M3NetFlow: A novel multi-scale multi-hop graph AI model for integrative multi-omic data analysis

Heming Zhang[1], S. Peter Goedegebuure[2,3], Li Ding[3,4], David DeNardo[3,4], Ryan C. Fields[2,3], Yixin Chen[6], Philip Payne[1], Fuhai Li[1,5#]

[1]Institute for Informatics, Data Science and Biostatistics (I2DB), [2]Department of Surgery, [3]Siteman Cancer Center, [4]Department of medicine, [5]Department of Pediatrics, Washington University School of Medicine, [6]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA. [#]Correspondence: Fuhai.Li@wustl.edu

Summary: Multi-omic data-driven studies, characterizing complex disease signaling system from multiple levels, are at the forefront of precision medicine and healthcare. The integration and interpretation of multi-omic data are essential for identifying molecular targets and deciphering core signaling pathways of complex diseases. However, it remains an open problem due the large number of biomarkers and complex interactions among them. In this study, we propose a novel Multi-scale Multi-hop Multi-omic graph model, *M3NetFlow*, to facilitate generic multi-omic data analysis to rank targets and infer core signaling flows/pathways. To evaluate M3NetFlow, we applied it in two independent multi-omic case studies: 1) uncovering mechanisms of synergistic drug combination response (defined as anchor-target guided learning), and 2) identifying biomarkers and pathways of Alzheimer 's disease (AD). The evaluation and comparison results showed *M3NetFlow* achieves the best prediction accuracy (accurate), and identifies a set of essential targets and core signaling pathways (interpretable). The model can be directly applied to other multi-omic data-driven studies. The code is publicly accessible at: https://github.com/FuhaiLiAiLab/M3NetFlow

1. Introduction

Multi-omic data-driven studies are at the forefront of precision medicine and healthcare. Recently, multi-omic datasets, like genetic, epigenetic, transcriptomic, and proteomic, have been being generated to characterize dysfunctional biological processes and signaling pathways from multiple levels/views, and to elucidate the panoramic view of the disease pathogenesis [1–5]. For example, The Cancer Genome Atlas (TCGA) program have generated multi-omic datasets of over 20,000 samples spanning 33 cancer types, to understand the key molecular targets and signaling pathways of cancer [6]. Moreover, the multi-omic data of >10,000 cancer cell lines were profiled in the Cancer Cell Line Encyclopedia (CCLE) project, which are valuable to investigate the mechanism of cancer response to given drugs and drug combinations[7]. In addition, the multi-omic data of Alzheimer's disease (AD) are generated and publicly available in the ROSMAP[11] project to uncover the pathogenesis of AD. Also, the exceptional longevity (EL), like the Long-Life Family Study (LLFS) project, have been generating multi-omics data[11–13] to identify protective biomarkers and pathways for long and healthy life. The multi-omic data are valuable and essential for understanding the key molecular targets and mechanisms of diseases, identifying novel therapeutic targets, predicting effect drugs and drug cocktails to guide the development of precision medicine.

However, it remains an open problem and a challenging task for integrative and interpretable omics data analysis, though a set of computational models has been proposed. A comprehensive review of existing multi-omics data integration analysis models was reported[16]. Specifically, these models were clustered into a few categories, like similarity, correlation, Bayesian, multivariate, fusion and network-based models. The PAthway Representation and Analysis by Direct Inference on Graphical Models (PARADIGM[17]) is one of the most widely used methods among these traditional computational methods. Aside from that, the NMTF-based method called iCell[18] was also proposed to integrate the multi-omics data and only the direct

connection was considered. The timeOmics[22] was recently developed to identify signaling patterns over time of biomarkers in multi-omics data, allowing for the analysis of time-series data in biological systems.

Problem formulation. The problem to be tackled in this study using graph AI models is to identify or rank disease or drug response associated molecular biomarkers from a large number of multi-omic features, and infer the potential signaling network among the selected biomarkers. The two specific biological problems to be tackled using the graph AI models are as follows. The first task is an anchor-biomarker (like known molecular targets, or drug targets) guided learning, e.g., to identify targets and pathways regulating drug cocktail response by integrating multi-omic data of cells and drug targets. The input data of the first case study is the multi-omic profiling of individual cell lines, kegg signaling pathways (graph), drug targets, and synergistic scores of a set of drug combinations. The output of the model is a set of top-ranked biomarkers and signaling pathways linking to the given anchor-biomarkers (information flowing to the anchor-biomarkers), which are informative to predict and explain the mechanism of drug combination response. In addition to drug-targets, the anchor-biomarkers can be generic given targets of interested to study the up- or down-stream signaling pathways of the anchor-biomarkers. The second task is to identify disease associated biomarkers and signaling pathways of Alzheimer's disease (AD). It is a generic biomarker ranking and pathway inference problem without the guide/constraint of anchor-biomarkers. The input is multi-omic data, like the AD and control/normal samples and kegg signaling pathways (graph). The output of the model is a set of top-ranked biomarkers and signaling pathways, which are informative to classify AD from normal samples explaining the pathogenesis of AD. The two biological problems/case studies represent two of the most needed multi-omic data analysis tasks.

Related work. Graph Neural Networks (GNNs) have gained prominence due to their capability to model relationships within graph-structured data[25–28]. And numerous studies have applied the Graph Neural Network (GNN) with the integration of the multi-omics data. MOGONET[29] initially creates similarity graphs among samples by leveraging each omics data, then employs a Graph Convolutional Network (GCN[25]) to learn a label distribution from each omics data independently. Subsequently, a Cross-omics discovery tensor is implemented to refine the prediction by learning the dependency among multi-omics data. MoGCN[30] adopts a similar approach by constructing a patient similarity network using multi-omics data and then using GCN to predict the cancer subtype of patients. GCN-SC[31] utilizes a GCN to combine single-cell multi-omics data derived from varying sequencing methodologies. MOGCL[32] takes this further by exploiting the potency of graph contrastive learning to pretrain the GCN on the multi-omics dataset, thereby achieving impressive results in downstream tasks with fine-tuning. Nevertheless, none of the aforementioned techniques contemplate incorporating structured signaling data like KEGG into the model. Moreover, general GNN models are limited by their expression power, i.e., the low-pass filtering or over-smoothing issues, which hampers their ability to incorporate many layers. The over-smoothing problem was firstly mentioned by extending the propagation layers in GCN[33]. Moreover, theoretical papers using Dirichlet energy showed diminished discriminative power by increasing the propagation layers[34]. And multiple attempts were made to compare the expressive power of the GCNs[28,35], and it is shown that WL subtree kernel[36] is insufficient for capturing the graph structure. Hence, to improve the expression powerful of GNN, the $\boldsymbol{K}$-hop information of local substructure was considered in various recent research[37–42]. However, none of these studies was specifically designed to well integrate the biological regulatory network and provide the interpretation with important edges and nodes.

In this study, we present a novel graph AI model, named ***M3NetFlow (Multi-scale, Multi-Hop, Multi-omics NETwork Flow inference)*** to address the challenges mentioned above.

Specifically, using an attention mechanism, our approach tackles these challenges by first incorporating local K-hop information within each subgraph by integrating the gene regulatory network such as KEGG[43] and BioGrid[44,45]. Subsequently, we establish global-level bi-directional propagation to facilitate message passing between genes/proteins. In the interpretation phase, we leverage the attention mechanism to aggregate the weights of all connected paths for genes, and a reweighting process inspired by (Term Frequency – Inverse Document Frequency) TF-IDF[46] is employed to redistribute the importance scores of genes across different cell lines. Afterward, a visualization tool, ***NetFlowVis,*** was developed for the better analysis of targets and signaling pathways of drugs and drug combinations. To assess and demonstrate the effectiveness of our proposed model, ***M3NetFlow***, applied it in two independent multi-omic case studies: 1) uncovering mechanisms of synergistic drug combination response (defined as anchor-target guided learning), and 2) identifying biomarkers and pathways of Alzheimer 's disease (AD). The evaluation and comparison results showed *M3NetFlow* achieves the best prediction accuracy (accurate), and identifies a set of essential targets and core signaling pathways (interpretable). The model can be directly applied to other multi-omic data-driven studies.

2. Methodology

Table 1 provided the links from which four types of datasets, drug combination effects, multi-omics data, gene-gene interactions, and drug-gene interactions, were collected to predict drug combinations for cancer cell lines. What's more, to investigate Alzheimer's Disease, we also obtained corresponding multi-omics datasets and clinical dataset from publicly available sources, particularly the ROSMAP datasets (see **Table 2**). Comprehensive details regarding the data and the preprocessing results are thoroughly documented in Appendix Section A. This section provides an in-depth explanation of the methodologies employed, as well as the specific parameters and outcomes obtained during the preprocessing stage.

Table 1. Multi-omics data and drug combination screening data of cancer cell lines.

Database Type	Database Name	Website Link
Drug Combination Effect Data	NCI ALMANAC[41] And O'Neil [42] drug combination	https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-ALMANAC and https://drugcomb.fimm.fi/
Multi-omics Data	Cell Model Passports[43] RNA-Seq	https://cog.sanger.ac.uk/cmp/download/rnaseq_20191101.zip
	Cell Model Passports[43] CNV	https://cog.sanger.ac.uk/cmp/download/cnv_20191101.zip
	CCLE[44] Methylation	https://data.broadinstitute.org/ccle/CCLE_RRBS_TSS1kb_20181022.txt.gz
	CCLE[44] Gene Amplification	https://data.broadinstitute.org/ccle_legacy_data/binary_calls_for_copy_number_and_mutation_data/CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct
	CCLE[44] Gene Deletion	https://data.broadinstitute.org/ccle_legacy_data/binary_calls_for_copy_number_and_mutation_data/CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct
Gene-Gene Interaction Data	KEGG[45] Gene Interactions Network	https://www.genome.jp/kegg/
Drug-Target Interaction Data	DrugBank[46]	https://go.drugbank.com/

Table 2. Multi-omics data and clinical data from ROSMAP database resources

Database Type	Database Name	Website Link
Clinical Data	ROSMAP_clinical	https://www.synapse.org/#!Synapse:syn3191087
Multi-omics Data	ROSMAP_arrayMethylation_imputed	https://www.synapse.org/#!Synapse:syn3168763
	ROSMAP.CNV.Matrix(Mutation)	https://www.synapse.org/#!Synapse:syn26263118
	ROSMAP_RNAseq_FPKM_gene	https://www.synapse.org/#!Synapse:syn3505720
	C2.median_polish_corrected_log2(Proteomics)	https://www.synapse.org/#!Synapse:syn21266454
Mapping Data	GEO GPL16304 Platform	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16304

