbf{\Delta}^{-1/2}$, where $\mathbf{\Delta}^{-1/2}$ is the element-wise inverse square root of the degree matrix $\mathbf{\Delta}$, defined as:

$$\mathbf{\Delta}^{-1/2} = \mathrm{diag}\left( \left\{ \frac{1}{\sqrt{\sum_{j=1}^{M} \mathbf{A}_{ij}}} \right\}_{i=1}^{M} \right) \tag{4}$$

To incorporate this information into the model, a Laplacian regularization term $\mathrm{Tr}(\overline{\mathbf{W}}^T \mathbf{L} \overline{\mathbf{W}})$ is introduced, where $\mathbf{L}$ is the normalized graph Laplacian matrix of the interaction network $\mathbf{A}$, defined as:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}^{(11)} & \mathbf{L}^{(12)} & \dots & \mathbf{L}^{(1D)} \\ \mathbf{L}^{(21)} & \mathbf{L}^{(22)} & \dots & \mathbf{L}^{(2D)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}^{(D1)} & \mathbf{L}^{(D2)} & \dots & \mathbf{L}^{(DD)} \end{bmatrix} = \mathbf{I} - \tilde{\mathbf{A}} \tag{5}$$

and $\overline{\mathbf{W}}$ denotes the concatenation of all omic-specific factor matrices $\mathbf{W}^{(d)}$ along the feature axis (i.e. stacked vertically across $d$).

By the definition of the trace operator and the block structure of $\mathbf{L}$, which follows the same $D \times D$ block partitioning as the interaction matrix $\mathbf{A}$, the Laplacian regularization term can be reformulated as:

$$\mathrm{Tr}(\overline{\mathbf{W}}^T \mathbf{L} \overline{\mathbf{W}}) = \sum_{p=1}^{D} \sum_{q=1}^{D} \mathrm{Tr}(\mathbf{W}^{(p)T} \mathbf{L}^{(pq)} \mathbf{W}^{(q)}) \tag{6}$$

By partitioning the interaction network into blocks, interactions between omics layers can be computed directly from the data or derived from prior knowledge when available (e.g. mRNA–miRNA interaction networks). This design enhances the model's robustness and adaptability across diverse omics datasets. By combining the NMF formulation with sparsity constraints (Equation (2)) and graph Laplacian regularization (Equation (6)), the final objective function of the model is defined as:

$$\begin{aligned} \min_{\mathbf{H},\mathbf{W}^{(d)}} f = \min_{\mathbf{H},\mathbf{W}^{(d)}} & \frac{1}{2} \sum_{d=1}^{D} \left\| \mathbf{X}^{(d)} - \mathbf{W}^{(d)} \mathbf{H} \right\|_F^2 \\ & + \frac{\alpha}{2} \sum_{p=1}^{D} \sum_{q=1}^{D} \mathrm{Tr}(\mathbf{W}^{(p)T} \mathbf{L}^{(pq)} \mathbf{W}^{(q)}) \\ & + \sum_{d=1}^{D} \beta_d \left\| \mathbf{W}^{(d)} \right\|_1 + \sum_{i=1}^{N} \gamma_i \left\| \mathbf{H}_{\cdot,i} \right\|_1 \end{aligned} \tag{7}$$

This objective function consists of four key components: the reconstruction error term, the cross- and intra-omics network regularization term, the omic-specific sparsity regularization term, and the sample-specific sparsity regularization term.

## 2.2 Optimization strategy and convergence criteria

The objective function in Equation (7) is not jointly convex in all variables $\mathbf{W}^{(d)}$ and $\mathbf{H}$ and therefore must be optimized iteratively by updating one matrix while keeping the others fixed. The algorithm alternately updates the variables $\mathbf{W}^{(d)}$ and $\mathbf{H}$ until either a maximum number of iterations is reached, or a convergence criterion is met. The convergence criterion is defined as the absolute change in the objective function falling below a pre-defined threshold. Further details on the iterative update formulas are provided in Section 2, available as supplementary data at *Bioinformatics* online.

For initialization, each $\mathbf{W}^{(d)}$ is first initialized using Nonnegative Double Singular Value Decomposition (NNDSVD) as in standard NMF. The sample factor matrix $\mathbf{H}$ is then initialized based on the initialized $\mathbf{W}^{(d)}$ using the Lasso-based update rule from PIntMF (Pierre-Jean *et al.* 2022). When $\mathbf{W}^{(d)}$ is fixed, the objective function in Equation (7) reduces to Equation (8), which is solved with respect to each sample vector $\mathbf{H}_{\cdot,i}$. This results in an independent Lasso problem for each $i = 1, \dots, N$

