www.ijecs.in
*International Journal Of Engineering And Computer Science*
*Volume 14 Issue 12 December 2025, Page No. 28062-28069*
*ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i12.5359*

# Discovering Significant Cancer Related Metabolite Correlations through Statistical Measures and Graph Attention Network Model

**Dr. R. Porkodi [1], M. Nandhini [2]**

[1] Department of Computer Science Bharathiar University Coimbatore-641046, INDIA
porkodi_r76@buc.edu.in
[2] Department of Computer Science Bharathiar University Coimbatore-641046, INDIA
nandhudass12@gmail.com

**Abstract**
Metabolites are small molecules that play vital roles in sustaining biological functions and supporting energy production. Investigating disease-associated metabolites allows researchers to gain deeper insights into their contribution to disease development. This research is focused on exploring metabolite–disease associations through similarity analysis. Cancer-related diseases and their corresponding metabolites are retrieved from the HMDB database. The metabolite and disease data are converted into numerical vectors using the TF-IDF method to enable similarity analysis. Three statistical similarity measures—Pearson correlation, Bhattacharyya distance, and Chebyshev distance—are applied to the TF-IDF dataset. The main objective of this research is to develop a graph-based model that predicts the correlation strength between metabolites and diseases using statistical measures. The research is conducted in three phases. In the first phase, the primary objective is to identify the most effective statistical similarity measure for detecting metabolite–disease associations. In the second phase, the similarity dataset is used as input to a Graph Attention Network model, which classified associations as strong, moderate, or weak. This framework provided valuable insights into the degree of relationship between metabolites and diseases, highlighting how metabolite imbalances may contribute to disease onset and facilitate early diagnosis. Based on the comparative evaluation of the three measures, Pearson correlation proved to be the most suitable method for metabolite–disease analysis. Furthermore, in the overall results, the GAT model predicted the metabolite–disease associations as moderate and achieved an accuracy of 99%. In the third phase, the TF-IDF dataset is directly given to the GAT model to predict association between metabolites and diseases. Finally, results from the Pearson similarity dataset with the GAT model and results from the TF-IDF method with the GAT model are compared. While the TF–IDF approach within the GAT framework achieved 33% accuracy, this research findings indicate that the GAT model with Pearson similarity, is particularly well-suited for examining how individual metabolites are linked to multiple diseases in terms of correlation strength.

**Keywords:** Metabolites, Cancer, Diseases, Statistical Measures, TF- IDF, GAT, Correlation Strength

## 1. Introduction

Metabolomics is the comprehensive study of small molecules, or metabolites, within cells, tissues, and biofluids. These metabolites represent the downstream products of gene expression and protein activity, thereby serving as sensitive indicators of physiological and pathological states. By profiling the dynamic metabolome, researchers can capture real-time snapshots of biochemical processes, offering insights into health, disease progression, and therapeutic responses. Metabolism plays a central role in disease development, as disruptions in metabolic pathways often precede clinical symptoms. Altered glucose utilization in diabetes, dysregulated lipid metabolism in cardiovascular disorders, and abnormal amino acid turnover in cancer exemplify how metabolic imbalance drives disease creation. The association between specific metabolites and diseases is crucial, as these biomarkers not only reveal underlying mechanisms but also enable early diagnosis, risk stratification, and personalized treatment strategies. Cancer is linked to

changes in the body's normal balance of metabolites. Tumor cells change how they use energy by shifting processes like sugar breakdown, amino acid use, and fat production. These changes cause unusual levels of metabolites, which spread through the body, disturb overall balance, and connect cancer with other health problems such as diabetes, obesity, and heart disease. Statistical measures play an important role in analysing the relationship between metabolites and diseases. They provide a means of quantifying the strength of associations, helping researchers determine whether a metabolite is strongly, moderately, or weakly linked to a particular disease. Statistical methods ensure that findings are both reliable and meaningful. Moreover, these measures support biomarker discovery and serve as a foundation for integrating advanced models, such as Graph Attention Networks, with additional approaches like TF–IDF method.

Understanding the significance of metabolites in relation to diseases motivates that the identification of correlation levels, which consequently contributes to disease prevention when metabolite imbalances arise.