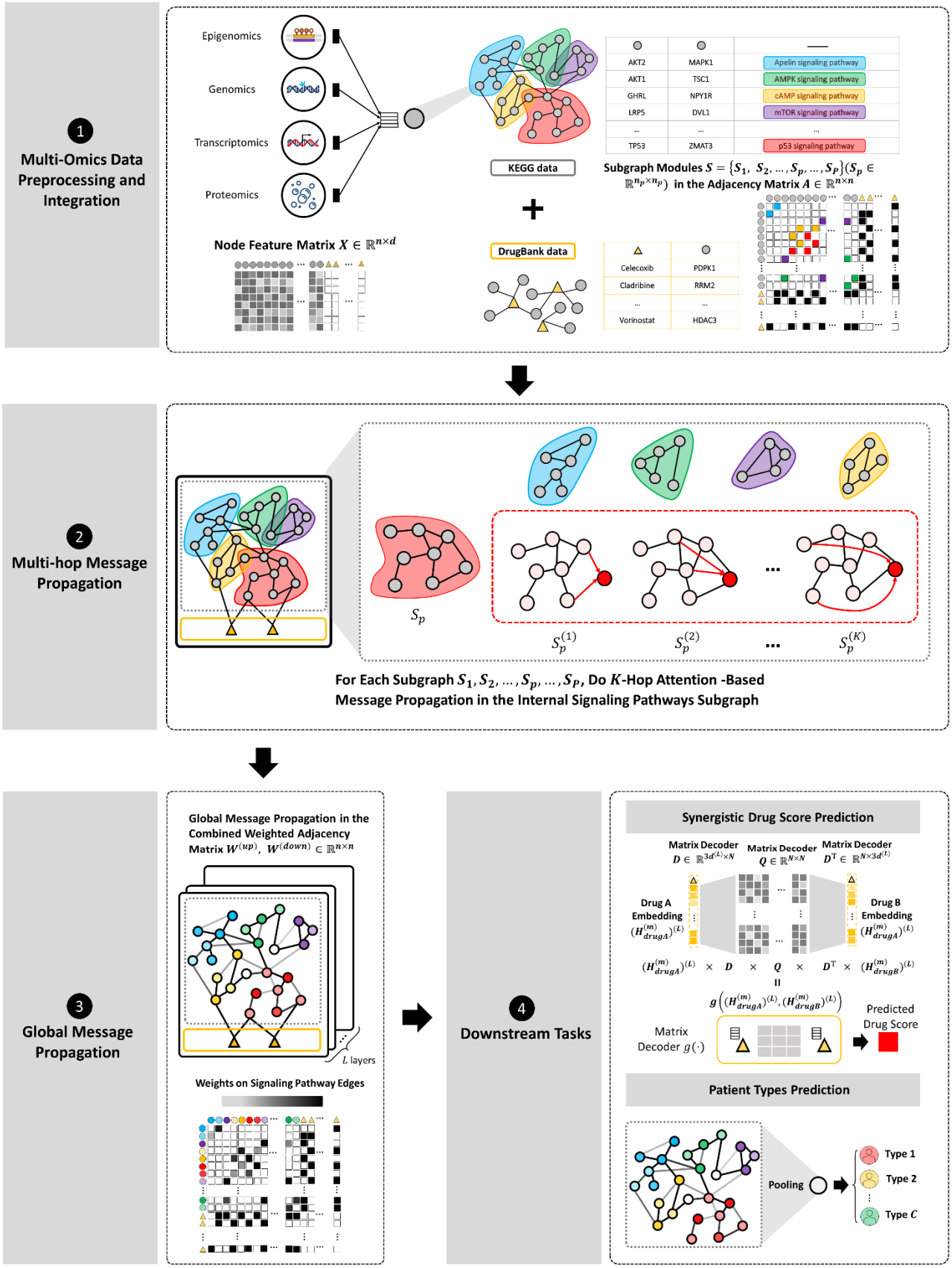
1
Multi-Omics Data Preprocessing and Integration
Epigenomics
Genomics
Transcriptomics
Proteomics
Node Feature Matrix $X \in \mathbb{R}^{n \times d}$
KEGG data
AKT2 MAPK1 Apelin signaling pathway
AKT1 TSC1 AMPK signaling pathway
GHRL NPY1R cAMP signaling pathway
LRP5 DVL1 mTOR signaling pathway
TP53 ZMAT3 p53 signaling pathway
Subgraph Modules $S = \{S_1, S_2, \dots, S_p, \dots, S_P\}(S_p \in \mathbb{R}^{n_p \times n_p})$ in the Adjacency Matrix $A \in \mathbb{R}^{n \times n}$
DrugBank data
Celecoxib PDPK1
Cladribine RRM2
Vorinostat HDAC3
2
Multi-hop Message Propagation
S_p
$S_p^{(1)}$ $S_p^{(2)}$ ... $S_p^{(K)}$
For Each Subgraph $S_1, S_2, \dots, S_p, \dots, S_P$, Do K-Hop Attention -Based Message Propagation in the Internal Signaling Pathways Subgraph
3
Global Message Propagation
Global Message Propagation in the Combined Weighted Adjacency Matrix $W^{(up)}$, $W^{(down)} \in \mathbb{R}^{n \times n}$
L layers
Weights on Signaling Pathway Edges
4
Downstream Tasks
Synergistic Drug Score Prediction
Matrix Decoder $D \in \mathbb{R}^{3d^{(L)} \times N}$
Matrix Decoder $Q \in \mathbb{R}^{N \times N}$
Matrix Decoder $D^{\mathrm{T}} \in \mathbb{R}^{N \times 3d^{(L)}}$
Drug A Embedding $(H_{drugA}^{(m)})^{(L)}$
Drug B Embedding $(H_{drugA}^{(m)})^{(L)}$
$(H_{drugA}^{(m)})^{(L)} \times D \times Q \times D^{\mathrm{T}} \times (H_{drugB}^{(m)})^{(L)}$
$g\left((H_{drugA}^{(m)})^{(L)}, (H_{drugB}^{(m)})^{(L)}\right)$
Matrix Decoder $g(\cdot)$
Predicted Drug Score
Patient Types Prediction
Pooling
Type 1
Type 2
Type C

Figure 1. Model architecture of *M3NetFlow.* ①Integrate the one-hot encoded and multi-omics features into a vector for each node. Afterwards, Merge drug-gene and gene-gene interactions into an adjacency matrix. ② Multi-hop attention-based propagation was performed in the subgraphs. ③ Use combined weighted adjacency matrix for global signaling propagation.④ Downstream tasks. (1) Decode the pair of drug nodes to predict the drug synergy scores. (2) Use pooling strategy to predict the patient outcomes.

2.1 Model Architecture of M3NetFlow

Figure 1 shows the schematic architecture of the proposed ***M3NetFlow*** model. The model input parameters are: $\mathcal{X} = \{(X^{(1)}, T^{(1)}),\ (X^{(2)}, T^{(2)}), \dots, (X^{(m)}, T^{(m)}), \dots, (X^{(M)}, T^{(M)})\}\ (X^{(m)} \in \mathbb{R}^{n\times d}, T \in \mathbb{R}^{n\times 2}), A \in \mathbb{R}^{n\times n}, S = \{S_1,\ S_2, \dots, S_p, \dots, S_P\}(S_p \in \mathbb{R}^{n_p\times n_p}),\ D_{in} \in \mathbb{R}^{n\times n},\ D_{out} \in \mathbb{R}^{n\times n}$, where M represents number of data points in the drug screening dataset. To predict the drug combination effect scores Y $(Y \in \mathbb{R}^{M\times 1})$, we would like to build up the machine learning model $f(\cdot)$ with $f(\mathcal{X}, A, S, D_{in}, D_{out}) = Y$, where $\mathcal{X}$ denotes all of the data points in the dataset and $(X^{(m)}, T^{(m)})$ is m-th data points in the dataset, where $X^{(m)}$ denotes the node features matrix with n nodes of d features and $T^{(m)}$ denotes the one-hot encoding of the number of drugs targeted on those n nodes. The matrix A is the adjacency matrix that demonstrates the node-node interactions, and the element in adjacency matrix A such as a_{ij} indicates an edge from i to j, and S is a set of subgraphs that partition the whole graph adjacent matrix A into multiple subgraphs with $S_p \in \mathbb{R}^{n_p\times n_p}$ of nodes interactions between its internal n_p nodes and each subgraph has its own corresponding subgraph node feature matrix $X_p \in \mathbb{R}^{n_p\times d}$. D_{in} is an in-degree diagonal matrix for nodes in directed graph, and D_{out} is an out-degree diagonal matrix for nodes in directed graph.

Network Modular and Multi-hop Message Propagation In the graph message passing stages of our architecture (see **Figure 1** step3 and step4), the multi-scale design, i.e., the local network module/subgraph module message passing stage for each signaling pathway, and the global message passing stage were designed. The multi-scale design will ensure the message fully interacts in the internal subgraph. Moreover, we opted for a K-hop attention-based graph neural

network because it further allows for the consideration of longer distance information, which can be interpreted as signaling flow in a single message propagation layer. With those initial embedding features $X^{(m)} \in \mathbb{R}^{n\times d}, (m = 1, 2, \ldots, M)$, the K-hop attention-based graph neural network was built to incorporate the long-distance information for each subgraph with

$$\left(\alpha_{ij}^{(m)}\right)_p^{(k)} = \mathrm{ATT}_{hop}\left[X_p^{(m)}\right] \#(1)$$

$$\left(H_p^{(m)}\right)_i = \mathrm{MSG}_{hop}\left[X_p^{(m)}, \left(\alpha_{ij}^{(m)}\right)_p^{(k)}\right] \#(2)$$

, where $\left(\alpha_{ij}^{(m)}\right)_p^{(k)}$ is the attention score between node i and node j in the k-th kop of the subgraph S_p for the data point m and the updated node feature $\left(H_p^{(m)}\right)_i$ for node i will be generated via K-hop message propagation (see **Appendix B.1** for details).

Global Bi-directional Message Propagation Following the message propagation in the multiple internal subgraphs, the global weighted bi-directional message propagation will be performed, where nodes-flow contains both 'upstream-to-downstream' (from up-stream signaling to drug targets) and 'downstream-to-upstream' (from drug targets to down-stream signaling) (see **Figure 1** step 3). Before the global level message propagation, the node feature for data point m in each subgraph $H_p^{(m)}$ $(p = 1, 2, \ldots, P)$ will be combined into a new unified node features matrix $\left(H^{(m)}\right)^{(0)}$ as the initial node features with

$$\left(H^{(m)}\right)^{(0)}{}_i = \frac{1}{\sum_{p=1}^{P} I[\mathcal{V}_i \in S_p]} \sum_{p=1}^{P} \left(H_p^{(m)}\right)_i \cdot I[\mathcal{V}_i \in S_p] \quad (3)$$

, where V_i represents the node/vertex i in the graph and $I\left[V_i \in S_p\right]$ is the indicator function, whose value will be one if $\mathcal{V}_i \in S_p$. In the end, the initial node features for global level propagation will be $\left(H^{(m)}\right)^{(0)} \in \mathbb{R}^{n\times d^{(0)}}, d^{(0)} = d' + d$. Then $(H^{(m)})^{(L)}$ will be generated by weighted bi-directional message propagation via

$$\left(H^{(m)}\right)^{(L)} = \mathrm{MPN}\left(\left(H^{(m)}\right)^{(0)}\right) \#(4)$$

, where MPN is the global bi-directional message propagation network (see **Appendix B.2** for details).

