MetabOmics: Metabolism-Oriented Omics Data Integration

Aycan Şahin
Istanbul Technical University
Istanbul, Turkiye
sahinay21@itu.edu.tr

Mehmet Ali Erdoğan
Istanbul Technical University
Istanbul, Turkiye
erdoganmeh25@itu.edu.tr

Utku Sabri Kaya
Istanbul Technical University
Istanbul, Turkiye
kayau17@itu.edu.tr

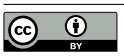
Ali Çakmak*
Istanbul Technical University
Istanbul, Turkiye
ali.cakmak@itu.edu.tr

Abstract

Biological processes arise from complex interactions across multiple molecular layers, yet bridging the gap between disparate omics data types remains a significant challenge. This paper introduces MetabOmics, a comprehensive metabolism-oriented integrated multi-omics analysis method structurally designed to accommodate genomics, transcriptomics, proteomics, and metabolomics datasets. Our methodology centers on the construction of an integrated multi-omic interaction network that incorporates a wide range of biological interactions, including gene expression, translation, transcription factor activity, and post-transcriptional regulation via microRNAs. To capture the cascading effects of molecular changes, we map measured biological entities onto this network and utilize information diffusion models, such as Linear Threshold Diffusion, to propagate fold-changes throughout the system. These propagated measurements are then used to update the lower and upper bounds of metabolic reactions within a genome-scale metabolic model in a personalized manner. Finally, we apply an extended metabolic flux analysis algorithm to compute reaction and pathway differentiation scores.

To demonstrate the empirical efficacy of this framework, we evaluated our approach using paired transcriptomics and metabolomics data across six different cancer cohorts. To further validate the framework's capacity for deep multi-omics integration, we additionally applied MetabOmics to the MayoRNASeq Progressive Supranuclear Palsy (PSP) cohort, successfully integrating transcriptomics, metabolomics, and proteomics. Our results demonstrate that our network-based integration achieves highly competitive classification performance compared to unconstrained multi-omics baselines, and significantly outperforms single-omics approaches. Crucially, we quantitatively establish that MetabOmics produces vastly more stable and biologically concordant feature selections; it achieves robust literature concordance across all evaluated cohorts, whereas simple data concatenation frequently fails to identify disease-relevant pathways.

*Corresponding author.



BCB '26, Rende (CS), Italy

ACM ISBN 979-8-4007-2653-8/26/06
https://doi.org/10.1145/3807503.3819462

Keywords

Multi-omics integration, Metabolic networks, Information diffusion models, Genome-scale metabolic modeling, Cancer metabolism

ACM Reference Format:
Aycan Şahin, Mehmet Ali Erdoğan, Utku Sabri Kaya, and Ali Çakmak. 2026. MetabOmics: Metabolism-Oriented Omics Data Integration. In *17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '26), June 30–July 03, 2026, Rende (CS), Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3807503.3819462

1 Introduction

Biological processes and functions arise from the interactions of tens of thousands of molecules, making them inherently complex. The advent of big data and high-throughput technologies in the area of biological sciences has ushered in an era of unprecedented opportunities and challenges. A key concern is how to effectively connect multiple disparate omic data types to attain a comprehensive understanding of the intricate biological functions operating within organisms. The Big Data to Knowledge (BD2K) grand challenge highlights the need to integrate diverse data types into a cohesive, biologically meaningful structure[4]. The remarkable progress in omics technologies, spanning genomics, transcriptomics, proteomics, metabolomics, and more, has enabled researchers to quantitatively track changes in biological processes with unparalleled detail. Despite these advancements, bridging the gap between different omics data types and linking them to the phenotypic characteristics of organisms remains an elusive task.

Using metabolic models as frameworks for analyzing high-throughput data—such as transcriptomics, proteomics, and metabolomics—allows for the identification of condition-dependent shifts in an organism's metabolic activity. One of the significant challenges in metabolic network modeling is developing computational methods that can predict metabolic flux by integrating these diverse data sources.

Previous studies have utilized Constraint-Based Modeling (CBM) to qualitatively combine high-throughput molecular datasets with metabolic networks. For example, Åkesson et al. (2004) and Becker and Palsson (2008) used gene expression data to identify genes that are absent or likely absent in specific contexts, aiming to uncover metabolic states that hinder or minimize flux through related reactions. Additionally, Shlomi et al. (2008) integrated data on genes with both low and high expression levels to determine the probability of associated reactions carrying metabolic flux.

While metabolomics adds an extra layer of information known as metabolic regulation, transcriptomics and proteomics provide valuable insights into the hierarchical regulation of metabolic flux, which represents the control over enzyme activity.