## 2. Literature Review

**Table I. Key Results from Reviewed Studies**

| S.No | Title | Authors | Key Findings |
|---|---|---|---|
| **1.** | MetaPredictor: in silico prediction of drug metabolites based on deep language models with prompt engineering | Keyun Zhu, Mengting Huang, Yimeng Wang, Yaxin Gu, Weihua Li, Guixia Liu, Yun Tang | Applies deep language models with prompt engineering to predict drug metabolites. Improves generalization and scalability compared to traditional computational methods. [2] |
| **2** | DMoVGPE: predicting gut microbial associated metabolites profiles with deep mixture of variational Gaussian Process experts | Qinghui Weng, Mingyi Hu, Guohao Peng, Jinlin Zhu | Proposes a deep mixture of variational Gaussian Process experts model to predict gut microbial metabolite profiles from sequencing data. Offers cost-effective alternatives to direct metabolomics. [3] |
| **3** | Explore potential disease-related metabolites based on latent factor model | Yongtian Wang, Liran Juan, Jiajie Peng, Tao Wang, Tianyi Zang, Yadong Wang | Uses latent factor modeling to identify disease-related metabolites. Demonstrates utility in uncovering metabolic signatures for diagnosis and mechanistic insights. [4] |
| **4** | Predicting Disease–Metabolite Associations Based on the Metapath Aggregation of Tripartite Heterogeneous Networks | Wenzhi Liu, Pengli Lu | Employs metapath aggregation in heterogeneous networks to predict disease–metabolite associations. Outperforms traditional experimental methods in efficiency and scalability. [5] |

| | | | |
|---|---|---|---|
| **5** | MicrobeRX: a tool for enzymatic-reaction-based metabolite prediction in the gut microbiome | Angel J. Ruiz-Moreno, Ángela Del Castillo-Izquierdo, Isabel Tamargo-Rubio, Jingyuan Fu | Introduces *MicrobeRX*, integrating thousands of human, microbial, and drug metabolic reactions to predict gut microbiome-derived metabolites and their impact on host health. [6] |

**Table II. Cancer-Related Diseases and their Corresponding Metabolites**

## 3. Materials and methods

This section outlines the resources, methodological approaches, and analytical tools employed in assessing the strength of correlations between metabolites and diseases.

Metabolite data associated with cancer are retrieved from the HMDB database, with metabolite names and corresponding disease names extracted as features. The metabolite–cancer dataset is constructed using two approaches. First, a dataset is generated through the TF-IDF method, which is subsequently transformed into similarity-based datasets. To evaluate correlation strength between metabolites and diseases, a Graph Attention Network (GAT) model is developed. Three statistical similarity measures—Bhattacharyya, Pearson, and Chebyshev—is

applied to the TF-IDF dataset, yielding three distinct similarity datasets. Comparative analysis revealed that Pearson similarity Outperformed than Bhattacharyya and Chebyshev in predicting metabolite–disease correlation strength. The GAT model is trained and tested using both the TF-IDF dataset and the Pearson similarity dataset to assess the correlation strength between metabolites and cancer related diseases.

## 4. MDCS Analysis Framework

This metabolites and diseases correlation strength analysis framework is designed to uncover associations between metabolites and diseases through statistical evaluation and to predict their
correlations using a Graph Attention Network (GAT) model. This framework is designed to uncover associations between metabolites and diseases through statistical evaluation and to predict their correlations using a Graph Attention Network (GAT) model.

### *Data Collection*

A comprehensive corpus comprising 24,002 metabolites and diseases are initially retrieved from the HMDB database. For the present investigation, 907 metabolites are identified as being connected to 12 cancer-associated diseases. These include colorectal, ovarian, pancreatic, lung, thyroid, prostate, kidney, stomach, endometrial, and cervical cancers, together with metastatic bone disease and perillyl alcohol administration in cancer treatment. A representative subset of these cancer-related diseases and their linked metabolites is presented
below.

### *Feature Extraction*

The analysis of metabolite–disease associations is conducted using only two parameters: the metabolite name and the disease name.

### *Numerical vectors creation*

The metabolite name and disease name are in the form of text data. So, have to convert it into numerical values for any model learning process.it is accomplished by statistical method TF-IDF (*Term Frequency – Inverse Document Frequency).*

It highlights specific terms and documents relationships and prioritize important documents and terms pairs.