2.2 Downstream tasks

Drug combination response prediction by decoding drug node embeddings. After obtaining the embedded node features $(H^{(m)})^{(L)} \in \mathbb{R}^{n\times 3d^{(L)}}$ from the global message passing network, the features for drug A are represented as $(H^{(m)}_{drugA})^{(L)} \in \mathbb{R}^{1\times 3d^{(L)}}$ and the features for drug B are represented as $(H^{(m)}_{drugB})^{(L)} \in \mathbb{R}^{1\times 3d^{(L)}}$. Utilizing the decagon decoder[47], the prediction of combo score will be calculated in the following equation:

$$g\left((H^{(m)}_{drugA})^{(L)}, (H^{(m)}_{drugB})^{(L)}\right) = (H^{(m)}_{drugA})^{(L)} D U D^{\mathrm{T}} ((H^{(m)}_{drugB})^{(L)})^{\mathrm{T}} \ \#(5)$$

, where $D \in \mathbb{R}^{3d^{(L)}\times E}$ and $U \in \mathbb{R}^{E\times E}$ are trainable decoder matrices, as illustrated in Figure 1 from step 4.

Patient outcome predictions. With the embedded node features, the global mean pooling strategy was applied to predict the patient outcome with

$$\widehat{y^{(m)}} = \arg\max\left(\mathrm{MLP}\left(\mathrm{AVG}\left[\left(H^{(m)}\right)^{(L)}\right]\right)\right) \#(6)$$

, where $\widehat{y^{(m)}} \in \mathbb{R}^{C}$ and C is the number of sample types (see Figure 1 step 4).

2.3 Drug combination response interpretation using subgraph attention scoring

As aforementioned, we added the K-hop subgraph message propagation with attention in each subgraph, and the attentions or weights on each edge can potentially indicate and interpret the

signaling flow on the signaling network to affect the drug combination response. Specifically, putting the attentions/weights on each edge will use the trained parameters $a \in \mathbb{R}^{2d'}$ and $W \in \mathbb{R}^{d \times d'}$. And for cell lines set, the set $\mathcal{C} = \{C_1,\ C_2, \dots, C_r, \dots, C_R\}$ contains $R = 39$ cell lines, and we can choose the input node feature matrix within a specific cell line to form the cell-line specific attention matrix in k-hop by

$$(\alpha_{ij}^{(C_r)})^{(k)} = \sum_{p=1}^{P} \left(\alpha_{ij}^{(m)}\right)_p^{(k)}, \left(\forall X^{(m)} \in C_r\right) \tag{7}$$

, where $(\alpha_{ij}^{(C_r)})^{(k)}$ will be the element in $\left(\mathcal{A}^{(C_r)}\right)^{(k)}$, the k-hop attention matrix for cell line C_r. And above formula shows the calculation for the 1-head attention matrix. For the multi-head attention mechanism, the averaged value will be used for the attention matrix.

3. Results

3.1 Experimental Setup

To validate the proposed model on drug prediction task, we implemented a 5-fold cross-validation approach. This method involved the use of a four-element vector, <D_A, D_B, C_C, S_{ABC}>, as the model input. In this tuple, D_A and D_B denote the specific drug combinations being analyzed. C_C represents the cell line name, which is integrated with multi-omics data to provide comprehensive contextual information. Lastly, S_{ABC} reflects the synergistic drug effect observed on the cell line C_C. In total, there are 2788 model input data points for the NCI ALMANAC dataset and 1008 model input data points for the O'Neil dataset. In each fold model, 4 folds of the data points will be used as the training dataset, and the rest 1-fold will be used as the testing dataset. For those 1489 genes, their 8 features indicate the RNA sequence number, copy number variation, gene amplification, gene deletion, gene methylation maximum and minimum value and whether they have a connection spanning D_A and D_B respectively. Regarding the AD

samples outcome prediction on ROSMAP dataset, we utilized 138 samples from the ROSMAP dataset, categorized by disease status (74 AD, 64 non-AD). To address the data imbalance, we performed downsampling for the classification task. For the AD vs. non-AD classification, we downsampled the AD samples to match the 64 non-AD samples, resulting in a dataset with 64 AD and 64 non-AD samples. Afterwards, 5-fold cross-validation was leveraged to evaluate the performance of our model. For those 2144 genes, their 10 features indicate the methylation values on upstream, distal promoters, proximal promoters, core promoters and downstream, the genetic mutations with (number of duplicates, deletions and mCNV), the gene expression values and protein expression values. For each drug pair, their 8 features in drug prediction task or 10 features in ROSMAP sample outcome prediction task were initiated with zeros. Furthermore, to indicate connections between nodes, adjacency matrices were also created. Those matrices were formed from the KEGG, which contains gene pairs with sources and destinations**.** For drug-gene edges, the connections are bidirectional, which means that in adjacency matrices, those elements are symmetric.

3.2 Hyperparameters

Subsequently, a model was developed by using pytorch and torch geometric. For both drug prediction and ROSMAP sample outcome prediction tasks, the learning rate started at 0.002 and was reduced equally within each batch for a certain epoch stage. And the epochs after 60 will keep the learning rate at 0.0001. Adam optimizer was chosen for optimization with eps=1e-7 and weight_decay=1e-20. We empirically set the ***K-hop Subgraph Message Propagation*** part with $K = 3$ and the ***Global Bi-directional Message Propagation*** part with $L = 3$. Afterward, the feature dimensions will vary at the different layers and will be denoted by: $(d^{(1)}, d^{(2)}, \dots, d^{(l)}, \dots, d^{(L)})$. At the global message propagation part, the layer will concatenate biased node features and transformed node features of the previous layer for both upstream

and downstream, generating concatenated dimensions for the output dims being $3 \times d^{(l)}$ in l-th layer. Output dims of the previous layer served as the input dims for the current layer, as follows: (1) First layer (input dims, output dims): $(d^{(0)}, 3d^{(1)})$; (2) Second layer (input dims, output dims): $(3d^{(1)}, 3d^{(2)})$; (3) Third layer (input dims, output dims): $(3d^{(2)}, 3d^{(3)})$. The final embedded drug node dims were $3d^{(3)}, (L = 3)$. For drug prediction task, the decoder trainable transformation matrix dims: $D \in \mathbb{R}^{3d^{(L)} \times E}$ and $U \in \mathbb{R}^{E \times E}$ were used as trainable decoder matrices, with changeable parameters of E to adapt model performance and the model used $E = 150$. In ROSMAP sample prediction task, the trainable graph mean pooling was leveraged to predict the sample outcome. As for the LeakyReLU function, the parameter α was set as 0.1 for both tasks.

3.3 *M3NetFlow* improves drug combination synergy and AD prediction accuracy.

To evaluate the model performance in terms of synergy score prediction for drug combinations and predictions on ROSMAP AD samples, we conducted 5-fold cross-validation. As shown in **Table 4**, the average prediction (using the Pearson correlation coefficient), was about 61% Pearson correlation using the test data in the NCI ALMANAC dataset and was about 64% Pearson correlation using the test data in the O'Neil dataset. Regarding the ROSMAP dataset, the average accuracy was about 66% using the test data in the ROSMAP dataset (see **Table 5**). These prediction results are comparable with existing deep learning models[47,48]. Moreover, we also compared our proposed model M3NetFlow with other deep learning models (see **Tables 4-5** and **Figure 2a**),), which included the GCN[49] , Graph Attention network[50] (GAT), UniMP[51], MixHop[39], Principal Neighborhood Aggregation[52] (PNA) and GIN[53]. By checking the p values over 5-fold cross validation, the performance of the M3NetFlow have significant improvement over most of the GNN-based methods (see **Tables 4-5**).

Table 4. Model comparison using average Pearson correlation of 5-fold cross-validation using

both NCI ALMANAC and O'Neil datasets.

Dataset	NCI ALMANAC	NCI ALMANAC P value	O'Neil	O'Neil P value
GCN	51.93% ± 3.24%	0.001692	44.47% ± 6.02%	0.000338
GAT	49.16% ± 2.07%	0.060917	57.06% ± 3.40%	0.003399
UniMP	49.02% ± 4.21%	0.324693	55.84% ± 9.98%	0.116502
MixHop	57.78% ± 3.34%	0.000418	27.15% ± 9.14%	0.000166
PNA	55.63% ± 2.26%	0.179165	62.20% ± 2.07%	0.151676
GIN	53.76% ± 2.25%	0.000991	33.12% ± 9.03%	0.000387
M3NetFlow	**60.72% ± 0.70%**	**-**	**64.36% ± 2.31%**	**-**

Table 5. Model comparison using average accuracy of 5-fold cross-validation using ROSMAP dataset

Dataset	Accuracy	P value
GCN	59.43% ± 4.53%	0.049650832
GAT	62.80% ± 7.11%	0.332946723
UniMP	61.83% ± 3.78%	0.124477379
MixHop	57.20% ± 3.92%	0.014849014
PNA	57.83% ± 1.82%	0.017236825
GIN	49.83% ± 5.71%	0.001882831
M3NetFlow	67.34% ± 5.12%	**-**