$$\begin{aligned} \min_{\mathbf{H}_{\cdot,i}} g &= \min_{\mathbf{H}_{\cdot,i}} \frac{1}{2} \sum_{d=1}^{D} \left\| \mathbf{X}^{(d)} - \mathbf{W}^{(d)} \mathbf{H} \right\|_F^2 + \sum_{i=1}^{N} \gamma_i \left\| \mathbf{H}_{\cdot,i} \right\|_1 + \text{const} \\ &= \min_{\mathbf{H}_{\cdot,i}} \sum_{i=1}^{N} \left( \frac{1}{2} \left\| \overline{\mathbf{X}} - \overline{\mathbf{W}} \mathbf{H}_{\cdot,i} \right\|_F^2 + \gamma_i \left\| \mathbf{H}_{\cdot,i} \right\|_1 \right) + \text{const} \end{aligned} \tag{8}$$

$\overline{\mathbf{X}}$ denotes the concatenation of all omics layers along the feature axis.

## 2.3 Intra- and cross-omics interaction network construction

The interaction network $\mathbf{A}$ is structured as a block matrix, where each block $\mathbf{A}^{(pq)}$ is a binary submatrix representing interactions between omics layers $p$ and $q$. This flexible design allows known interactions to be incorporated directly when available. In cases where computation is necessary, the interactions for a given omics pair $(p, q)$, whether intra- or cross-omics, are computed using the absolute value of the Pearson correlation coefficient (PCC) across all samples and then thresholded to match the average density of the existing interaction networks. This approach ensures that the resulting interaction network remains consistent with existing cross-omics interaction data, thereby preserving the biological relevance of the interactions.

## 2.4 Evaluation methods

The performance of X-intNMF is evaluated based on its ability to classify disease outcomes, assess survival or disease-free time and analyze cancer subtypes using the learned sample factor matrix $\mathbf{H}$. The underlying assumption is that higher-quality representations derived from the multi-omics data in $\mathbf{H}$ will lead to improved predictive performance in downstream analyses.

### 2.4.1 Classification

In the binary classification tasks, the model is evaluated on two subtasks: (1) disease phenotype prediction and (2) classification of patients into pre-defined survival or disease-free duration categories (e.g. short-term versus long-term survival). Performance is assessed using two commonly used metrics: area under the receiver operating characteristic curve (AUC) and Matthews correlation coefficient (MCC). AUC evaluates the model's ability to distinguish between positive and negative classes across various threshold settings, making it robust to class imbalance. MCC considers all four confusion matrix categories, true positives, true negatives, false positives, and false negatives, and provides a more balanced and informative measure, particularly in imbalanced datasets.

### 2.4.2 Survival analysis

Beyond the binary classification of cancer phenotype and survival duration range, a separate survival analysis is conducted using a Cox proportional hazards model with an Elastic Net

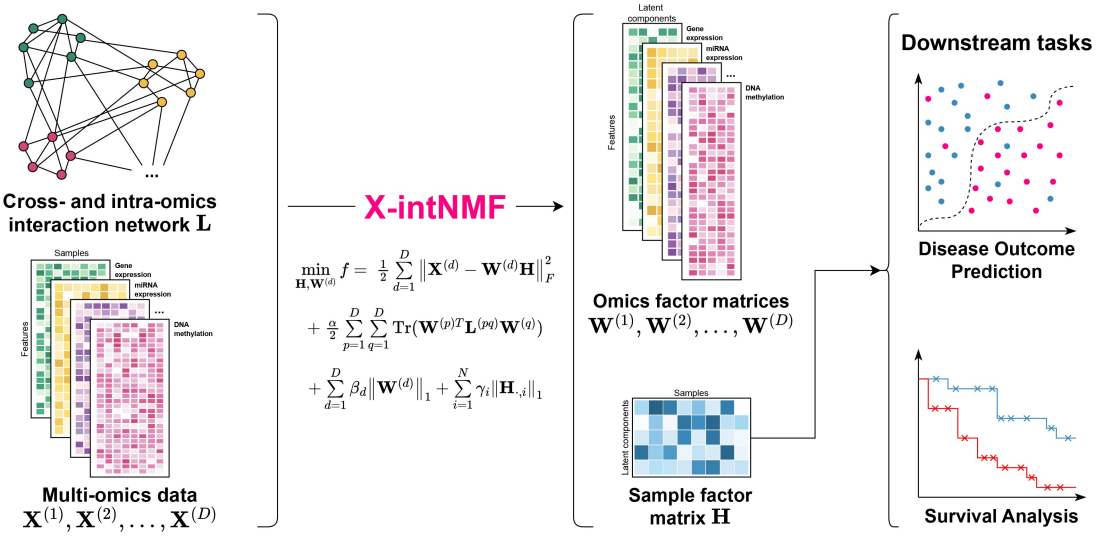


**Figure 1** Overview of the proposed X-intNMF framework. X-intNMF integrates multi-omics data with both cross-omics and intra-omics interaction networks to jointly learn a shared sample factor matrix **H** and omic-specific factor matrices $\mathbf{W}^{(d)}$ via NMF with graph-based regularization. The resulting low-dimensional representations, **H** for samples and $\mathbf{W}^{(d)}$ for omics features, are then utilized for downstream analyses, including classification and survival prediction.