The presence of extensive multi-omics data has prompted the development of diverse methodologies and algorithms for their integrated analysis. Existing multi-omics methodologies generally fall into data concatenation [19], correlation-based analysis [13], multivariate techniques like PLS [8], pathway-based mappings [25], and network-based approaches [24]. More recently, deep learning architectures, such as multi-omics variational autoencoders, have emerged to capture complex non-linear relationships across layers [1, 28]. While purely data-driven integration methods (e.g., deep learning, data concatenation) achieve high predictive performance, they operate as 'black boxes' lacking mechanistic interpretability. MetabOmics addresses this by acting as a biological regularizer, constraining omics features within a genome-scale metabolic model.

In this paper, we present a novel metabolism-oriented integrated multi-omics data analysis method that addresses these gaps by constructing an interaction network based on genomics, transcriptomics, proteomics, and metabolomics data. Initially, we map the measured omics entities such as genes, proteins, and metabolites onto this network and compute fold-changes to quantify the impact of different conditions. We then propagate these changes through the network using information diffusion models. To personalize the analysis, we introduce a flux variability analysis tailored to individual variability in metabolic reactions, enabling us to adjust the metabolic bounds in response to propagated changes. To robustly evaluate this framework, first, we present an extensive pan-cancer analysis integrating transcriptomics and metabolomics data. Second, we incorporated a third layer (proteomics) using a Progressive Supranuclear Palsy cohort.

2 Methods

In this section, we describe our methods that entail (i) constructing a global interaction network, (ii) propagating the omics measurements over this network to the metabolism, and (iii) computing the metabolic changes. An overarching end-to-end flowchart of our proposed pipeline is illustrated in Figure 1.

2.1 Constructing an integrated multi-omic network

We build an integrated multi-omic network that supports major genomic interactions such as expression, translation, transcription factors, post-transcriptional regulation (e.g., through miRNAs), metabolic interactions, etc. The nodes in the network represent various genomic entities such as genes, proteins, transcription factors, miRNAs, metabolites, and reactions. The edges represent different types of interactions between these entities, including gene expression regulation, protein-protein interactions, transcriptional regulation by transcription factors, post-transcriptional regulation by miRNAs, metabolite production and subsumption in biochemical reactions, etc. By using Recon3D[5], the metabolites are added as node and these nodes connect the metabolites to reactions depending on whether produced or consumed.

2.2 Propagation of Multi-Omics Measurements

Information diffusion and random walk models are well-established in systems biology for uncovering complex molecular mechanisms. For example, Random Walk with Restart (RWR) algorithms are widely used for disease-gene prioritization [16], while heat diffusion models like HotNet and HotNet2 identify significantly mutated subnetworks in cancer [15]. Furthermore, methods such as TieDIE utilize network diffusion to effectively connect genomic perturbations to transcriptional changes [17]. Building upon the proven ability of these algorithms to capture cascading downstream effects across biological networks, our framework employs diffusion models to propagate multi-omics fold-changes across an integrated interaction network.

2.2.1 Selecting diffusion model. We employ the Linear Threshold Model for the diffusion process.

- Any node that has a measurement is considered active initially, and all the remaining ones are considered to be inactive.
- At each time step, consider an individual i who is in the inactive state. Let θ_i, a value between 0 and 1, represent the threshold value for i. The state of i will switch to active if the proportion of its neighbors who are in the active state meets or exceeds θ_i.

$$I_i = \frac{\sum_{w \in \mathcal{N}(i)} b_{w,i} \cdot x_w}{\sum_{w \in \mathcal{N}(i)} b_{w,i}} \tag{1}$$

Where I_i is the total influence value, $b_{w,i}$ is the weight of the edge between i and its neighbor w, x_w is the node state (active or inactive, $x_w \in \{0, 1\}$), and $\mathcal{N}(i)$ is the set of neighbors of node i. The diffusion process finishes when the number of active individuals becomes stable [30].

2.3 Predicting the value of a new active node

We predict the value of new active nodes using information diffusion. While we evaluated four approaches (i.e., Sum, Max, Mean, and Linear Threshold diffusion—Supplementary Section S8), Max Diffusion yielded the most robust empirical performance. In the Max Diffusion model, the value of a node n is is the maximum weighted value among its activator neighbors minus the maximum weighted value among its repressor neighbors.

$$\text{val}(n) = \max_{a \in \mathcal{A}(n)} \{w_{n,a} \cdot \text{val}(a)\} - \max_{r \in \mathcal{R}(n)} \{w_{n,r} \cdot \text{val}(r)\} \tag{2}$$

Where $\text{val}(n)$ is the predicted value of node n; $\mathcal{A}(n)$ and $\mathcal{R}(n)$ are the sets of activator and repressor neighbors of node n; $w_{n,r}$ is the weight of the edge between nodes n and r; and $\text{val}(a)$ or $\text{val}(r)$ represent the values of those active neighbors.