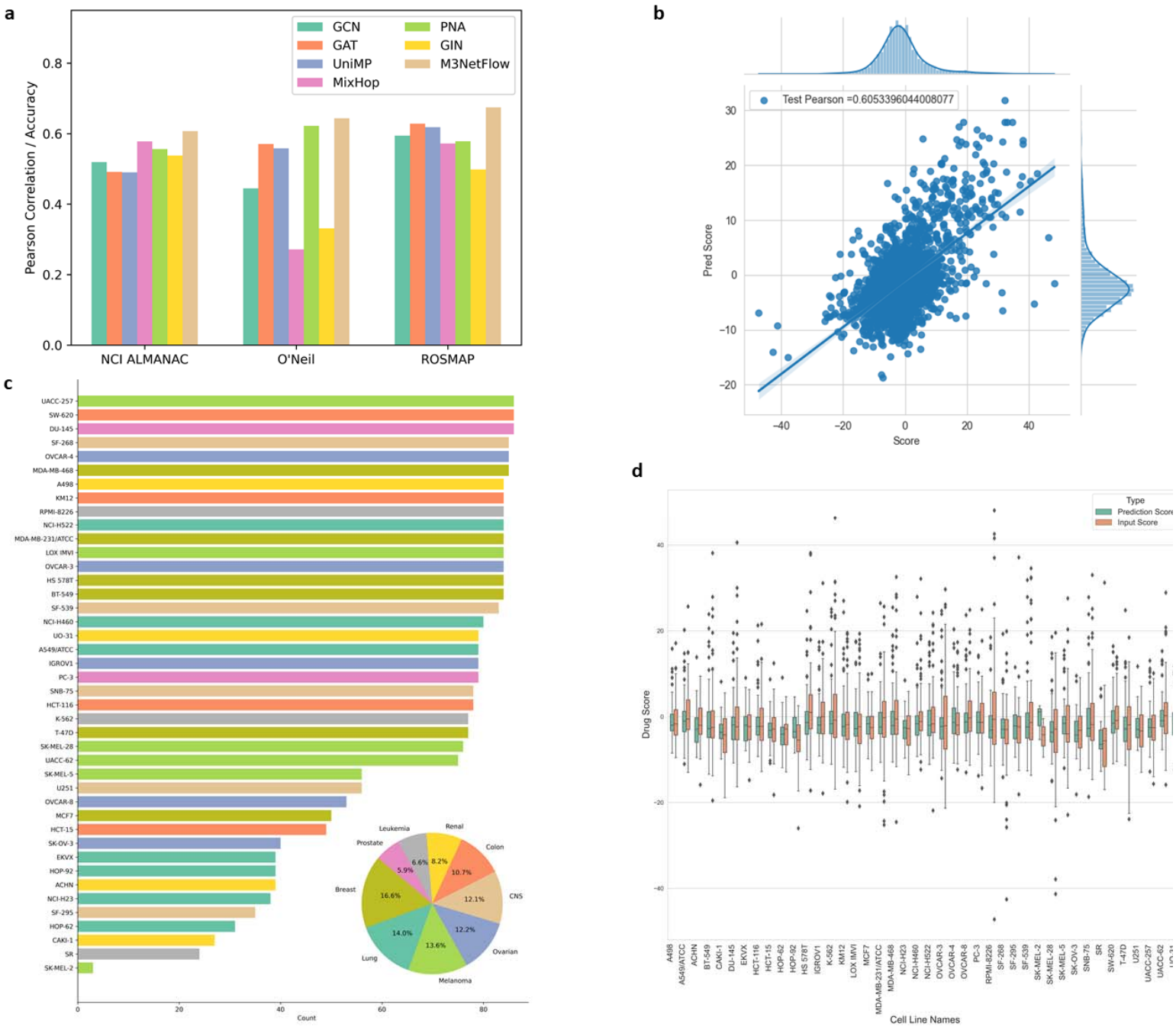

Figure 2. Model performance and overview of input dataset NCI ALMANAC, O'Neil (drug combination multi-omic data) and ROSMAP (AD multi-omic data). **a.** Pearson correlation comparisons for GCN, GAT, UniMP, MixHop, PNA, GIN and ***M3NetFlow*** models for NCI ALMANAC, O'Neil and ROSMAP dataset. **b.** Scatter plot of the model with data points in whole NCI ALMANAC dataset. **c.** Distributions of all cell lines in whole NCI ALMANAC dataset. **d.** Box plots across all cell lines in the NCI ALMANAC dataset.

3.4 *M3NetFlow* ranks important targets via attention score

[54]Based on the attention matrix of each sample (like a cell line or an AD patient sample) between neighboring nodes (genes) on the graph in each fold of cross validation, the average attention matrix in was calculated. Therefore, the weighted importance of each node (gene) of each sample will be calculated based on the attention with

$$\mathcal{D}_g^{(C_r)} = \sum_i^n \left(\bar{\mathcal{A}}_{ig}^{(C_r)}\right)^{(1)} + \sum_j^n \left(\bar{\mathcal{A}}_{gj}^{(C_r)}\right)^{(1)} \tag{8}$$

, where $\left(\bar{\mathcal{A}}_{ig}^{(C_r)}\right)^{(1)}$ is the element of averaged 5-fold attention matrix from a sample C_r in the i-th row and j-th column and $\mathcal{D}_g^{(C_r)}$ is the node importance score for gene g in the sample C_r. Combining all the node importance vectors, the matrix $\mathcal{D} \in \mathbb{R}^{n \times R}$ will be generated, where n is the number of nodes, and R is the number of samples. However, some of genes may show the importance of relatively higher scores in each sample, which weakens the analysis of the sample-specific analysis. In this way, the idea of reweighting the gene importance score was created based on the TF-IDF to reweight the node importance in each sample with

$$\mathcal{W}_g = Log\left(\frac{R+1}{S_g + 0.01}\right) \tag{9}$$

$$\mathcal{D}'_g = \mathcal{W}_g \cdot \mathcal{D}_g \tag{10}$$

, where S_g is the number of samples with a higher node importance score than the threshold $\mathcal{S}$ in all samples. Here, the 95 percentile of node degree values in the whole matrix $\mathcal{D}$ was used for $\mathcal{S}$. Finally, the reweighted node degree (importance score) matrix $\mathcal{D}'$ will be generated. Based on the reweighted matrix $\mathcal{D}'$, the node (gene) importance score will be calculated. The sum of importance score of individual genes across samples of disease will be used as the disease importance score of genes.

3.5 Synergistic drug combinations are associated with the identified key targets

For each cell line, the highest 5, and lowest 5 drug scores and their corresponding drugs and

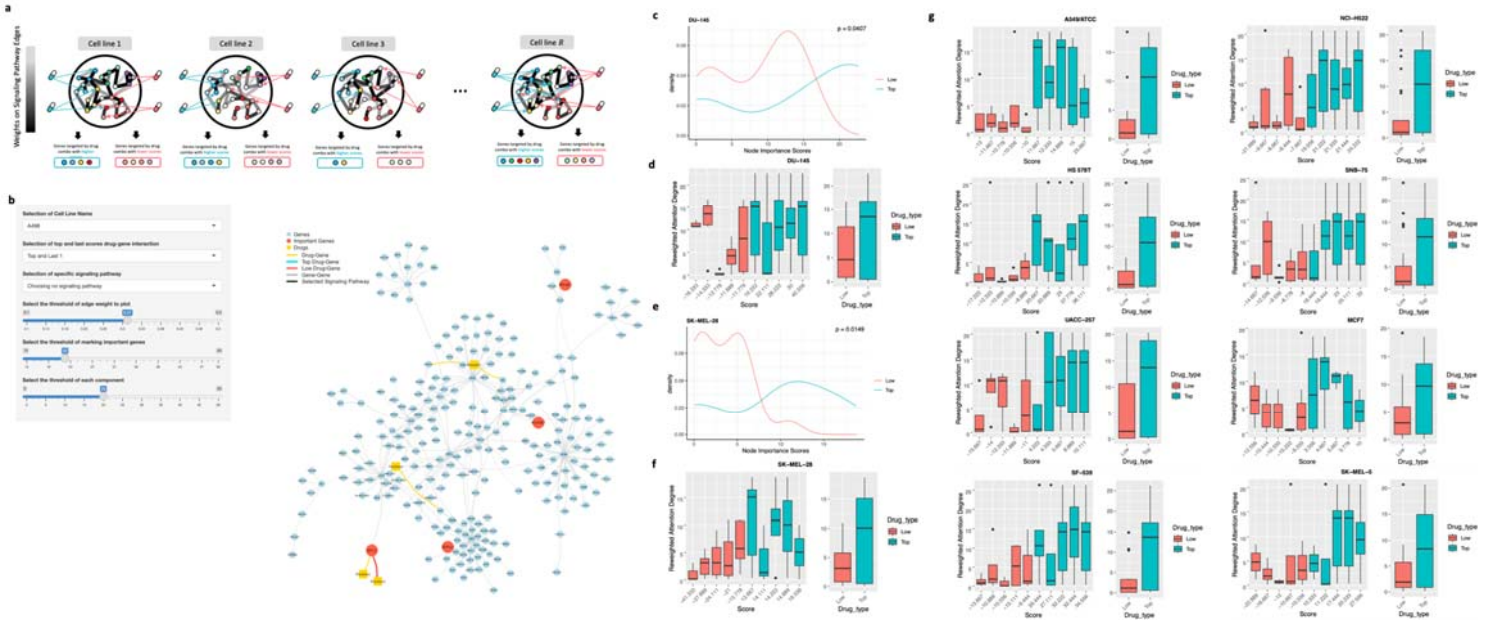

Figure 3. Validation and visualization through important node analysis **a.** Procedures of analyzing the important genes targeted by by the highest 5 and lowest 5 drug combinations **b.** Visualization tool ***NetFlowVis*** for core signaling network interactions of cell line DU-145 **c-f.** Density and box plots by comparing the genes targeted by the highest 5 and lowest 5 drug combinations in DU-145 and SK-MEL-28 cell lines **g.** Selected box plots by comparing the genes targeted by the highest 5 and lowest 5 drug combinations in cell lines with 8 lowest p-values (check **Table S1** for details of the p-values by comparing the gene important score lists targeted by the highest 5 and lowest 5 drug combinations.

drug-targeted genes were collected (check **Figure 3a**). Comparing the targeted genes by drugs from the highest 5 drug scores and lowest drug scores, the differences are statistically significant with p values in more than half cell lines. Since the SK-MEL-2 cell line only has 3 drug screen data points and all of them are smaller than 0, this cell line was excluded from the statistical analysis.

In total, there are 27 out of 41 (~65%) cell lines have p values smaller than 0.1 (filled with deep orange color in **Table S1**) and 32 out of 41 (~78%) cell lines have p value smaller than 0.3 (filled with light orange color in **Table S1**). And 37 out of 41 (~90%) cell lines have higher gene

importance scores on both mean and median values (filled with light green in **Table S1**). **Figures 4-6** provided more explicit evidence of a trend where genes targeted by drugs with higher scores exhibited higher node degrees (gene importance scores). In each cell line, the left boxplots compared the node degree distribution of genes targeted by the top 5 drugs with the highest scores against those targeted by the bottom 5 drugs with the lowest scores. The right boxplots in