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

Where TF (td)= frequency of term t in document d

$$IDF(t) = \log(N/(df(t)+1)) + 1 \quad (2)$$

***GAT Model Development***

Graph Attention Networks (GATs) provide a powerful framework for

**Table III. TF – IDF values of Metabolites and Diseases**

| Metabolite | Disease |
|---|---|
| Citrulline | Colorectal cancer |
| Citrulline | Pancreatic cancer |
| 5,8,11-Eicosatrienoic acid | Colorectal cancer |
| LysoPC(14:0/0:0) | Ovarian cancer |
| Pyroglutamylglycine | Colorectal cancer |
| Undecanedioic acid | Colorectal cancer |
| Valeric acid | Colorectal cancer |
| Suberic acid | Colorectal cancer |
| 1-Methylguanosine | Colorectal cancer |

| Disease | 10 | 10z | 11 | 11d | 11z |
|---|---|---|---|---|---|
| Lung Cancer | 0 | 0 | 0 | 0 | 0 |
| Ovarian cancer | 0 | 0.181877 | 0 | 0 | 0.363754 |
| Pancreatic cancer | 0 | 0 | 0 | 0 | 0 |
| Stomach cancer | 0 | 0 | 0 | 0 | 0 |
| Thyroid cancer | 0 | 0 | 0 | 0 | 0 |

investigating relationships because they can capture the fact that not all associations are equally strong. By assigning different weights to connections, GATs provide graded predictions that reflect biological reality, distinguishing strong, moderate, and weak links. This weighting improves statistical rigor by preventing all associations from being treated the same. In addition, the attention mechanism produces interpretable outputs, such as attention maps, which highlight the most influential metabolites for a given disease.

The workflow of GAT model is explained below.

*BEGIN*

*LOAD dataset from CSV*
*EXTRACT unique metabolites and diseases*
*ASSIGN unique IDs to each metabolite and disease*
*FOR each row in dataset:*
*GET metabolite ID (u)*
*GET disease ID (v)*
*ADD edge (u, v)*
*COMPUTE absolute Pearson correlation (r)*
*IF r ≥ 1 THEN label = Strong (2)*

*ELSE IF r ≥ -1 THEN label = Moderate (1)*
*ELSE label = Weak (0)*
*STORE metabolite, disease, correlation*

*CONVERT edges and labels to tensors*
*CREATE one-hot encoded node features*
*BUILD PyTorch Geometric Data object with nodes, edges, and labels*

*INITIALIZE GAT with:*
*- Layer 1: GATConv (input → hidden)*
*- Activation: ELU*

*- Layer 2: GATConv (hidden → output classes)*
*FOR epoch = 1 to 50:*
*PERFORM forward pass to compute node embeddings*
*FOR each edge (u, v):*
*COMPUTE edge embedding = element-wise product of node embeddings*
*APPLY log-softmax to edge embeddings*
*COMPUTE loss using negative log likelihood*
*BACKPROPAGATE and update model weights*
*IF epoch MOD 10 == 0 THEN PRINT loss and accuracy*

*AFTER training:*
*COMPUTE final predictions for all edges*
*MAP numeric labels to text (Weak, Moderate, Strong)*
*SAVE results to CSV file*