2.4 Setting Reaction Flux Bounds

The diffusion process operates as follows: if there is an experimental result for a node, then the diffused value of the node is set to that measured value, and it does not change during the diffusion process. Otherwise, the value of the node is estimated according to its neighbors, and the propagation continues based on that value. The distinction between experimentally determined and propagated values allows the framework to maintain flexibility and robustness

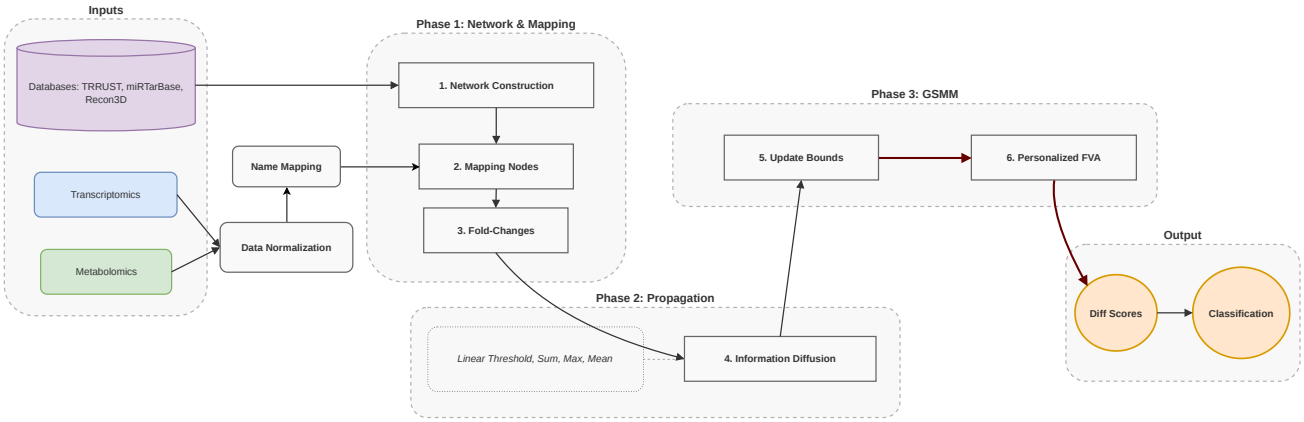


Figure 1: End-to-end workflow of the MetabOmics pipeline. Phase 1: Construction of the interaction network and mapping of measured omics entities. Phase 2: Propagation of multi-omics fold-changes via information diffusion models. Phase 3: Personalized metabolic integration and flux variability analysis to compute final differential scores.

in scenarios where empirical data might be sparse or incomplete. This ensures that the diffusion process can still proceed and that nodes without direct experimental data can be inferred through the network's diffusion dynamics.

As the network diffusion reaches its final stage, we evaluate the gene reaction rules (see Figure 1, Phase 3). Based on the diffused values, we then adjust the lower and upper bounds of the reactions in the genome-scale metabolic network model to reflect the inferred biological states. When the diffused value of a node is positive, the lower bound of the corresponding reaction is increased proportionally to the magnitude of the positive diffused value. Conversely, when the diffused value of a node is negative, both the lower and upper bounds of the reaction are reduced proportionally.

2.5 Configuring Personalized Objective Func.

We configure a personalized objective function by integrating patient-specific metabolomics data [6]. In particular, individual-specific reaction weights are computed based on metabolite fold-changes:

$$C_r^{(i)} = \sum_{m \in M_r^{prod}} \mathrm{mfc}_m^{(i)} \cdot \frac{S_{m,r}}{\sum_{r' \in P_m} |S_{m,r'}|}$$

where $mfc_m^{(i)}$ is the fold change for metabolite m, P_m is the set of reactions producing metabolite m, $S_{m,r}$ is the stoichiometric coeff. of metabolite m in reaction r, $S \in \mathbb{R}^{M \times R}$ is the stoichiometric matrix (metabolites $\times$ reactions), and $v \in \mathbb{R}^R$ is the reaction flux vector (variables). We solve the personalized optimization problem:

$$\max_v \ (C^{(i)})^\top v \quad \text{s.t.} \quad Sv = 0,\ v_{lb} \le v \le v_{ub}$$

where v_{lb} and v_{ub} are upper and lower reaction bounds.