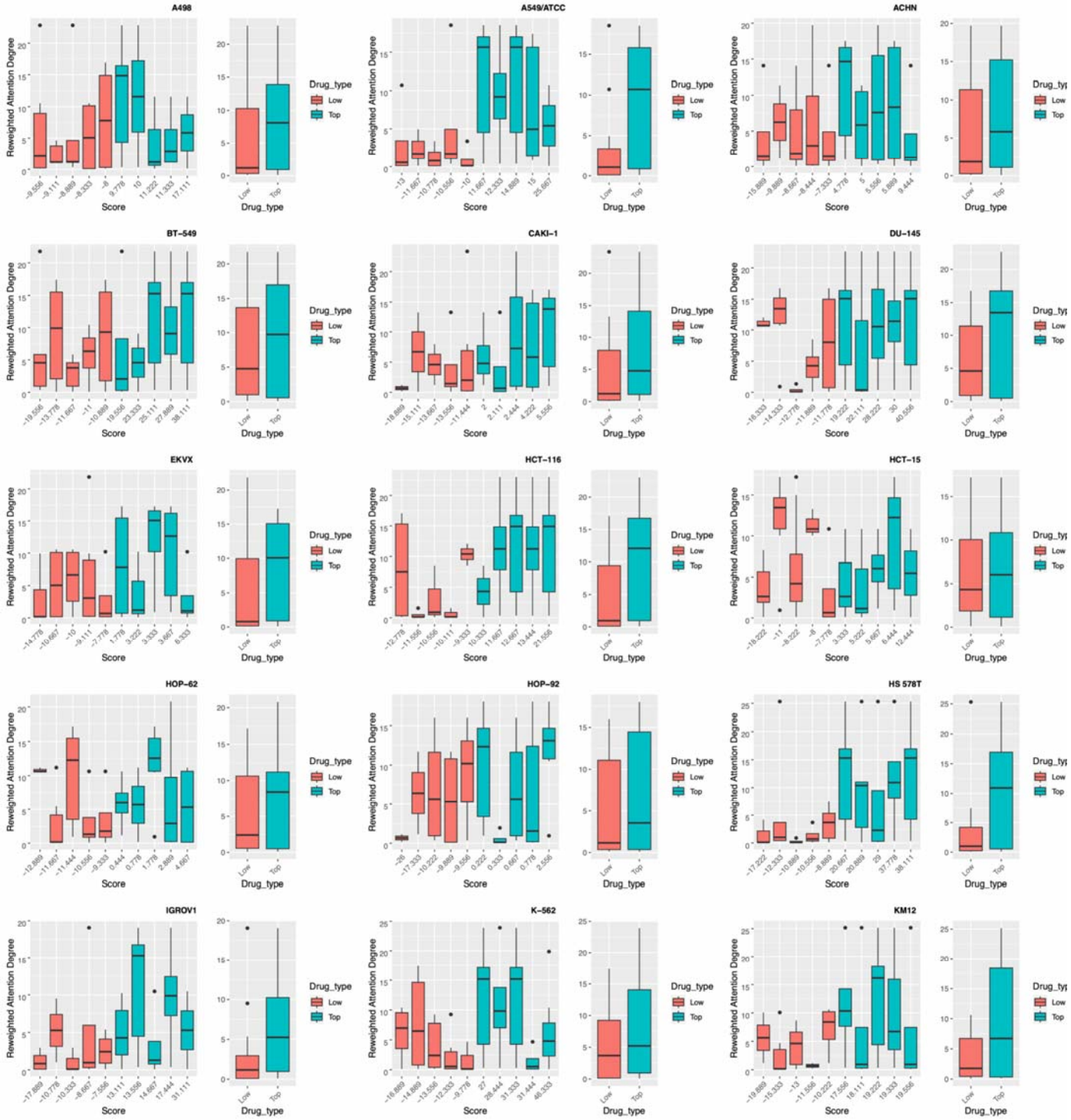

Figure 4 Cell line A498, A549/ATCC, ACHN, BT-549, CAKI-1, DU-145, EKVX, HCT-116, HCT-15, HOP-62, HOP-92, HS 578T, IGROV1, K-562, KM12 gene degree distribution targeted by highest 5 and lowest 5 drug scores

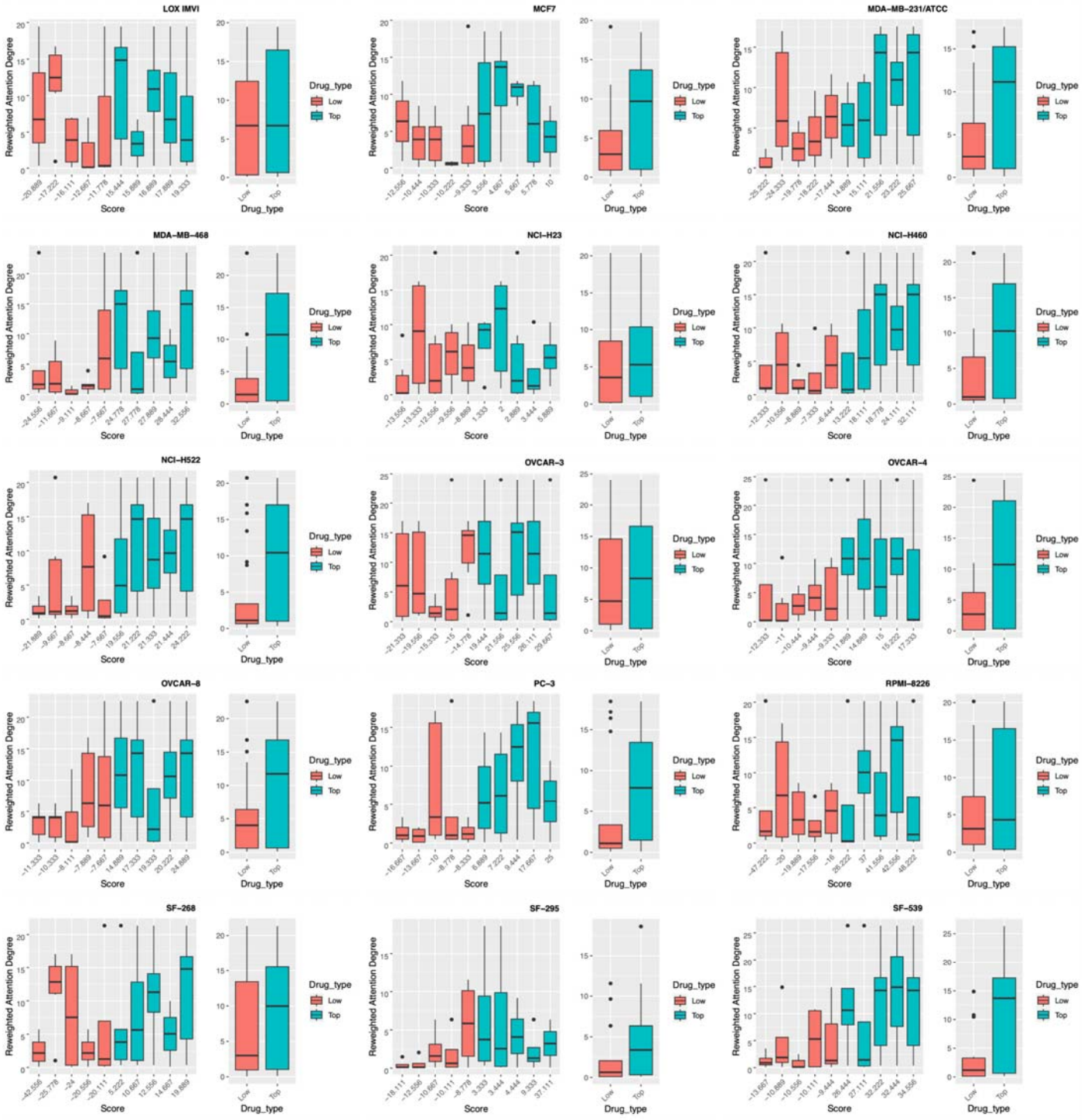

Figure 5 Cell line LOX IMVI, MCF7, MDA-MB-231/ATCC, MDA-MB-468, NCI-H23, HCI-H460, NCI-H522, OVCAR-3, OVCAR-4, OVCAR-8, PC-3, RPMI-8226, SF-268, SF-295, SF-539 gene degree distribution targeted by highest 5 and lowest 5 drug scores

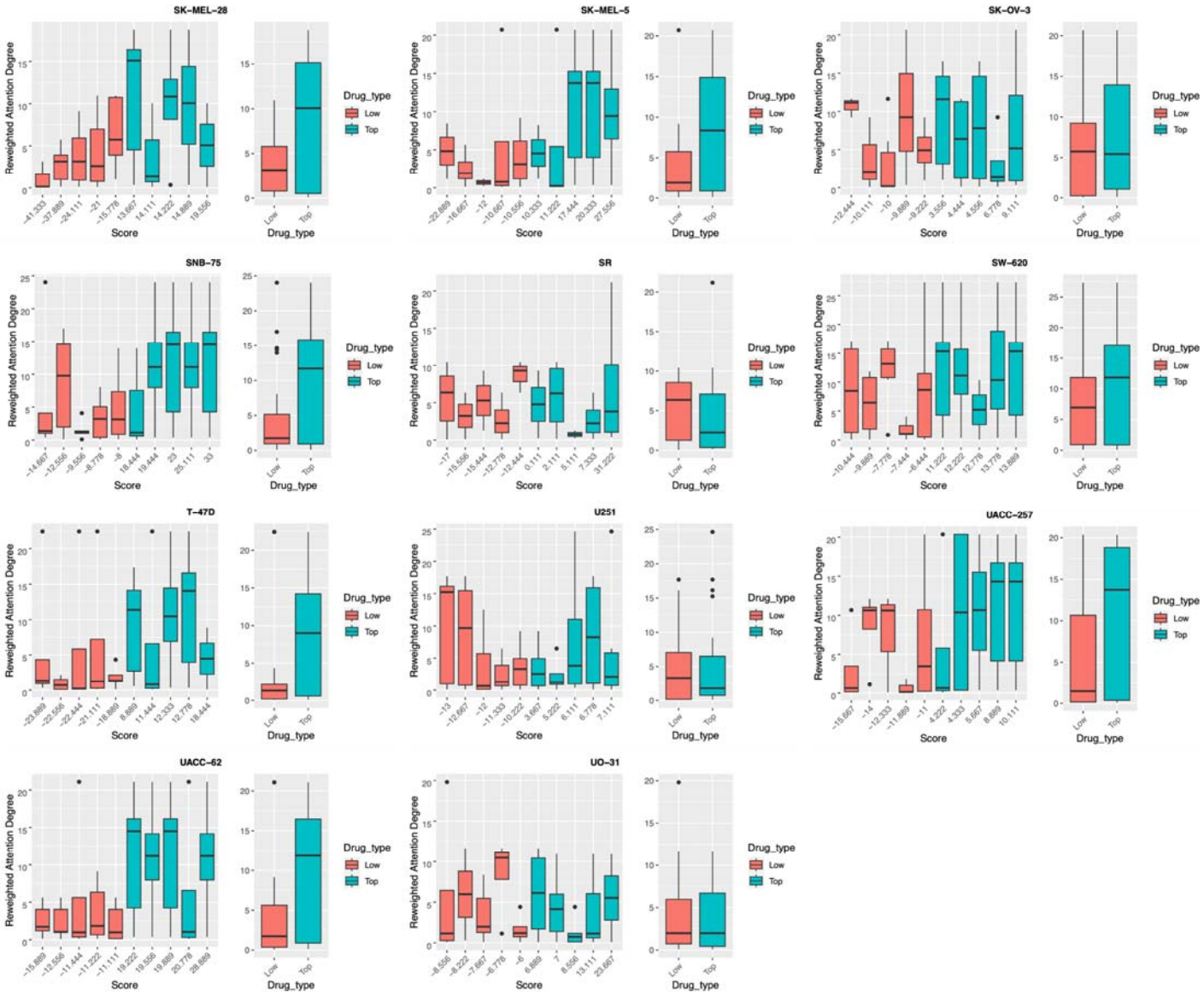

Figure 6 Cell line SK-MEL-28, SK-MEL-5, SK-OV-3, SNB-75, SR, SW-620, T-47D, U251, UACC-257, UACC-62, UO-31 gene degree distribution targeted by highest 5 and lowest 5 drug scores

each cell line represented the overall comparisons between the top 5 and bottom 5 drug scores. Consequently, we could assign prior targets to genes with the highest node degree (gene importance scores) using our calculations. This allowed us to rank the genes in each cell line according to their node degree (importance scores), which can be found in **Table S2**. We generated a list of the top 20 genes in each cell line, available in **Table S3**. Subsequently, we

identified overlapping genes among the cell lines of the same cancer type based on the top 20 genes in each line, as depicted in **Supplementary Figure S1**.