penalty (Simon *et al.* 2011). This analysis aims to investigate the correlation between a patient's overall survival, overall disease-free status, and the learned sample factor matrix **H**. The model optimizes the log-likelihood function $L(\lambda)$ while incorporating both $l_1$-norm and $l_2$-norm penalties on the regression coefficients $\lambda$, resulting in the following optimization problem:

$$\log L(\lambda) - \xi\left(\eta||\lambda||_1 + \frac{1-\eta}{2}||\lambda||_2^2\right) \tag{9}$$

where $\xi \geq 0$ is the shrinkage control parameter, and $\eta \in [0,1]$ is the mixing parameter that balances the $l_1$-norm and $l_2$-norm penalties. After training, the estimated coefficient $\lambda$ is applied to the test set to compute the prognostic index (**PI**) for each patient:

$$\mathbf{PI} = \lambda^T \mathbf{H}_{\text{test}} \tag{10}$$

where $\mathbf{H}_{\text{test}}$ represents the dimensionality-reduced learned sample factor matrix of the test set. The median value of **PI** is then used as a threshold to classify patients into high-risk and low-risk groups. To evaluate the predictive performance, Kaplan–Meier survival curves are generated, and log-rank tests are performed to assess the statistical significance of survival or disease-free status differences between the high-risk and low-risk groups.

### 2.4.3 Cancer subtype analysis

Besides binary classification and survival analysis, the model is further evaluated based on its ability to cluster patients into different cancer subtypes. First, a sample-by-sample absolute Pearson correlation matrix is computed from the learned sample factor matrix **H**. Then, hierarchical clustering is performed to identify subtype-specific patterns. All clustering procedures are implemented using the `scipy` and `seaborn` libraries.

## 2.5 Baselines

X-intNMF is compared with the following baselines, using their default configurations:

- MOFA2 (Argelaguet *et al.* 2020) is a Bayesian model for unsupervised multi-omics data integration.
- iNMF (Yang and Michailidis 2016) is a NMF-based multi-omics integration method using a partitioned factorization structure that captures homogeneous and heterogeneous effects on multi-modal data.
- iGMFNA (Gao *et al.* 2019) is a NMF-based multi-omics integration method which integrating multi-omics data with only intra-omics interactions information.
- MOGONET (Wang *et al.* 2021) is a supervised deep learning model for multi-omics data integration. It employs a View Correlation Discovery Network to capture cross-omics correlations in the label space. Unlike X-intNMF, MOGONET is supervised and does not incorporate known external cross-omics interaction networks.
- MOMA (Moon and Lee 2022) is a supervised multi-task attention learning algorithm that uses a geometric approach to identify important modules in multi-omics data through an attention mechanism. The output probabilities of the two prediction tasks are averaged to generate the final classification.
- MCRGCN (Chen *et al.* 2024) is a supervised graph contrastive learning model for cancer subtype identification. It leverages the unique feature distributions of each omics layer and their interactions to enhance classification accuracy.

## 2.6 Parameter tuning

The framework has four hyperparameters: the interaction network regularization parameter $\alpha$, the number of latent components $K$, the omic-specific sparsity regularization parameters $\beta_d$,

and the sample-specific sparsity regularization parameters $\gamma_i$. The parameters $\gamma_i$ are computed directly during the initialization of **H** using Lasso cross-validation (`LassoCV`) from the *scikit-learn* package. The remaining parameters, $K$, $\alpha$, and $\beta_d$, are tuned via grid search in conjunction with a specific downstream task. In this study, hyperparameter optimization is performed based on the classification task with logistic regression used as the downstream classifier. The resulting **H** from the best-performing configuration selected from the lists below is then reused for additional downstream analyses, including survival/disease-free outcome prediction and cancer subtype analysis.

- $K$: 10, 25, 50, 100, 200
- $\alpha$: 0, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
- $\beta_d$: 1, 0.1, 0.01, 10, 100 (applied uniformly for all $d$)

For each classification target, 100 test cases are generated by randomly splitting the data into 80% training and 20% testing subsets. Within each test case, five-fold cross-validation is performed on the training portion to evaluate all hyperparameter combinations. For each combination, the AUC scores across the five folds are averaged, and the set of parameters yielding the highest average AUC is selected. These optimal parameters are then used to train the model on the full training set of that test case, and evaluation, including AUC and MCC, is performed on the corresponding test set. For the baseline models, since MOFA2, iNMF, and iGMFNA produce latent representations, their classification setups were designed to match that of the proposed model.