3 Experimental Evaluation

In this section, we present an empirical evaluation of our proposed multi-omic integrated analysis approach on real datasets. We first describe our metrics and the datasets, followed by performance evaluation on six different datasets. To ensure reproducibility, the core MetabOmics pipeline is available via a repository: https://github.com/anonymousscodee/anoncoder123.

The primary metrics for evaluation were the F1-score (the harmonic mean of precision and recall), along with precision and recall, computed across all cross-validation folds to assess generalizability.

3.1 Dataset

We evaluated 764 tumor and 224 adjacent normal samples across six cancer cohorts (BRCA, COAD, PRAD, ccRCC3/4, PDAC) [3], alongside a 98-sample 3-omics (transcriptomics, metabolomics, proteomics) PSP cohort [2]. Detailed cohort statistics are in Suppl. Table S3.

3.2 Patient Stratification Performance

In this study, we implemented a machine-learning pipeline to analyze different types of cancer-related metabolomics and gene expression data. The primary objective was to classify samples accurately as 'healthy' or 'cancer' based on transformed features derived from metabolomic and gene expression profiles. The evaluation of the models' performance was conducted using stratified cross-validation to ensure robust and generalizable results.

3.2.1 Dual Omics Setting: Table S27 (Suppl. Material) displays the F1 scores of various classifiers on multiple cancer datasets, utilizing reaction differentiation scores as features. Each classifier, including logistic regression, random forest, SVM, and MLP, was fine-tuned to optimize performance. Logistic regression achieved the highest average F1 score (0.86) across datasets. This suggests that logistic regression is well-suited to reaction diff scores, likely due to its generalizability across cancer types. Random forest which is ensemble-based methods, demonstrates mixed performance. Random forest achieved competitive F1 scores in certain cases, such as kidney cancer (ccRCC3) and colon cancer, with F1 scores of 0.83 and 0.79, respectively. SVM and MLP demonstrated moderate performance, with SVM excelling in specific datasets but underperforming

overall. MLP had stable, though unremarkable, performance across datasets, suggesting limited benefit from its neural architecture with pathway diff scores. These results show that logistic regression consistently performed best.

3.2.2 Triple Omics Setting: To validate the framework's capacity for higher orders of omics integration, we additionally evaluated MetabOmics on the MayoRNASeq Progressive Supranuclear Palsy (PSP) cohort, simultaneously integrating transcriptomics, metabolomics, and proteomics data. As shown in Table 2, Logistic Regression achieves the strongest result (0.899 ± 0.035), demonstrating that the additional proteomics layer provides a complementary biological signal that improves both predictive accuracy and stability.

3.2.3 Comparison of Different Information Diffusion Methods. We evaluated four diffusion approaches (Max, Mean, Sum, and Linear Threshold) across multiple classifiers (e.g., Logistic Regression, Random Forest). The max diffusion (Max Diff.) approach achieved the highest average F1-score (0.886) across the six cancer datasets, performing exceptionally well in kidney (0.975) and colon (0.904) cancers. Mean Diffusion (0.884) and Linear Threshold (0.875) also showed highly competitive performance, while Sum Diffusion (0.842) struggled to capture detailed pathway signals. Because Max Diffusion isolates the strongest regulatory influences—effectively capturing key molecular cascading events without diluting the signal—it was selected as the primary diffusion model for all subsequent analyses. Detailed classifier performance tables for all diffusion methods are provided in Supplementary Section S8.

3.3 Biological Validation of Discovered Features

Significant pathways ($p < 0.05$) identified by each method were validated against a literature-curated ground truth for the cohorts of breast cancer [18, 26], prostate cancer [7, 11, 23], pancreatic cancer [20, 22, 27], clear cell renal cell carcinoma [9, 10, 21], colorectal cancer [14, 29, 31]. We utilized Precision, Recall, and F1-score.

Compared to standalone single-omics baselines, *MetabOmics* achieved superior precision, recall, and literature concordance F1-scores across almost all evaluated cohorts by ensuring better consistency with established biological hallmarks. For example, MetabOmics doubled the literature concordance F1-score of deltaFBA and Metabolitics in both the PDAC (0.400) and ccRCC3 (0.200) cohorts. These results demonstrate that the integration of multi-omic layers allows the framework to capture complex regulatory shifts and directional flux changes that single-omic approaches inherently misinterpret.