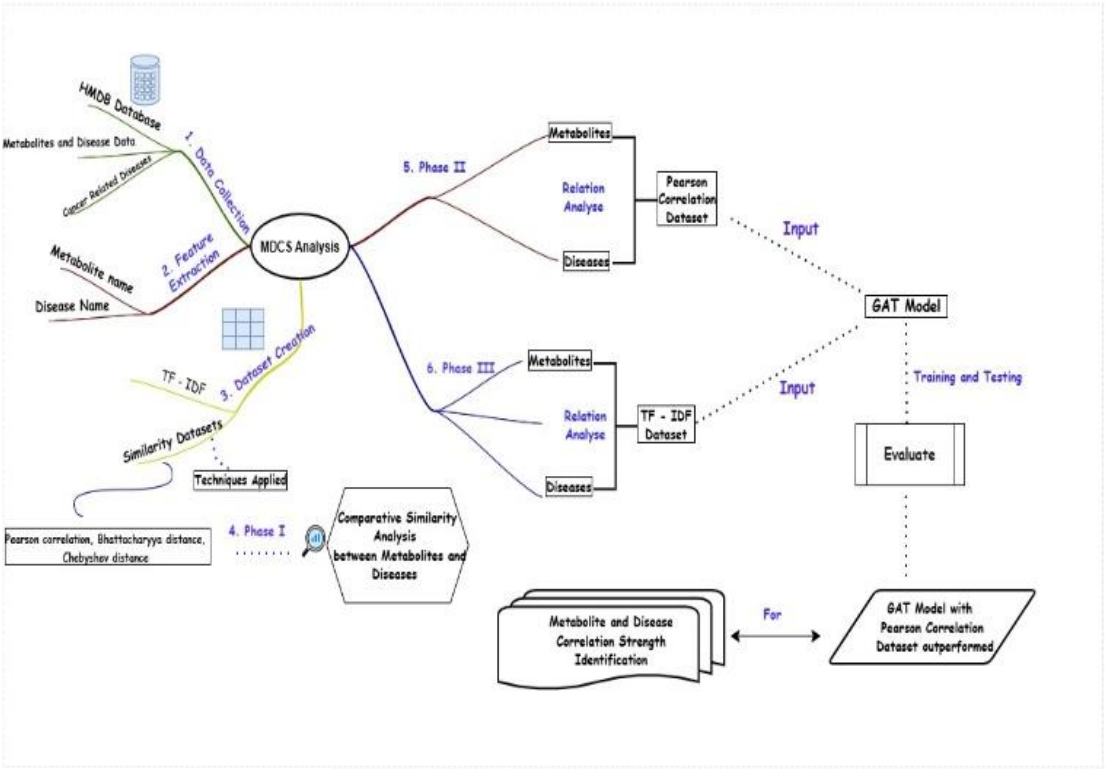


Fig. 1 The workflow of MDCS

The MDCS (Metabolite and Disease Correlation Strength) Analysis pipeline presents a structured, multi-phase approach to uncovering relationships between metabolites and diseases using advanced computational techniques. Beginning with data extraction from the HMDB database, the process isolates key features—metabolite and disease names—which are then used to construct two distinct datasets: one based on TF-IDF textual similarity and another leveraging numerical similarity metrics such as Pearson correlation, Euclidean distance, and Chebyshev distance. These datasets undergo three analytical phases: Phase I applies the similarity metrics; Phase II explores relationships using the Pearson Correlation Dataset; and Phase III

investigates associations via the TF-IDF Dataset. Both datasets are subsequently input into a Graph Attention Network (GAT) model, which is trained and evaluated to determine the strength of metabolite–disease correlations. The analysis concludes that the GAT model trained on the Pearson Correlation Dataset yields superior performance, highlighting its effectiveness in capturing biologically meaningful associations.

## 5. Metabolites and diseases correlation strength analysis

This section illustrates the application of a Graph Attention Network (GAT) model to metabolite–disease data, incorporating statistical measures including TF-IDF, Bhattacharyya, Pearson, and Chebyshev to assess correlation strength.

### *(i) Results of Metabolite to Disease Similarity Calculation*

**Table IV. Metabolite to Disease Similarity**

| Metrics | Min | Max | Mean | Std Dev |
|---|---|---|---|---|
| Bhattacharyya Distance | 0.0000 | 3.7256 | 2.9795 | 0.7877 |
| Pearson Correlation | -0.0131 | 1.0000 | 0.0000 | 0.0364 |
| Chebyshev Distance | 0.0000 | 1.0000 | 0.9944 | 0.0340 |

Pearson is the most useful because:

It gives low variation (0.0364). **Range (-0.0131 to 1.0000)** → Pearson can capture both weak negative and strong positive links.

The analysis revealed that not every metabolite is associated with every disease; in fact, most metabolite–disease pairs are uncorrelated. This neutral background is valuable, as it highlights the few pairs that exhibit strong positive or negative correlations. These distinct signals allow researchers to spot the important, biologically meaningful connections hidden within the data, providing a clearer path to understanding metabolite–disease relationships.

**Mean**

The **mean** is the average value of all the scores. It shows the **overall tendency** of the measure. In this Example, Pearson's mean = **0.0000** → most metabolite–disease pairs are uncorrelated. Bhattacharyya's mean = **2.9795** → on average, the distance is high, meaning less similarity. Chebyshev's mean = **0.9944** → almost all pairs look maximally different.