3.7 Alzheimer's Disease associated biomarkers and pathways

Based on the attention score derived node importance score (see Section 3.4), a set of AD associated targets and pathways are identified (see **Figure 7**). To evaluate the top-ranked targets, the top-ranked 20 AD associated biomarkers were further analyzed via pathway enrichment analysis. Interestingly, a set of AD associated signaling pathways are identified. As seen in **Figure 8**, the top-ranked targets are involved in a set of signaling transduction pathways, which indicates the importance of these targets. For example, the Phospholipase D signaling pathway[55,56] and sphingolipid signaling pathway[51] play important roles in the pathogenesis of AD. The dysfunction of mammalian target of rapamycin (mTOR) signaling is associated with $A\beta$ and Tau aggregation in AD[57]. Moreover, the PI3K/AKT signaling pathway, interacting with mTOR, is also correlated with the biological process of neurodegeneration [58]. Moreover, the MAPK signaling pathways is related to glucose metabolism, and affect $A\beta$ aggregation. In addition to these signaling pathways, the immune signaling pathways, i.e., [59]the T cell receptor (TCR) signaling pathway and Fc epsilon RI signaling pathway. It has also been implicated in AD, with studies showing that these immune pathways are associated with neuroinflammation[60,61,62]. Moreover, the *long-term depression* **(**LTD) pathway is enriched, which plays important roles in synaptic plasticity, a process essential for learning and memory. For example, studies have shown that amyloid-beta (Aβ) oligomers, which accumulate in the brains of Alzheimer's patients, can enhance LTD by interacting with NMDA receptors, leading to synaptic weakening and memory deficits[63]. The overactivation of metabotropic glutamate receptors (mGluRs) in the LTD pathway has been implicated in AD, where it promotes Aβ production and further synaptic depression[64]. In addition, the insulin signaling pathway (hsa04910) has been increasingly recognized for its role in the pathogenesis of Alzheimer's disease (AD). Disruption of insulin

signaling, often seen in type 2 diabetes, can exacerbate AD progression by impairing neuronal glucose metabolism and increasing oxidative stress[65]. For example, GSK3B is involved in tau phosphorylation, a process that leads to the formation of neurofibrillary tangles[66]. Also, the activation of MAPK8 (JNK) and MAP2K1 (MEK1) can lead to apoptotic signaling and further neuronal damage, contributing to cognitive decline[67]. Additionally, the insulin receptor (INSR) and related molecules like PIK3R3 and FOXO1 are involved in regulating synaptic plasticity and neuronal survival, which are crucial for maintaining cognitive function. In summary, the identified genes are involved in the complex pathogenesis of AD.

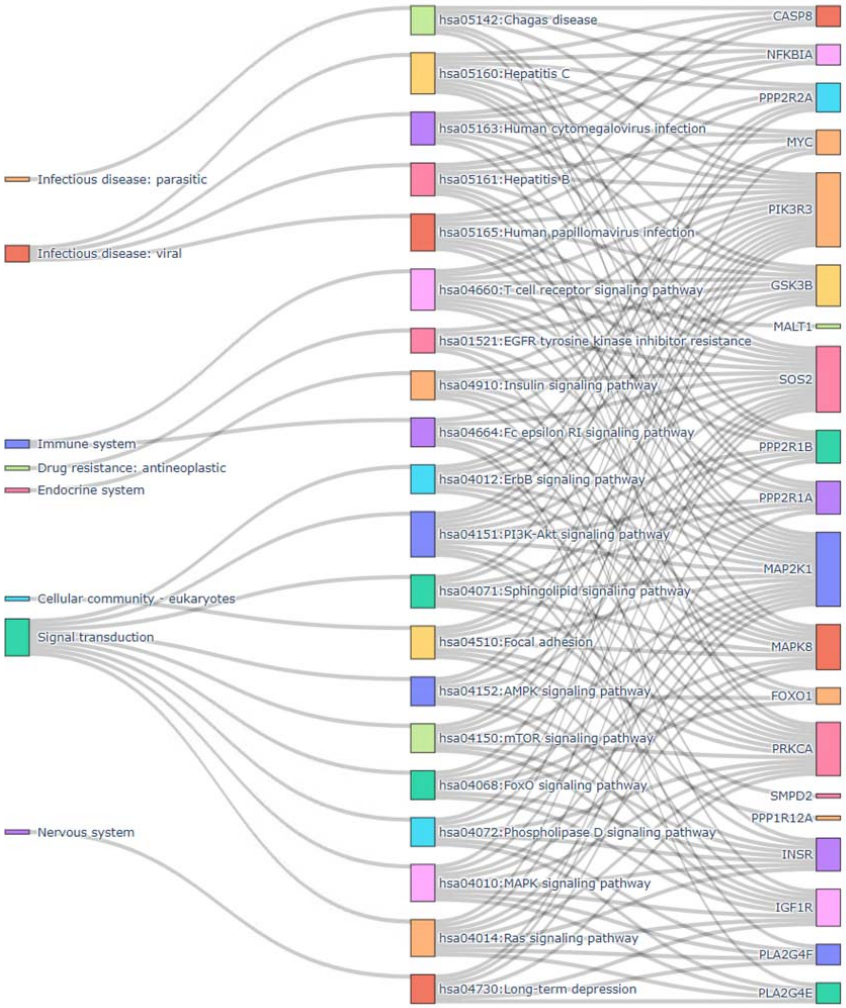

Figure 7. Sankey plot of enriched signaling pathways of the identified AD associated genes.

3.6 *NetFlowVis*: a tool to visualize core signaling pathways associated with synergistic drug combinations on specific cancer cell lines.

To visualize the results and gain a better understanding of the underlying mechanism of drug effects, we generated a core network of signaling interactions by applying a threshold. To enhance the interactions, we utilized the RShiny package and developed a visualization tool called ***NetFlowVis*** (The website link has been provided in the code availability part). This tool

allows users to control the edge threshold and set a minimum number of nodes in each network component. Users have the option to select the desired cell line for visualization and mark specific signaling pathways for detailed analysis. As an example, **Figure 4b** demonstrates the visualization of the cell line DU-145. Furthermore, by referring to the top 20 prior gene targets and their corresponding gene degree (node importance scores) for cell line DU-145, we observed that the majority of gene targets were included in the filtered signaling network interactions depicted in **Figure 4b**.

By utilizing the ***NetFlowVis*** visualization tool, users could take control of the size of the core signaling network. In the case of the DU-145 cell line, an edge threshold of 0.31 was selected, which corresponded to approximately the 98.5th percentile for edge weight among all edges. Additionally, a threshold of 23 was chosen for marking important nodes/genes in red, representing ~99.9th percentile for node importance scores across all nodes. To gain a deeper understanding of the underlying mechanism, the interactions between drugs and genes were analyzed, with the highest and lowest drug-gene interactions being represented by the colors purple and green, respectively. The analysis revealed that the drug ***Paclitaxel*** targeted the important gene ***BCL2***, while the drug ***Celecoxib*** targeted less essential genes (see **Figure 4b**). This finding demonstrated that targeting the important genes selected by the ***M3NetFlow*** model led to higher drug scores, as depicted in **Figure 4c-f**. Moreover, by choosing a specific signaling pathway within the core signaling network, it was observed that the ***IL-17 signaling pathway*** was downstream of the critical gene ***TRAF6***. Interestingly, experimental validation has shown the involvement of the ***IL-17 signaling pathway*** in promoting migration and invasion of ***DU-145*** cell lines through the upregulation of ***MTA1*** expression in prostate cancer[68].

4. Discussion and conclusion

Along with the advancement of next-generation sequencing (NGS) technology, multi-omics data

of diseases are being generated to characterize the dysfunctional molecular targets and signaling pathways of complex diseases, like cancer and AD, which are valuable and essential for the development of personalized medicine or precision medicine prediction. However, it remains a challenging task to identify key molecular targets and core signaling pathways from a large signaling network with thousands of signaling targets and extensive signaling interactions. Herein, we present a novel graph AI model, named M3NetFlow, for generic multi-omic data integrative analysis. In this model, multi-omic data are used as the numerical features of individual proteins, and the protein-protein interactions form a large-scale signaling graph. Unlike existing graph models, we divided the large-scale network into signaling modules and applied a multi-hop strategy to better infer the essential proteins and their interactions. The experimental evaluation on two real multi-omic data analysis applications showed that the proposed model outperformed existing graph models in accuracy and is able to identify essential targets and core signaling pathways, demonstrating its interpretability for Alzheimer's disease or drug combination synergy. The proposed model can be applied to 1) anchor-target guided or 2) generic target and pathway inference, as indicated by the two multi-omic data analysis applications that represent widely conducted tasks in omic data analysis. In the first scenario, a set of targets of interest, such as drug targets or known molecular targets, are used as the anchor nodes. Then, the signaling flows/information from neighboring graph nodes propagate to the anchor nodes. The embeddings of the anchor nodes are used as input for decoders to predict sample classes or drug combination responses. In the second scenario, the embeddings of graph nodes are pooled together as input for decoders to predict sample classes. Based on the attention scores of graph nodes, the essential targets and signaling pathways are further identified. Therefore, the proposed model can be applied for generic, integrative, and interpretable multi-omic data analysis tasks, and the code is publicly accessible.

It is still an exploratory study for multi-omic data analysis. There are some limitations that need

further investigation. For example, more signaling pathways and larger protein-protein interactions should be evaluated. Moreover, dividing large signaling graphs into subnetworks or network modules can be achieved by using biologically meaningful annotations, such as gene ontology (GO) terms. Additionally, more multi-omic datasets are being generated. Combining multi-omic data from different diseases can provide a larger sample size than individual disease datasets, which could improve the training or pre-training of graph AI models and help identify pan-disease or disease-specific targets. It is also interesting to expand graph models from tissue-level multi-omic data to single-cell multi-omic data, which can be more challenging due to the large number of single-cell samples. Therefore, novel and improved graph AI models are needed to integrate and interpret multi-omic datasets, identify and infer key molecular targets and signaling pathways of complex diseases, and guide the development of precision medicine.

Appendix

Section A. Data Collection and Preprocessing

A.1 Drug combination screening datasets

In this paper, we utilized the NCI ALMANAC and O'Neil[69] drug screening datasets to train the deep learning models. The NCI ALMANAC dataset consists of combo scores for various permutations of 104 FDA-approved drugs, representing their impact on tumor growth in NCI60 human tumor cell lines. For evaluating the synergy score of two drugs on a specific tumor cell line, we utilized the average combo-score of two drugs at different doses, employing a 4-element tuple: $<D_A, D_B, C_C, S_{ABC}>$. On the other hand, the O'Neil dataset, obtained from the DrugComb[70] platform, served as another drug combination screening dataset. We extracted the processed dataset from this platform, which provided average synergy Loewe scores. These scores assessed the synergy score of two drugs on a given tumor cell line, employing a 4-element tuple: $<D_A, D_B, C_C, S_{ABC}>$.

A.2 Multi-omics data of cancer cell lines

In this study, multi-omics data comprising RNA-seq, copy number variation, gene methylation, and gene mutation data for a total of 1,489 genes were incorporated into the model. These data were obtained from the Cell Model Passports[71] and CCLE[72] databases. Through the identification of overlapping cell lines from these databases, we identified 42 cell lines that were present in both the Cell Model Passports database and the NCI ALMANAC dataset, which are A498, A549/ATCC, ACHN, BT-549, CAKI-1, DU-145, EKVX, HCT-116, HCT-15, HOP-62, HOP-92, HS 578T, IGROV1, K-562, KM12, LOX IMVI, MCF7, MDA-MB-231/ATCC, MDA-MB-468, NCI-H23, NCI-H460, NCI-H522, OVCAR-3, OVCAR-4, OVCAR-8, PC-3, RPMI-8226, SF-268, SF-295, SF-539, SK-MEL-2, SK-MEL-28, SK-MEL-5, SK-OV-3, SNB-75, SR, SW-620, T-47D, U251, UACC-257, UACC-62, UO-31. Additionally, we found 24 cell lines that were present in both the Cell Model Passports database and the O'Neil dataset, which are A2058, A2780, A375, CAOV3, HCT116, HT144, LOVO, MDAMB436, NCI-H460, NCIH1650, NCIH2122, NCIH23, OV90, OVCAR3, RKO, RPMI7951, SK-OV-3, SKMEL30, SKMES1, SW-620, SW837, T-47D, UACC62, VCAP.