For instance, **Arachidonic acid metabolism** is significantly **increased** in **Colon Cancer (COAD)**, driven by the upregulation of the COX-2 pathway, which promotes inflammation and tumor cell survival. While standalone metabolomics analysis (i.e., *Metabolitics*) may identify metabolites in this pathway as altered, it often fails to provide the necessary regulatory context for flux direction. Similarly, while *deltaFBA* identifies the importance of this pathway based on gene expression bounds, it lacks the integrated metabolite-driven flux pressure required to determine the correct direction, often resulting in directional discordance. In contrast, *MetabOmics* correctly identifies both the importance and the upregulation of this pathway, consistent with the literature indicating that eicosanoid signaling is a central hallmark of CRC progression.

Similarly, in **Kidney Cancer (ccRCC)**, **Fatty acid synthesis** is **increased** while its **degradation (oxidation)** is **repressed** via HIF-mediated signaling. This synchronized regulation leads to the formation of large cytoplasmic lipid droplets, which is the defining "clear cell" morphological hallmark of this cancer. While standalone flux balance analysis often fails to distinguish these specific synthesis requirements from general maintenance, *MetabOmics* accurately predicts the increased flux through the synthesis pathway, outperforming baseline methods in directional accuracy.

In the **Breast Cancer** cohort, **Nucleotide interconversion** activity is **increased** to meet the heightened demand for DNA replication in highly proliferative tumor cells. *MetabOmics* captures the correct flux direction of this process, whereas standalone analysis via *deltaFBA* frequently fails to prioritize this pathway correctly or assigns an incorrect direction because it evaluates gene expression independent of the metabolic neighborhood. By utilizing information diffusion, *MetabOmics* ensures that regulatory signals from rate-limiting enzymes are propagated through the network's topology, yielding a physiologically accurate representation. For **Pancreatic Cancer (PDAC)**, **Nucleotide interconversion** and lipid reprogramming pathways such as **Fatty acid synthesis** are significantly **increased** to support anabolic growth under hypoxic conditions. *MetabOmics* demonstrates high biological fidelity by correctly predicting the increased flux through these hallmarks, yielding the highest F1-score (0.400) among the tested methods. In this cohort, *deltaFBA* typically identifies the importance of these anabolic shunts but fails to resolve the increased flux direction due to the lack of metabolite-driven metabolic pressure.

The relatively lower performance in the **Prostate Cancer (PRAD)** cohort can be attributed to data sparsity. The PRAD dataset features only 387 measured metabolites, significantly fewer than cohorts like ccRCC (966), which restricts the framework's ability to effectively anchor diffused transcriptomic signals onto the metabolic network. This limitation particularly hinders the baseline *deltaFBA* model, which relies exclusively on transcriptomic boundaries without the anchoring benefit of measured metabolite levels to resolve complex regulatory states like the 'reverse Warburg' effect. Note: The literature concordance for the 3-omics PSP cohort is evaluated alongside the interpretability baseline comparisons in Section 3.5 (Table 4).

3.4 Comparison with State-of-art

To compare different approaches for mapping omics data into pathway-level features, we evaluated Metabolitics, MetabOmics, Partial Least Squares Discriminant Analysis (PLS-DA), Canonical Correlation Analysis (CCA), and Data Concatenation methods under the same experimental setup. As shown in Table 1, simple Data Concatenation achieves the highest raw F1 scores across multiple cohorts for the dual-omics setting. Results for the triple-omics setting are provided in Table 2, where MetabOmics performance is superior to both simple data concatenation and Metabolitics.

While purely data-driven approaches, such as PLS-DA, CCA, and simple data concatenation, often achieve higher raw F1 scores in our evaluations, these models remain primarily statistical tools

Table 1: Model Performance Comparison for Dual-Omics Setting (Weighted F1-score)

Dataset	Metabolitics	deltaFBA	PLS-DA	Corr-Net (CCA)	MetabOmics	Data Concat.	1D CNN
Breast (BRCA)	0.772 ± 0.182	0.6709 ± 0.0655	0.908	0.862	0.803 ± 0.114	**0.9075 ± 0.0705**	0.80
Kidney (ccRCC3)	0.936 ± 0.060	0.7556 ± 0.1342	0.956	**0.982**	0.946 ± 0.044	0.9226 ± 0.0845	0.84
Kidney (ccRCC4)	0.973 ± 0.055	0.8333 ± 0.1044	**0.987**	0.961	0.961 ± 0.059	0.9274 ± 0.0563	0.82
Colon (COAD)	0.790 ± 0.133	0.6654 ± 0.1879	0.907	**0.961**	0.887 ± 0.094	0.9412 ± 0.0973	0.86
Pancreas (PDAC)	0.546 ± 0.263	0.5704 ± 0.0614	**0.842**	0.827	0.832 ± 0.107	0.8357 ± 0.0671	0.84
Prostate (PRAD)	0.832 ± 0.098	0.6140 ± 0.0757	0.734	0.801	**0.889 ± 0.069**	0.8116 ± 0.0688	0.77
Average	0.808	0.685	0.889	**0.899**	0.886	0.891	0.822