Regarding $\xi$ and $\eta$ in the Cox proportional hazards model, $\eta$ is fixed at 0.5, while $\xi$ is tuned using five-fold cross-validation on the training set.

# 3 Results

The performance of X-intNMF is evaluated through experiments on The Cancer Genome Atlas (TCGA) datasets, incorporating multi-omics profiles and interaction networks. This section begins with a detailed description of the datasets and data preprocessing procedures. Next, classification results are presented for both cancer phenotype prediction and survival duration classification, along with comparisons to baseline models. An ablation study is then conducted to assess the contribution of the interaction network to model performance. Additionally, results from survival and disease-free outcome analyses, as well as breast cancer subtyping, are reported to further demonstrate the effectiveness of the framework. Finally, the model's scalability is evaluated through experiments in a three omics setting.

## 3.1 Data preparation

X-intNMF was primarily evaluated using three TCGA datasets: breast cancer (BRCA), lung adenocarcinoma (LUAD), and ovarian cancer (OV). The omics layers included mRNA expression, miRNA expression, and DNA methylation data, all obtained from the UCSC Xena Hub (Goldman *et al.* 2020). The mRNA–miRNA interaction network was derived from TargetScanHuman (Agarwal *et al.* 2015). The model was assessed under two experimental settings: (1) a two omics setting combining mRNA and miRNA expression and (2) a three omics setting that additionally incorporated DNA methylation. Furthermore, to comprehensively evaluate the model's performance, 22 additional TCGA pan-cancer datasets were analyzed for overall survival and disease-free survival outcomes. Detailed descriptions of these datasets are provided in Table S6, available as supplementary data at *Bioinformatics* online.

All omics layers undergo preprocessing steps that include: removing samples not present across all omics layers, eliminating features with zero variance or all-zero values, and discarding mRNA/miRNA features not found in the interaction network. Due to the high dimensionality of the DNA methylation data, its features are further reduced by selecting the most variable features, limited to match the number of mRNA features after preprocessing. All retained samples are normalized by row-wise division using the maximum value in each row, scaling values to the $[0, 1]$ range. The dimensions of the processed omics layers are summarized in Table 1.

For the breast cancer dataset, the classification targets include four phenotypes: estrogen receptor (ER+/ER−), progesterone receptor (PR+/PR−), human epidermal growth factor receptor 2 (HER2+/HER2−), and triple-negative breast cancer (TN versus non-TN). For the lung cancer and ovarian cancer datasets, the classification targets are survival and disease-free duration categories. Detailed information on the datasets and classification targets is provided in Table 2.

## 3.2 X-intNMF outperforms baselines in breast cancer phenotype classification

Table 3 summarizes the breast cancer phenotype classification results across competing methods. X-intNMF outperforms most baselines in both AUC and MCC across all four targets (ER, PR, HER2, and TN). In particular, X-intNMF achieves consistently high AUC values for ER, PR, and TN, demonstrating its robust ability to distinguish between positive and negative cases. Even for HER2, typically the most challenging target where many models underperform, X-intNMF achieves a strong AUC of 0.7656, outperforming all baseline methods.

In addition to its strong AUC performance, X-intNMF also achieves consistently high MCC values on PR and HER2, reflecting a better balance between sensitivity and specificity. Although the MCC values are slightly lower than some baselines for ER and TN, the AUC results remain competitive. This improvement can be partly attributed to the effective integration of cross- and intra-omics interaction networks, which provide biologically meaningful context and enhance the quality of the learned representations.

## 3.3 X-intNMF demonstrates robust performance in predicting survival-related duration range

In addition to cancer phenotype classification, X-intNMF is also evaluated on the task of predicting survival and disease-free duration categories for the lung and ovarian cancer datasets. Tables 4 and 5 present the results for the lung and ovarian cancer datasets, respectively.

X-intNMF consistently outperforms baseline models in both survival and disease-free classification tasks for the two datasets, demonstrating strong performance in terms of both AUC and MCC, highlighting its robustness and accuracy in prognosis classification.

**Table 1** Number of features and samples after preprocessing for all datasets.

| | BRCA | LUAD | OV |
|---|---|---|---|
| Feature# | | | |
| mRNA | 10 400 | 10 400 | 10 400 |
| miRNA | 277 | 277 | 274 |
| DNAm | 10 400 | 10 400 | 10 400 |
| Sample# | | | |
| 2-omics | 830 | 466 | 304 |
| 3-omics | 674 | 456 | 295 |

The counts of mRNA, microRNA, and DNA methylation features are denoted as mRNA, miRNA, and DNAm, respectively.