A.3 Multi-omic data and clinical phenotypes datasets from ROSMAP

Following the acquisition of the datasets, they were reformatted into 2-dimensional data frames, with columns dedicated to sample identifiers, such as IDs and names, and rows corresponding to probes, gene symbols, and gene IDs. To successfully integrate the multi-omics data with clinical information, it was essential to match identical samples across the various datasets. This required standardizing the row data—including probes, gene symbols, and gene IDs—into a unified gene-level format, either by aggregating gene-specific measurements or by resolving duplicates from gene synonyms. Genes were subsequently mapped to a reference genome to ensure precise annotation within the multi-omics datasets. Gene counts were then normalized across the datasets, with missing values imputed using zeros or negative ones as needed. After

aligning all columns to standard sample IDs and rows to standardized gene IDs, and ensuring a consistent number of samples and genes, the data was prepared for integration into Graph Neural Network (GNN) models, where epigenomics, genomics, transcriptomics and proteomics data were employed as features for nodes.

A.4 KEGG Signaling Pathways

About 59,241 gene-gene interactions for over 8,000 genes across different signaling pathways were collected from KEGG[73] database. And 48 signaling pathways in the KEGG dataset were selected as the ground truth of subgraphs in the gene-gene interactions, which were AGE-RAGE signaling pathway in diabetic complications, AMPK signaling pathway, Adipocytokine signaling pathway, Apelin signaling pathway, B cell receptor signaling pathway, C-type lectin receptor signaling pathway, Calcium signaling pathway, Chemokine signaling pathway, ErbB signaling pathway, Estrogen signaling pathway, Fc epsilon RI signaling pathway, FoxO signaling pathway, Glucagon signaling pathway, GnRH signaling pathway, HIF-1 signaling pathway, Hedgehog signaling pathway, Hippo signaling pathway, Hippo signaling pathway - multiple species, IL-17 signaling pathway, Insulin signaling pathway, JAK-STAT signaling pathway, MAPK signaling pathway, NF-kappa B signaling pathway, NOD-like receptor signaling pathway, Neurotrophin signaling pathway, Notch signaling pathway, Oxytocin signaling pathway, PI3K-Akt signaling pathway, PPAR signaling pathway, Phospholipase D signaling pathway, Prolactin signaling pathway, RIG-I-like receptor signaling pathway, Rap1 signaling pathway, Ras signaling pathway, Relaxin signaling pathway, Signaling pathways regulating pluripotency of stem cells, Sphingolipid signaling pathway, T cell receptor signaling pathway, TGF-beta signaling pathway, TNF signaling pathway, Thyroid hormone signaling pathway, Toll-like receptor signaling pathway, VEGF signaling pathway, Wnt signaling pathway, cAMP signaling pathway, cGMP-PKG signaling pathway, mTOR signaling pathway, p53 signaling pathway. After preprocessing the dataset, 17,259 gene-gene interactions for 1,489 genes were obtained.

And 28,843 gene-gene interactions for 2,144 genes were obtained for ROSMAP AD dataset.

A.5 Drug-Target interactions derived from DrugBank

Drug-target information was extracted from the DrugBank[74] database (version 5.1.5, released 2020-01-03). In total, 15,263 drug-target interactions were obtained for 5435 drugs/investigational agents and 2775 targets. Further, 17 drugs with known targets to the 1489 genes previously identified were selected for use in our model for NCI ALMANAC[54], specifically: Celecoxib, Cladribine, Dasatinib, Docetaxel, Everolimus, Fulvestrant, Gefitinib, Lenalidomide, Megestrol acetate, Mitotane, Nilotinib, Paclitaxel, Romidepsin, Sirolimus, Thalidomide, Tretinoin, Vorinostat. And 10 drugs with known targets to the 1489 genes previously identified were selected for use in our model, specifically: Dasatinib, Erlotinib, Lapatinib, Sorafenib, Sunitinib, Vorinostat, Geldanamycin, Metformin, Paclitaxel, Vinblastine.

Section B. Method Details

B.1 K-hop attention-based graph neural network

To incorporate the long-distance information, for each subgraph K-hop attention based mechanism was constructed with

$$\left(\alpha_{ij}^{(m)}\right)_p^{(k)} = \frac{\exp\left(\mathrm{LeakyReLU}(a[W\left(X_p^{(m)}\right)_i || W\left(X_p^{(m)}\right)_j])\right)}{\sum_{q\in\mathcal{N}_i^{(k)}} \exp\left(\mathrm{LeakyReLU}(a[W\left(X_p^{(m)}\right)_i || W\left(X_p^{(m)}\right)_q])\right)} \tag{10}$$

$$(H_p^{(m)})_i = \frac{1}{KQ_i}\sum_{k}^{K}\sum_{q\in\mathcal{N}_i^{(k)}}[\left(\alpha_{iq}^{(m)}\right)_p^{(k)} W\left(X_p^{(m)}\right)_q] \tag{11}$$

, where $\left(\alpha_{ij}^{(m)}\right)_p^{(k)}$ is the attention score mentioned in equation (1) and $\mathcal{N}_i^{(k)}$ is the neighbor nodes calculated by the adjacency matrix in k-th hop $A^{(k)}$ for the node i (See **Appendix B.3** for the algorithm of k-th hop adjacency matrix). The linear transformation vector $a \in \mathbb{R}^{2d'}$ was also

defined. At the same time, the linear transformation for features of each node will be defined as $W \in \mathbb{R}^{d \times d'}$. And $\left(X_p^{(m)}\right)_t$, $\left(H_p^{(m)}\right)_t$ represent the feature and updated feature of node t $(t = 1, 2, \ldots, n)$. Aside from that, Q_t represents the number of signaling pathway the node t $(t = 1, 2, \ldots, n)$ belongs to. And above formula shows the calculation for the 1-head attention. The number of head h will be modified in the model, and the node embeddings for each subgraph will take the average of embeddings in every head attention.

B.2 Weighted Bi-directional Message Propagation

The weight matrices are represented as $W_{up}^{(l)} \in \mathbb{R}^{n \times n}$, $W_{down}^{(l)} \in \mathbb{R}^{n \times n}$ for each of the L global message propagation layers , $(l = 1, 2, \ldots, L)$. The weighted adjacency matrices are defined as follows:

$$A'^{(l)}_{up} = W_{up}^{(l)} \cdot A \ \#(12)$$

$$A'^{(l)}_{down} = W_{down}^{(l)} \cdot A^{\mathrm{T}} \ \#(13)$$

Here, $A'^{(l)}_{up} \in \mathbb{R}^{n \times n}$ and $A'^{(l)}_{down} \in \mathbb{R}^{n \times n}$ maintain the parameters for drug-gene relationships as in the original adjacency matrix. The mean aggregation of each node's neighbors' features from upstream to downstream from layer $l - 1$ to layer l is achieved using the following equation:

$$\left(H_{up_neigh}^{(m)}\right)^{(l)} = D_{in}^{-1}(A'^{(l)}_{up})^{\mathrm{T}} \left(H^{(m)}\right)^{(l-1)} Z_{up}^{(l)} \ \#(14)$$

Similarly, the mean aggregation of each node's neighbors' features from downstream to upstream is computed as:

$$\left(H_{down_neigh}^{(m)}\right)^{(l)} = D_{out}^{-1}(A'^{(l)}_{down})^{\mathrm{T}} \left(H^{(m)}\right)^{(l-1)} Z_{down}^{(l)} \ \#(15)$$

The biased transformation of nodes and their features is given by:

$$\left(H_{self}^{(m)}\right)^{(l)} = \left(H^{(m)}\right)^{(l-1)} B^{(l)} \ \#(16)$$

In these equations, $Z_{up}^{(l)} \in \mathbb{R}^{d^{(l-1)} \times d^{(l)}}$ and $Z_{down}^{(k)} \in \mathbb{R}^{d^{(l-1)} \times d^{(l)}}$ are the linear transformation

matrices for each layer l, $(l = 1, 2, \ldots, L)$. For each layer, the model employs a normalization function N: $\mathbb{R}^{n\times 3d^{(l)}} \rightarrow \mathbb{R}^{n\times 3d^{(l)}}$ and a LeakyReLU activation function with parameter α to map the concatenated node features $\left(H_{concat}^{(m)}\right)^{(l)} = \left(\left(H_{up_neigh}^{(m)}\right)^{(l)} \mid \left(H_{down_neigh}^{(m)}\right)^{(l)} \mid \left(H_{self}^{(m)}\right)^{(l)}\right) \in \mathbb{R}^{n\times 3d^{(l)}}$ to $H^{(l)} \in \mathbb{R}^{n\times 3d^{(l)}}$ as follows

$$(H^{(m)})^{(l)} = \text{LeakyReLU}(N(\left(H_{concat}^{(m)}\right)^{(l)}))\#(17)$$

, where the normalization function N performs L2 normalization on the demo matrix $V \in \mathbb{R}^{p\times q}$ along the row axis using the equation:

$$v'_{ij} = \frac{v_{ij}}{\sqrt{\sum_{j=1}^{q} v_{ij}^2}} \#(18)$$

, where v'_{ij} is an element of the new matrix $V' \in \mathbb{R}^{p\times q}$.

B.3 K-hop adjacency matrix generation algorithm

Algorithm: $\boldsymbol{k}$-th hop adjacency matrix

Input: initial adjacency matrix $A \in \mathbb{R}^{n\times n}$
Step 1: $A_{sum} = A$
Step 2: For k in $2, 3, \ldots, K$:
Step 3: $A^{(k)} = A^k$
Step 4: For i in $1, 2, 3, \ldots, n$:
Step 5: For j in $1, 2, 3, \ldots, n$:
Step 6: If $A_{ij}^{(k)} > 1$:
Step 7: $A_{ij}^{(k)} = 1$
Step 8: End If
Step 9: If $i = j$:
Step 10: $A_{ij}^{(k)} = A_{ij}^{(k)} - (A_{sum})_{ij} - 1$
Step 11: Else:
Step 12: $A_{ij}^{(k)} = A_{ij}^{(k)} - (A_{sum})_{ij}$
Step 13: End if
Step 14: If $A_{ij}^{(k)} < 0$:
Step 15: $A_{ij}^{(k)} = 0$
Step 16: End If
Step 17: End For
Step 18: End For
Step 19: $A_{sum} = A_{sum} + A^{(k)}$
Step 20: End For

Output: $A^{(2)}, A^{(3)}, \ldots, A^{(K)}$

Section C Code Availability

M3NetFlow code: https://github.com/FuhaiLiAiLab/M3NetFlow

NetFlowVis app: https://m3netflow.shinyapps.io/NetFlowVis/

Acknowledgment

This study was partially supported by NIA R56AG065352 (to Li), NIA 1R21AG078799-01A1 (to Li/Province), NINDS 1RM1NS132962-01 (to Dickson/Marco/Cooper/Li), Children's Discovery Institute (CDI) M-II-2019-802, and NLM 1R01LM013902-01A1 to Li.