Table 2: Model Performance for PSP — Triple-Omics Integration (Weighted F1-score)

Method	XGBoost	LogReg	RF	SVM	MLP
MetabOmics	0.820	**0.899**	0.838	0.740	0.814
Metabolitics	0.802	**0.835**	0.762	0.771	0.806
Data Concat.	**0.8451**	0.7050	0.7262	0.7050	0.7186

that are highly open to overfitting. This is particularly critical given the high-dimensional nature of multi-omics data and the relatively small sample sizes of the available cohorts. By feeding thousands of unconstrained raw features directly into a machine learning classifier, the model is likely to memorize noise rather than learn generalizable biological patterns. In contrast, MetabOmics acts as a biological regularizer by propagating signals through an integrated interaction network and constraining them within the bounds of a genome-scale metabolic model (GSMM). Although this biological regularization may slightly reduce raw predictive metrics on the training set, it prevents the model from acting as a "black box". More importantly, it provides crucial mechanistic interpretability, enabling the identification of specific dysregulated metabolic reactions and pathways driving the classification, a fundamental objective of systems biology that simple concatenation cannot fulfill.

3.5 Interpretability Analysis

While the data concatenation baseline achieves comparable or superior F1 scores in binary classification, we argue that predictive performance alone is an insufficient criterion for evaluating multi-omics integration methods. A biologically meaningful method must also produce stable, reproducible feature selections that align with established disease mechanisms. To investigate whether these 'black-box' models are learning genuine biology or memorizing noise, we conducted feature stability and literature concordance analyses.

3.5.1 Feature Stability. To assess the consistency of feature selection across cross-validation folds, we computed the mean pairwise Jaccard similarity of the top-30 features selected by XGBoost importance in each fold. A higher Jaccard index indicates that the model consistently relies on the same biological features regardless of data splits, which is a hallmark of a generalizable biological signal rather than statistical noise.

As shown in Table 3, MetabOmics achieves substantially higher Jaccard indices than data concatenation in six out of seven cohorts. This demonstrates that the biological network constraints imposed by MetabOmics force the model to consistently select the same core metabolic pathways across folds, whereas data concatenation selects largely different, high-variance genomic features in each fold. The exception is PDAC, where the small sample size (n=39) limits the stability of both methods.

Table 3: Feature Stability: Mean Pairwise Jaccard Similarity of Top-30 Features Across Folds

Dataset	MetabOmics	Data Concat.
Breast (BRCA)	**0.579**	0.123
Colon (COAD)	**0.612**	0.273
Prostate (PRAD)	**0.422**	0.323
Kidney (ccRCC3)	**0.586**	0.408
Kidney (ccRCC4)	**0.805**	0.650
Pancreas (PDAC)	0.386	**0.565**
PSP (MayoRNASeq)	**0.507**	0.154

3.5.2 Literature Concordance with Data Concatenation Baseline. To further quantify the interpretability advantage of MetabOmics beyond classification accuracy, we extended the literature concordance analysis to include the data concatenation baseline (Table 4). For this comparison, the top-30 features selected by XGBoost from the data concatenation model were mapped to metabolic pathways using Recon3D, and the resulting pathway sets were evaluated against the same literature ground truth.

Table 4: Literature Concordance: MetabOmics vs. Data Concatenation (Top-30, Pathway Level)

Dataset	Metric	MetabOmics	Data Concat.
Breast (BRCA)	Precision	**0.190**	0.286
	Recall	**0.400**	0.200
	F1-Score	**0.258**	0.235
Colon (COAD)	Precision	0.000	0.000
	Recall	0.000	0.000
	F1-Score	0.000	0.000
Prostate (PRAD)	Precision	**0.067**	0.000
	Recall	**0.125**	0.000
	F1-Score	**0.087**	0.000
Kidney (ccRCC3)	Precision	**0.250**	1.000
	Recall	**0.250**	0.125
	F1-Score	**0.250**	0.222
Kidney (ccRCC4)	Precision	**0.200**	0.000
	Recall	**0.500**	0.000
	F1-Score	**0.286**	0.000
Pancreas (PDAC)	Precision	**0.200**	0.000
	Recall	**0.333**	0.000
	F1-Score	**0.250**	0.000
PSP (MayoRNASeq)	Precision	**0.042**	0.000
	Recall	**0.167**	0.000
	F1-Score	**0.067**	0.000