**Table 2** Number of samples for each classification target.

| Dataset | Target | 2-omics Sample# | 3-omics Sample# |
|---|---|---|---|
| BRCA | Phenotype: ER+/ER− | 185/54 | 153/45 |
| | Phenotype: PR+/PR− | 159/80 | 130/68 |
| | Phenotype: HER2+/HER2− | 39/200 | 28/170 |
| | Phenotype: TN/non-TN | 46/193 | 40/158 |
| LUAD | Survival: $\leq 24/\geq 30$ months | 93/134 | 93/134 |
| | Disease-free: $<24/\geq 24$ months | 54/99 | 54/99 |
| OV | Survival: $<18/\geq 18$ months | 40/207 | 38/202 |
| | Disease-free: $\leq 12/\geq 18$ months | 17/71 | 15/68 |

Counts are shown as "left/right," indicating the number of samples in each of the two classes.

**Table 3** Breast cancer phenotype classification results.

| Methods | ER | | PR | | HER2 | | TN | |
|---|---|---|---|---|---|---|---|---|
| | MCC | AUC | MCC | AUC | MCC | AUC | MCC | AUC |
| X-intMF | 0.7615 | **0.9612** | **0.7303** | **0.9068** | **0.2821** | **0.7656** | 0.6887 | **0.9652** |
| MOFA2 | 0.7544 | 0.9471 | 0.6688 | 0.8860 | 0.2318 | 0.6922 | 0.6430 | 0.9471 |
| iNMF | 0.6725 | 0.9373 | 0.5776 | 0.8543 | 0.2235 | 0.6870 | **0.7070** | 0.9606 |
| iGMFNA | **0.7664** | 0.9571 | 0.6734 | 0.8827 | 0.2677 | 0.6984 | 0.7001 | 0.9542 |
| MOMA | 0.6990 | 0.9496 | 0.6613 | 0.9004 | 0.1022 | 0.7366 | 0.6252 | 0.9585 |
| MOGONET | 0.5730 | 0.9076 | 0.4372 | 0.8206 | 0.0134 | 0.6130 | 0.5266 | 0.9176 |
| MCRGCN | −0.0031 | 0.4961 | 0.0017 | 0.4954 | −0.0064 | 0.5051 | 0.0013 | 0.4897 |

Bold values indicate the best performance for each metric.

These results suggest that X-intNMF is a powerful framework for predicting cancer outcomes, effectively leveraging multi-omics data to provide reliable and interpretable predictions.

## 3.4 Impact of cross- and intra-omics feature interaction networks on model performance

To further validate the effectiveness of incorporating cross- and intra-omics feature interaction information into the learned multi-omics sample representations, an ablation study is conducted by setting the interaction network regularization parameter $\alpha$ to zero, which effectively disables the contribution of the interaction network during training. The results, presented in Table 6, illustrate the impact of this modification by comparing the classification performance of the ablated model with that of the full X-intNMF framework. Across all tasks, removing the interaction network leads to a notable decline in performance, particularly in terms of MCC. This decline is especially pronounced in complex and heterogeneous cases, such as HER2 classification in breast cancer and survival duration prediction in ovarian cancer. These findings confirm the importance of modeling cross- and intra-omics interactions in multi-omics data integration, demonstrating that incorporating such relationships enhances classification performance and robustness across diverse clinical contexts.

## 3.5 X-intNMF demonstrates strong capability in survival analysis

In addition to classification tasks, survival analysis is conducted using the learned sample factor matrix **H** to evaluate overall survival

**Table 4** Lung cancer survival and cancer-free duration range classification results.

| Methods | Survival: $\leq 24/\geq 30$ | | Disease-free: $<24/\geq 24$ | |
|---|---|---|---|---|
| | MCC | AUC | MCC | AUC |
| X-intMF | **0.2952** | **0.6832** | **0.3008** | **0.6672** |
| MOFA2 | 0.1812 | 0.6352 | 0.0298 | 0.5911 |
| iNMF | 0.1331 | 0.5982 | −0.0008 | 0.5200 |
| iGMFNA | 0.2431 | 0.6552 | −0.0081 | 0.5621 |
| MOMA | 0.2348 | 0.6802 | 0.0184 | 0.5570 |
| MOGONET | 0.0692 | 0.5676 | 0.0084 | 0.5202 |
| MCRGCN | 0.0026 | 0.5057 | −0.0177 | 0.5008 |

Bold values indicate the best performance for each task.