**Standard Deviation**

The **Std Dev** shows how much the values vary around the mean. It explains about **consistency and spread**. Pearson's Std Dev = **0.0364** → very small spread, values are stable and consistent. Bhattacharyya's Std Dev = **0.7877** → large spread, values differ a lot. Chebyshev's Std Dev = **0.0340** → small spread, but since the mean is near 1, almost all values are equally dissimilar. This visualization underscores the distinct statistical behavior of each metric and emphasizes the importance of considering both central tendency (mean) and variability (how much values change around the mean) when interpreting model or data characteristics.

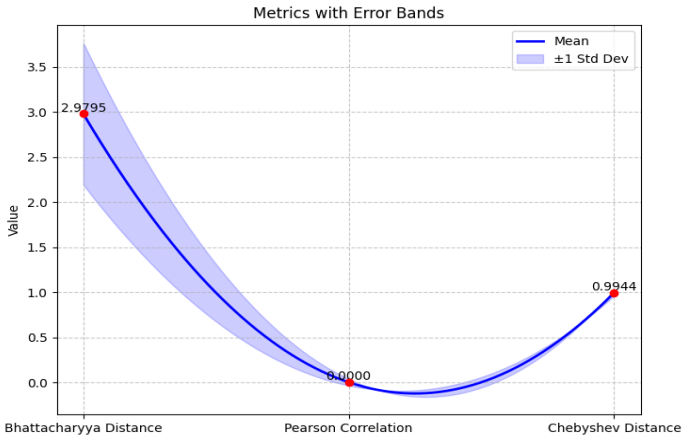


Fig. 2 Metabolite to Disease Similarity

This visualization underscores the distinct statistical behavior of each metric and emphasizes the importance of considering both central tendency (mean) and variability (how much values change around the mean) when interpreting model or data characteristics.

*(ii) Results of Metabolite to Disease Similarity using GAT model with Pearson similarity dataset*

Epoch 0, Loss: 1.0986, Accuracy: 0.4509
Epoch 10, Loss: 0.1623, Accuracy: 0.9999
Epoch 20, Loss: 0.0012, Accuracy: 0.9999
Epoch 30, Loss: 0.0016, Accuracy: 0.9999
Epoch 40, Loss: 0.0013, Accuracy: 0.9999
Training complete.

**Overall Model Accuracy: 0.9999**

*(iii) Results of Metabolite to Disease Similarity using GAT model with TF-IDF dataset*

Epoch 0, Loss: 1.0830, Train Acc: 0.4444, Test Acc: 0.0000
Epoch 5, Loss: 0.8286, Train Acc: 0.6667, Test Acc: 0.0000
Epoch 10, Loss: 0.6856, Train Acc: 0.7778, Test Acc: 0.0000
Epoch 15, Loss: 0.5444, Train Acc: 0.7778, Test Acc: 0.0000
Epoch 20, Loss: 0.3391, Train Acc: 0.8889, Test Acc: 0.0000
Epoch 25, Loss: 0.2296, Train Acc: 0.8889, Test Acc: 0.0000

**Final Test Accuracy: 0.3333333333333333**

Cancer with metastatic bone disease: Predicted association strength = 0
Cervical cancer: Predicted association strength = 2
Colorectal cancer: Predicted association strength = 1
Endometrial cancer: Predicted association strength = 0
Kidney cancer: Predicted association strength = 0
Lung Cancer: Predicted association strength = 1
Ovarian cancer: Predicted association strength = 1
Pancreatic cancer: Predicted association strength = 1
Perillyl alcohol administration for cancer treatment: Predicted association strength = 1
Prostate cancer: Predicted association strength = 1
Stomach cancer: Predicted association strength = 2
Thyroid cancer: Predicted association strength = 0

***Findings***

This research established Pearson correlation as the most appropriate statistical similarity method for metabolite–disease association analysis, demonstrating greater effectiveness than the Bhattacharyya and Chebyshev approaches. Furthermore, the Graph Attention Network (GAT) model trained and tested on the

similarity-based dataset achieved superior performance compared to the GAT model trained and tested on the TF-IDF dataset in identifying specific metabolite–disease relationships.