These results reveal a critical interpretability gap. Despite its high raw classification accuracy, data concatenation achieves a literature concordance F1-score of zero in five out of seven cohorts. This quantitatively proves our hypothesis: flat concatenation models achieve high statistical performance by overfitting to high-variance, biologically spurious noise. In stark contrast, MetabOmics achieves non-zero, literature-validated F1-scores across all seven cohorts. By routing signals through a constrained metabolic network, MetabOmics acts as a vital biological regularizer, trading

a marginal drop in statistical F1 for a massive gain in mechanistic validity. Notably, in the PSP cohort, where three omics layers (transcriptomics, metabolomics, and proteomics) are integrated, MetabOmics successfully identifies Sphingolipid metabolism as a concordant pathway, consistent with known myelination defects in PSP pathology [12].

4 Conclusion and Future Work

In this paper, we introduced MetabOmics, a network-driven multi-omics integration framework that bridges the gap between disparate molecular layers and metabolic phenotypes. Rather than relying on unconstrained feature fusion, our approach contextualizes omics measurements within a comprehensive, 5-layer biological simulation scaffold. By leveraging algorithmic information diffusion, specifically Max Diffusion and Linear Threshold models, we effectively propagate molecular fold-changes across both measured and unmeasured intermediate regulators to personalize genome-scale metabolic models (GSMMs).

We conducted an extensive pan-cancer analysis integrating paired transcriptomics and metabolomics data across six distinct cancer cohorts. Besides, we validated the framework's capacity for deep, three-omics integration by simultaneously incorporating transcriptomics, metabolomics, and proteomics from the MayoRNASeq Progressive Supranuclear Palsy (PSP) cohort. Across these diverse datasets, MetabOmics achieved high performance.

Future work will focus on integrating metabolite imputation algorithms and external reference metabolomes to address sparse data constraints, alongside validating the framework on larger independent cohorts. Additionally, the authors plan to deploy Graph Neural Networks (GNNs) to natively ingest the network structure, allowing the model to learn localized propagation weights and complex non-linear metabolic relationships. Finally, the framework will be expanded to incorporate new regulatory layers (such as epigenomics and phosphoproteomics), explore dynamic metabolic formulations, and predict individualized patient responses to targeted metabolic inhibitors for clinical precision oncology.

Funding

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) through the EU Joint Programme (JPND) (Grant No. 124N069).

References

[1] Fadi Alharbi, Aleksandar Vakanski, Boyu Zhang, Murtada K Elbashir, and Mohanad Mohammed. 2025. Comparative Analysis of Multi-Omics Integration Using Graph Neural Networks for Cancer Classification. *IEEE Access* (2025).

[2] Mariet Allen et al. 2016. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Scientific Data* 3 (2016), 160089.

[3] E. Benedetti, E. M. Liu, C. Tang, F. Kuo, M. Buyukozkan, T. Park, and E. Reznik. 2023. A multimodal atlas of tumour metabolism reveals the architecture of gene–metabolite covariation. *Nature Metabolism* 5, 6 (2023), 1029–1044.

[4] B. Berger, J. Peng, and M. Singh. 2013. Computational solutions for omics data. *Nat. Rev. Genet.* 14 (2013), 333–346.

[5] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, et al. 2018. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology* 36, 3 (2018), 272–281.

[6] Ali Cakmak and M Hasan Celik. 2021. Personalized metabolic analysis of diseases. *IEEE/ACM transactions on computational biology and bioinformatics* 18, 3 (2021), 1014–1025.

[7] LC Costello et al. 2016. A comprehensive review of the role of zinc in normal prostate function and metabolism; and its implications in prostate cancer. *Archives of Biochemistry and Biophysics* 611 (2016), 100–112. doi:10.1016/j.abb.2016.04.014

[8] Yuehua Cui. 2022. Multivariate data integration using R: Methods and applications with the mixOmics package.

[9] Q Deng et al. 2024. Lipid reprogramming and ferroptosis crosstalk in clear cell renal cell carcinoma: metabolic vulnerabilities and therapeutic targeting. *Cell Communication and Signaling* (2024).

[10] TE Edo et al. 2021. Metabolic Reprogramming in Kidney Cancer: Implications for Therapy. *Frontiers in Oncology* (2021).