**Table 5** Ovarian cancer survival and cancer-free duration range classification results.

| Methods | Survival: $<18/\geq 18$ | | Disease-free: $\leq 12/\geq 18$ | |
|---|---|---|---|---|
| | MCC | AUC | MCC | AUC |
| X-intMF | **0.2381** | **0.7268** | **0.4624** | **0.8387** |
| MOFA2 | 0.0164 | 0.5668 | 0.1263 | 0.6940 |
| iNMF | 0.0123 | 0.5606 | 0.0395 | 0.5595 |
| iGMFNA | 0.0291 | 0.5767 | 0.0000 | 0.6450 |
| MOMA | 0.0085 | 0.5971 | 0.0198 | 0.6318 |
| MOGONET | 0.0411 | 0.6259 | −0.0111 | 0.4481 |
| MCRGCN | −0.0035 | 0.5069 | 0.0080 | 0.4838 |

Bold values indicate the best performance for each task.

**Table 6** Classification results from the ablation study.

| Target | X-intNMF | | X-intNMF ($\alpha = 0$) | |
|---|---|---|---|---|
| | MCC | AUC | MCC | AUC |
| Breast cancer (BRCA) | | | | |
| ER | 0.7615 | 0.9612 | 0.7449 | 0.9376 |
| PR | 0.7303 | 0.9068 | 0.7247 | 0.9067 |
| HER2 | 0.2821 | 0.7656 | 0.2668 | 0.7322 |
| TN | 0.6887 | 0.9652 | 0.6740 | 0.9528 |
| Lung cancer (LUAD) | | | | |
| Survival: $\leq 24/\geq 30$ | 0.2952 | 0.6832 | 0.1657 | 0.5916 |
| Disease-free: $\leq 24/\geq 24$ | 0.3008 | 0.6672 | 0.2027 | 0.6438 |
| Ovarian cancer (OV) | | | | |
| Survival: $<18/\geq 18$ | 0.2381 | 0.7268 | 0.1791 | 0.6880 |
| Disease-free: $\leq 12/\geq 18$ | 0.4624 | 0.8387 | 0.0995 | 0.6780 |

and disease-free status, as described in Section 2.4. The analysis is performed on all three datasets and compared against both the ablated model (without the interaction network) and the concatenated original input data. Figures 2 and 3 show the Kaplan–Meier survival curves for the lung cancer dataset, illustrating differences in survival and disease-free status between high-risk and low-risk groups. X-intNMF achieves the most statistically significant log-rank test *P*-values and demonstrates the clearest separation between risk groups, outperforming both baselines in both analyses. Additional results for the breast and ovarian cancer datasets are provided in Figs S1–S10, available as supplementary data at *Bioinformatics* online. These results highlight the robustness and clinical relevance of the learned representations. To further validate the effectiveness of the model, comparative survival analyses across 25 TCGA pan-cancer datasets against multiple baseline methods were conducted. The detailed results are

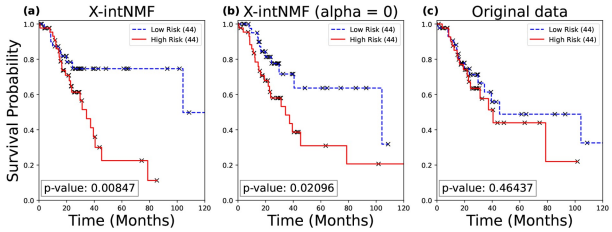


**Figure 2** Kaplan–Meier survival analysis for overall survival in lung cancer. The input data for the Cox model in (a) is the sample factor matrix **H** learned by the proposed framework; (b) uses the sample factor matrix **H** from the ablated model with $\alpha = 0$; and (c) uses the original input data.

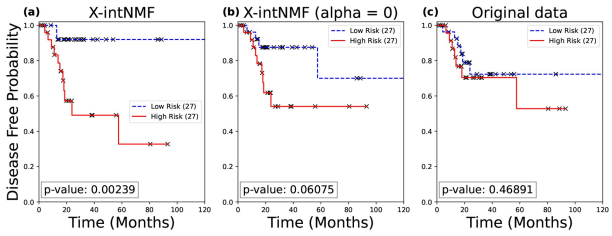


**Figure 3** Kaplan–Meier survival analysis for disease-free status in lung cancer. The input data for the Cox model in (a) is the sample factor matrix **H** learned by the proposed framework; (b) uses the sample factor matrix **H** from the ablated model with $\alpha = 0$; and (c) uses the original input data.

provided in Tables S7 and S8, available as supplementary data at *Bioinformatics* online.

## 3.6 X-intNMF effectively classifies breast cancer subtypes

In addition to the classification and survival analysis, the proposed framework is further evaluated for its ability to capture biologically meaningful subtypes using the breast cancer dataset. The subtype information, including Luminal A, Luminal B, HER2-enriched, and Basal-like, is obtained from the original TCGA study (The Cancer Genome Atlas Network 2012). Figure 4 presents the clustermap of the absolute Pearson correlation matrix computed from the learned sample factor matrix **H**. The clustering patterns in the heatmap reveal that patient samples with the same subtype labels tend to group together, indicating strong within-subtype correlation. This trend is especially pronounced for the Basal-like subtype, which forms a distinct cluster. These results demonstrate that the learned low-dimensional representation from multi-omics data by X-intNMF can effectively capture subtype-specific signals.