## 6. Conclusion

The findings highlight that correlation analysis effectively identifies links between metabolite and disease associations. Identifying the strength of these associations highlights potential pathways of disease development, supporting the idea that correcting metabolite imbalances may help prevent the onset of disease. This research demonstrated metabolite–disease association analysis through three developmental stages. In the first stage, the study provided insights showing that the capacity to capture metabolite–disease relationships varies depending on the similarity technique employed. For example, among the three similarity methods applied—Pearson, Bhattacharya, and Chebyshev—the Pearson correlation proved to be the most effective. In the second stage, the research advanced to the development of a Graph Attention Network (GAT) model capable of identifying associations between metabolites and diseases, categorized as strong, moderate, or weak. These outcomes suggested that meaningful knowledge discovery from metabolite–disease data is possible. In the third stage, instead of applying the Pearson similarity dataset to the GAT model, a TF-IDF dataset is used directly. The results indicated that the GAT model trained and tested on the Pearson similarity dataset performed better than the model trained and tested on the TF-IDF dataset in predicting metabolite–disease associations.

## References


1. Spelmen Vimalraj, S.; Porkodi Rajendran.: Convalecing the Process of Ranking Metabolites for Diseases using Subcellular Localization. Arabian Journal for Science and Engineering · Arabian Journal for Science and Engineering. 47:1619–1629 (2021).
2. Zhu, Keyun; Huang, Mengting; Wang, Yimeng; Gu, Yaxin; Li, Weihua; Liu, Guixia; Tang, Yun.: MetaPredictor: in silico prediction of drug metabolites based on deep language models with prompt engineering. Briefings in Bioinformatics · Briefings in Bioinformatics. 25(5): bbae374 (2024). https://doi.org/10.1093/bib/bbae374
3. Weng, Qinghui; Hu, Mingyi; Peng, Guohao; Zhu, Jinlin.: DMoVGPE: predicting gut microbial associated metabolite profiles with deep mixture of variational Gaussian Process experts. BMC Bioinformatics · BMC Bioinformatics. 26(1): 1–23 (2025). https://doi.org/10.1186/s12859-025-06110-7
4. Wang, Yongtian; Juan, Liran; Peng, Jiajie; Wang, Tao; Zang, Tianyi; Wang, Yadong.: Explore potential disease-related metabolites based on latent factor model. BMC Genomics · BMC Genomics. 23(Suppl 1): 269 (2022). https://doi.org/10.1186/s12864-022-08504-w
5. Liu, Wenzhi; Lu, Pengli.: Predicting Disease–Metabolite Associations Based on the Metapath Aggregation of Tripartite Heterogeneous Networks. Interdisciplinary Sciences: Computational Life Sciences · Interdisciplinary Sciences. 16: 829–843 (2024). https://doi.org/10.1007/s12539-024-00645-8
6. Ruiz-Moreno, Angel J.; Del Castillo-Izquierdo, Ángela; Tamargo-Rubio, Isabel; Fu, Jingyuan.: MicrobeRX: a tool for enzymatic-reaction-based metabolite prediction in the gut microbiome. Microbiome · Microbiome. 13: Article 2070 (2025). https://doi.org/10.1186/s40168-025-02070-5
7. Xiao, F.; Huang, C.; Chen, A.; Xiao, W.; Li, Z.: Identification of Metabolite–Disease Associations Based on Knowledge Graph. *Metabolomics · Metabolomics*. 21:32 (2025).
8. Ren, Sheng; Hinzman, Anna A.; Kang, Edward; Szczesniak, Robert D.; Lu, Long Jason: Computational and Statistical Analysis of Metabolomics Data. *Metabolomics · Metabolomics*. 11:1492–1513 (2015).
9. Sun, Feiyue; Sun, Jianqiang; Zhao, Qi: A Deep Learning Method for Predicting Metabolite–Disease Associations via Graph Neural Network. *Briefings in Bioinformatics · Briefings in Bioinformatics*. 23(4):bbac266 (2022).
10. Vaida, Maria; Wu, Jiawen; Himdiat, Eyad; Haince, Jean-François; Bux, Rashid A.; Huang, Guoyu; Tappia, Paramjit S.; Ramjiawan, Bram; Ford, W. Rand: M-GNN: A Graph Neural Network Framework for Lung Cancer Detection Using Metabolomics and Heterogeneous Graph Modeling. *International Journal of Molecular Sciences · IJMS*. 26(10):4655 (2025)