[11] E Eidelman et al. 2017. The Metabolic Phenotype of Prostate Cancer. *Frontiers in Oncology* 7 (2017), 131. doi:10.3389/fonc.2017.00131

[12] Lukas da Cruz Carvalho Iohan, Jean-Charles Lambert, and Marcos R. Costa. 2022. Analysis of modular gene co-expression networks reveals molecular pathways underlying Alzheimer's disease and progressive supranuclear palsy. *PLOS ONE* 17, 4 (2022), e0266405. doi:10.1371/journal.pone.0266405

[13] Min-Zhi Jiang, François Aguet, Kristin Ardlie, Jiawen Chen, Elaine Cornell, Dan Cruz, Peter Durda, Stacey B Gabriel, Robert E Gerszten, Xiuqing Guo, et al. 2023. Canonical correlation analysis for multi-omics: Application to cross-cohort analysis. *PLoS genetics* 19, 5 (2023), e1010517.

[14] CO Kelson and YY Zaytseva. 2024. Altered lipid metabolism in APC-driven colorectal cancer: the potential for therapeutic intervention. *Frontiers in Oncology* 14 (2024). doi:10.3389/fonc.2024.1343061

[15] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, et al. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* 47, 2 (2015), 106–114.

[16] Kathy Macropol, Tolga Can, and Ambuj K Singh. 2009. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics* 10, 1 (2009), 1–16.

[17] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, and Joshua M Stuart. 2013. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 29, 21 (2013), 2757–2764.

[18] Ian F Robey et al. 2008. Regulation of the warburg effect in early-passage breast cancer cells. *Neoplasia* 10, 7 (2008), 745–756.

[19] Rajib Roychowdhury, Soumya Prakash Das, Amber Gupta, Parul Parihar, Kottakota Chandrasekhar, Umakanta Sarker, Ajay Kumar, Devade Pandurang Ramrao, and Chinta Sudhakar. 2023. Multi-omics pipeline and omics-integration approach to decipher plant's abiotic stress tolerance responses. *Genes* 14, 6 (2023), 1281.

[20] Naiara Santana-Codina, Anjali A Roeth, et al. 2018. Oncogenic KRAS supports pancreatic cancer through regulation of nucleotide synthesis. *Nature Communications* 9 (2018), 4945.

[21] GL Semenza. 2007. HIF-1 mediates the Warburg effect in clear cell renal carcinoma. *Journal of Bioenergetics and Biomembranes* (2007).

[22] Jaekyoung Son, Costas A Lyssiotis, Haoqiang Ying, Xiaoxu Wang, Sujun Hua, Matteo Ligorio, et al. 2013. Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. *Nature* 496 (2013), 101–105.

[23] JV Swinnen et al. 2002. Overexpression of fatty acid synthase is an early and common event in the development of prostate cancer. *International Journal of Cancer* 98, 1 (2002), 19–22. doi:10.1002/ijc.10127

[24] Conghao Wang, Wu Lue, Rama Kaalia, Parvin Kumar, and Jagath C Rajapakse. 2022. Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. *Scientific Reports* 12, 1 (2022), 15425.

[25] Cecilia Wieder, Juliette Cooke, Clement Frainay, Nathalie Poupin, Russell Bowler, Fabien Jourdan, Katerina J Kechris, Rachel PJ Lai, and Timothy Ebbels. 2024. PathIntegrate: Multivariate modelling approaches for pathway-based multi-omics data integration. *PLOS Computational Biology* 20, 3 (2024), e1011814.

[26] David R Wise and Craig B Thompson. 2010. Glutamine addiction: a new therapeutic target in cancer. *Trends in biochemical sciences* 35, 8 (2010), 427–433.

[27] Haoqiang Ying, Alec C Kimmelman, Costas A Lyssiotis, Sujun Hua, Gerald C Chu, Eliot Fletcher-Sananikone, Jason W Locasale, Jaekyoung Son, Hailei Zhang, Jonathan L Coloff, et al. 2012. Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell* 149, 3 (2012), 656–670.

[28] Jiayang Zhang, Yilin Che, Rongrong Liu, Zhicheng Wang, and Weiwu Liu. 2025. Deep learning–driven multi-omics analysis: enhancing cancer diagnostics and therapeutics. *Briefings in Bioinformatics* 26, 4 (2025), bbaf440.

[29] J Zhang and S Zou. 2023. Metabolic reprogramming in colorectal cancer: regulatory networks and therapy. *Cell & Bioscience* 13 (2023), 1–22. doi:10.1186/s13578-023-00977-w

[30] Z. K. Zhang, C. Liu, X. X. Zhan, X. Lu, C. X. Zhang, and Y. C. Zhang. 2016. Dynamics of information diffusion and its applications on complex networks. *Physics Reports* 651 (2016), 1–34.

[31] R Zhou and F Wang. 2021. Glucose Metabolic Reprogramming in Colorectal Cancer: From Mechanisms to Targeted Therapy Approaches. *Cancer Medicine* (2021). doi:10.1002/cam4.71185