## 3.7 Scalability of X-intNMF beyond two omics modalities

To evaluate the framework's compatibility and scalability, it is also tested using three omics layers, mRNA, miRNA, and DNA methylation, across all three datasets for both classification and survival analysis. Table 7 presents the classification results for

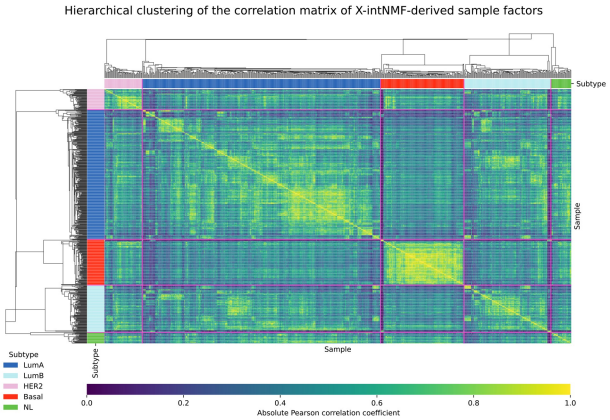


**Figure 4** Clustermap of the absolute Pearson correlation matrix computed from the learned sample factor matrix **H**. The pink line marks the boundary between subtype labels. The subtype annotation bar indicates five subtypes: Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2), Basal-like (Basal), and Normal (NL).

survival and disease-free outcomes in the lung cancer dataset, using the same label definitions as in the two omics setting (Table 2). The results demonstrate that the proposed framework continues to outperform baseline models in terms of both AUC and MCC, indicating that X-intNMF scales effectively to more complex multi-omics settings. Notably, MCRGCN yields near-random predictions on several targets in both the two- and three omics settings, further underscoring the difficulty of the task and the relative strength of X-intNMF. Additional results and label definitions for the breast and ovarian cancer datasets are provided in Tables S2 and S3, available as supplementary data at *Bioinformatics* online.

# 4 Discussion

The results confirm that incorporating cross- and intra-omics feature interaction networks improves model performance across classification, survival analysis, and cancer subtyping tasks, underscoring the importance of capturing meaningful biological relationships. Moreover, the framework demonstrates strong compatibility with more than two omics layers, effectively scaling to three modalities and potentially beyond. However, hyperparameter selection remains a challenge. The current grid search approach requires extensive computational resources to identify optimal settings. Alternative optimization algorithms, such as particle swarm optimization (Bonyadi and Michalewicz 2017), could reduce tuning time. Further improvements in data preprocessing and initialization strategies may also help stabilize learning dynamics and enhance model performance, robustness, and interpretability across diverse datasets.

While the framework achieves strong predictive performance, certain computational limitations persist. Objective function trajectories recorded via MLFlow (see Fig. S11, available as supplementary data at *Bioinformatics* online) show that the model converges rapidly during the first 1000 iterations, but the convergence rate slows significantly thereafter. This suggests that the model may benefit from adaptive parameter schedules to accelerate convergence. Whether continued training beyond

**Table 7** Lung cancer survival and cancer-free duration range classification results using the 3-omics setting.

| Methods | Survival: $\leq 24/\geq 30$ | | Disease-free: $< 24/\geq 24$ | |
|---|---|---|---|---|
| | MCC | AUC | MCC | AUC |
| X-intMF | **0.2323** | **0.6438** | **0.1808** | 0.6546 |
| $\mathrm{X-intNMF}_{\alpha=0}$ | 0.1182 | 0.5585 | 0.1803 | **0.6549** |
| MOGONET | 0.1193 | 0.5820 | 0.0216 | 0.5456 |
| MCRGCN | −0.0217 | 0.5053 | 0.0091 | 0.4950 |
| MOFA2 | 0.0994 | 0.6102 | 0.0872 | 0.5972 |
| iNMF | 0.1193 | 0.5804 | −0.0374 | 0.5061 |
| iGMFNA | 0.2020 | 0.6018 | 0.0631 | 0.5838 |

Bold values indicate the best performance for each task.

this point yields meaningful performance gains remains an open question. While GPU acceleration greatly enhances runtime efficiency, the optimization process may remain slow for large-scale datasets or when computations are limited to CPU execution. This bottleneck could potentially be addressed using methods such as the alternating direction method of multipliers (ADMM) (Gabay and Mercier 1976), which may offer faster convergence and improved numerical stability.

The current construction of the cross- and intra-omics interaction network uses Pearson correlation with a fixed threshold for omics pairs lacking prior knowledge. However, more tailored strategies may be appropriate. Future work could explore adaptive interaction functions and thresholds specific to each omic–omic relationship, both within and across omics layers, to better capture the unique characteristics of each data modality. Similarly, the omic-specific sparsity regularization parameters $\beta_d$ are currently set uniformly across layers; individualized tuning per omics type could further enhance performance. Another limitation is the exclusion of samples not present in all omics layers, which restricts the framework's applicability. Future research should investigate methods for integrating such samples, including imputation techniques or approaches that explicitly handle missing data, enabling more inclusive and comprehensive analyses.

Lastly, in addition to using the sample factor matrix $\mathbf{H}$ for main downstream tasks, the omic-specific factor matrices $\mathbf{W}^{(d)}$ can also be utilized for interpretability analyses, including gene set enrichment analyses. Further details are provided in Section 8 and Table S9, available as supplementary data at *Bioinformatics* online. Moreover, incorporating other data modalities, such as clinical records or imaging data, could broaden the framework's utility, making it a more powerful and comprehensive tool for multi-omics integration.

# 5 Conclusion

X-intNMF is a novel framework for multi-omics data analysis that integrates NMF with cross- and intra-omics feature interaction network regularization for cancer outcome classification, survival duration prediction, and cancer subtyping. The interaction network can be constructed using either prior biological knowledge or data-driven approaches, supporting both intra- and cross-omics relationships. This design enables the model to capture complex interdependencies across omics layers while maintaining the flexibility to adapt to diverse datasets. The framework is evaluated across multiple TCGA cancer datasets, demonstrating consistently improved performance compared to existing approaches. Results further confirm that X-intNMF is compatible with and scalable to three omics layers, yielding meaningful insights in survival analysis and cancer subtype identification. Future work could explore leveraging the omic-specific factor matrices, investigate alternative initialization strategies, improve hyperparameter tuning through genetic algorithms, and evaluate the impact of different interaction functions and network density settings. Overall, X-intNMF provides a flexible, interpretable, and effective solution for integrative multi-omics analysis in cancer research.

# Author contributions

Tien-Thanh Bui (Conceptualization [equal], Data curation [equal], Formal analysis [lead], Methodology [equal], Software [lead], Validation [lead], Writing—original draft [equal]), Rui Xie (Investigation [supporting], Methodology [supporting], Supervision [supporting], Writing—original draft [supporting]), and Wei Zhang (Conceptualization [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Resources [equal], Supervision [lead], Writing—original draft [equal])

# Supplementary material

Supplementary material is available at *Bioinformatics* online.

# Conflicts of interest

None declared.

# Funding

This work was supported by grants from the National Science Foundation (NSF) [NSF-III2246796 and NSF-III2152030].

# References


Abdelaziz EH, Ismail R, Mabrouk MS *et al.* Multi-omics data integration and analysis pipeline for precision medicine: systematic review. *Comput Biol Chem* 2024;**113**:108254.

Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Brief Funct Genomics* 2024;**23**:549–60.

Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 2015;**4**:e05005.

Ahmed KT, Sun J, Cheng S *et al.* Multi-omics data integration by generative adversarial network. *Bioinformatics* 2021;**38**:179–86.

Argelaguet R, Arnol D, Bredikhin D *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111.

Bonyadi MR, Michalewicz Z. Particle swarm optimization for single objective continuous space problems: a review. *Evol Comput* 2017;**25**:1–54.

Cai D, He X, Han J *et al.* Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell* 2011;**33**:1548–60.

Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One* 2017;**12**:e0176278.

Chen F, Peng W, Dai W *et al.* Supervised graph contrastive learning for cancer subtype identification through multi-omics data integration. *Health Inf Sci Syst* 2024;**12**:12.

Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput Math Appl* 1976;**2**:17–40.

Gao Y-L, Hou M-X, Liu J-X *et al.* An integrated graph regularized non-negative matrix factorization model for gene co-expression network analysis. *IEEE Access* 2019;**7**:126594–602.

Goldman MJ, Craft B, Hastie M *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020;**38**:675–8.

Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* 2022;**23**:bbab454.

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**:788–91.

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

Pierre-Jean M, Mauger F, Deleuze J-F *et al.* PIntMF: penalized integrative matrix factorization method for multi-omics data. *Bioinformatics* 2022;**38**:900–7.

Simon N, Friedman JH, Hastie T *et al.* Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw* 2011;**39**:1–13.

The Cancer Genome Atlas Network.Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.

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.

Yang T-H, Chen J-C, Lee Y-H *et al.* Identifying human miRNA target sites via learning the interaction patterns between miRNA and mRNA segments. *J Chem Inf Model* 2024;**64**:2445–53.

Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2016;**32**:1–8.



Bioinformatics, 2026, 42, 1–10
https://doi.org/10.1093/bioinformatics/btag046
Original Paper