Author contributions

FL conceived this study. FL, HZ designed the model and prepared the manuscript. HZ implemented the model and analyzed the data and results. FL, HZ, PG, LD, WK, DD, RF, YC, PP discussed the method and results.

Declaration of interests

The authors declare no competing interests.

References

1. TCGA. *https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga*.
2. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. doi:10.1038/nature11003
3. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013;41(D1):955-961. doi:10.1093/nar/gks1111
4. Li F, Eteleeb A, Buchser W, et al. Weakly activated core inflammation pathways were identified as a central signaling mechanism contributing to the chronic neurodegeneration in Alzheimer's disease. *bioRxiv*. Published online 2021.

5. Li F, Oh I, Kumar S, et al. Loss of estrogen unleashing neuro-inflammation increases the risk of Alzheimer's disease in women. *bioRxiv*. Published online 2022:2022-2029.
6. TCGA. *https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga*.
7. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019;569(7757):503-508. doi:10.1038/s41586-019-1186-3
8. Gazzoli I, Loda M, Garber J, Syngal S, Kolodner RD. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. *Cancer Res*. 2002;62(14):3925-3928.
9. Baylin SB, Jones PA. Epigenetic determinants of cancer. *Cold Spring Harb Perspect Biol*. 2016;8(9):a019505.
10. Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*. 2016;166(3):755-765.
11. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031
12. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights*. 2020;14:1177932219899051.
13. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237-i245.
14. Mihajlovi´cmihajlovi´c K, Ceddia G, Malod-Dognin N, et al. Multi-omics integration of scRNA-seq time series data predicts new intervention points for Parkinson's disease. Published online 2023. doi:10.1101/2023.12.12.570554
15. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods*. 2020;17(2):159-162.
16. Hu Y, Peng T, Gao L, Tan K. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv*. 2021;7(16):eabf1356.
17. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol*. 2022;40(10):1458-1466. doi:10.1038/s41587-022-01284-4
18. Bodein A, Scott-Boyer MP, Perin O, Lê Cao KA, Droit A. TimeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics*. 2022;38(2):577-579. doi:10.1093/bioinformatics/btab664
19. Zhou G, Li S, Xia J. Network-based approaches for multi-omics integration. *Computational methods and data analysis for metabolomics*. Published online 2020:469-487.
20. Zhang T, Zhang L, Payne P, Li F. Synergistic Drug Combination Prediction by Integrating Multi-omics Data in Deep Learning Models. *arXiv preprint arXiv:181107054*. Published online 2018.
21. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907*. Published online 2016.
22. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst*. 2017;2017-Decem(Nips):1025-1035.

23. Veličković P, Casanova A, Liò P, Cucurull G, Romero A, Bengio Y. Graph attention networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. Published online 2018:1-12.
24. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? *arXiv preprint arXiv:181000826*. Published online 2018.
25. 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(1):3445.
26. Li X, Ma J, Leng L, et al. MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front Genet*. 2022;13:806842.
27. Gao H, Zhang B, Liu L, Li S, Gao X, Yu B. A universal framework for single-cell multi-omics data integration with graph convolutional networks. *Brief Bioinform*. 2023;24(3):bbad081.
28. Rajadhyaksha N, Chitkara A. Graph Contrastive Learning for Multi-omics Data. *arXiv preprint arXiv:230102242*. Published online 2023.
29. Li G, Muller M, Thabet A, Ghanem B. DeepGCNs: Can GCNs go as deep as CNNs? *Proceedings of the IEEE International Conference on Computer Vision*. 2019;2019-Octob:9266-9275. doi:10.1109/ICCV.2019.00936
30. Cai C, Wang Y. A Note on Over-Smoothing for Graph Neural Networks. Published online 2020.
31. Morris C, Ritzert M, Fey M, et al. *Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks*. www.aaai.org
32. Weisfeiler YuB, Leman A. The reduction of a graph to canonical form and the algebra which appears therein. *Nauchno-Technicheskaya Informatsia (NTI Series)*. 1968;2(9):12-16.
33. Chien E, Peng J, Li P, Milenkovic O. Adaptive universal generalized pagerank graph neural network. *arXiv preprint arXiv:200607988*. Published online 2020.
34. Nikolentzos G, Dasoulas G, Vazirgiannis M. k-hop graph neural networks. *Neural Networks*. 2020;130:195-205.
35. Abu-El-Haija S, Perozzi B, Kapoor A, et al. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In: *International Conference on Machine Learning*. PMLR; 2019:21-29.
36. Brossard R, Frigo O, Dehaene D. Graph convolutions that can finally model local structure. *arXiv preprint arXiv:201115069*. Published online 2020.
37. Wang G, Ying R, Huang J, Leskovec J. Multi-hop attention graph neural network. *arXiv preprint arXiv:200914332*. Published online 2020.
38. Feng J, Chen Y, Li F, Sarkar A, Zhang M. *How Powerful Are K-Hop Message Passing Graph Neural Networks*.
39. Kanehisa MGS, Goto S. KEGG: kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28:27-30. doi:10.1093/nar/28.1.27
40. Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47(D1):D529-D541.
41. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34(suppl_1):D535-D539.
42. Salton G. Modern information retrieval. *(No Title)*. Published online 1983.

43. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: Predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*. 2018;34(9):1538-1546. doi:10.1093/bioinformatics/btx806
44. Zhang T, Zhang L, Payne P, Li F. Synergistic Drug Combination Prediction by Integrating Multi-omics Data in Deep Learning Models. *arXiv preprint arXiv:181107054*. Published online 2018.
45. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. Published online 2017:1-14.
46. Veličković P, Casanova A, Liò P, Cucurull G, Romero A, Bengio Y. Graph attention networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. Published online 2018:1-12.
47. Shi Y, Huang Z, Feng S, Zhong H, Wang W, Sun Y. *Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification*.
48. Corso G, Cavalleri L, Beaini D, Liò P, Veličkovi´ P, Deepmind V. *Principal Neighbourhood Aggregation for Graph Nets*.
49. Thekumparampil KK, Wang C, Oh S, Li LJ. *Attention-Based Graph Neural Network for Semi-Supervised Learning*.; 2018.
50. Holbeck SL, Camalier R, Crowell JA, et al. The National Cancer Institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res*. 2017;77(13):3564-3576. doi:10.1158/0008-5472.CAN-17-0489
51. Rosene MJ, Hsu S, You SF, et al. Phospholipase D3 contributes to Alzheimer's disease risk via disruption of Aβ clearance and microglia response to amyloid plaques. *medRxiv*. Published online January 1, 2022:2022.01.31.22270175. doi:10.1101/2022.01.31.22270175
52. Wang X, Li H, Sheng Y, et al. The function of sphingolipids in different pathogenesis of Alzheimer's disease: A comprehensive review. *Biomedicine & Pharmacotherapy*. 2024;171:116071. doi:https://doi.org/10.1016/j.biopha.2023.116071
53. Davoody S, Asgari Taei A, Khodabakhsh P, Dargahi L. mTOR signaling and Alzheimer's disease: What we know and where we are? *CNS Neurosci Ther*. 2024;30(4):e14463. doi:https://doi.org/10.1111/cns.14463
54. Razani E, Pourbagheri-Sigaroodi A, Safaroghli-Azar A, Zoghi A, Shanaki-Bavarsad M, Bashash D. The PI3K/Akt signaling axis in Alzheimer's disease: a valuable target to stimulate or suppress? *Cell Stress Chaperones*. 2021;26(6):871-887. doi:https://doi.org/10.1007/s12192-021-01231-3
55. Adinolfi LE, Nevola R, Lus G, et al. Chronic hepatitis C virus infection and neurological and psychiatric disorders: An overview. *World J Gastroenterol*. 2015;21(8):2269-2280. doi:10.3748/wjg.v21.i8.2269
56. Gate D, Saligrama N, Leventhal O, et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature*. 2020;577(7790):399-404. doi:10.1038/s41586-019-1895-7
57. Merlini M, Kirabali T, Kulic L, Nitsch RM, Ferretti MT. Extravascular CD3+ T Cells in Brains of Alzheimer Disease Patients Correlate with Tau but Not with Amyloid Pathology: An

Immunohistochemical Study. *Neurodegener Dis*. 2018;18(1):49-56. doi:10.1159/000486200
58. Theoharides TC, Zhang B. Neuro-inflammation, blood-brain barrier, seizures and autism. *J Neuroinflammation*. 2011;8. doi:10.1186/1742-2094-8-168
59. Li S, Hong S, Shepardson NE, Walsh DM, Shankar GM, Selkoe D. Soluble Oligomers of Amyloid β Protein Facilitate Hippocampal Long-Term Depression by Disrupting Neuronal Glutamate Uptake. *Neuron*. 2009;62(6):788-801. doi:10.1016/j.neuron.2009.05.012
60. Chen X, Lin R, Chang L, et al. Enhancement of long-term depression by soluble amyloid β protein in rat hippocampus is mediated by metabotropic glutamate receptor and involves activation of p38MAPK, STEP and caspase-3. *Neuroscience*. 2013;253:435-443. doi:10.1016/j.neuroscience.2013.08.054
61. Craft S. *Insulin Resistance and Alzheimer's Disease Pathogenesis: Potential Mechanisms and Implications for Treatment*. Vol 4.; 2007.
62. Hernandez F, Lucas JJ, Avila J. GSK3 and tau: Two convergence points in Alzheimer's disease. *Journal of Alzheimer's Disease*. 2013;33(SUPPL. 1). doi:10.3233/JAD-2012-129025
63. Zhu X, Rottkamp CA, Hartzler A, et al. Activation of MKK6, an upstream activator of p38, in Alzheimer's disease. *J Neurochem*. 2001;79(2):311-318. doi:10.1046/j.1471-4159.2001.00597.x
64. Guo N, Shen G, Zhang Y, Moustafa AA, Ge D, You Z. Interleukin-17 promotes migration and invasion of human cancer cells through upregulation of MTA1 expression. *Front Oncol*. 2019;9(JUN):1-11. doi:10.3389/fonc.2019.00546
65. O'Neil J, Benita Y, Feldman I, et al. An unbiased oncology compound screen to identify novel combination strategies. *Mol Cancer Ther*. 2016;15(6):1155-1162. doi:10.1158/1535-7163.MCT-15-0843
66. Zagidullin B, Aldahdooh J, Zheng S, et al. DrugComb: An integrative cancer drug combination data portal. *Nucleic Acids Res*. 2019;47(W1):W43-W51. doi:10.1093/nar/gkz337
67. Van Der Meer D, Barthorpe S, Yang W, et al. Cell Model Passports - a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res*. 2019;47(D1):D923-D929. doi:10.1093/nar/gky872
68. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. doi:10.1038/nature11003
69. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. Published online 1999:28. doi:10.1093/nar/27.1.29
70. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074-D1082. doi:10.1093/nar/gkx1037

Supplementary Figure 1

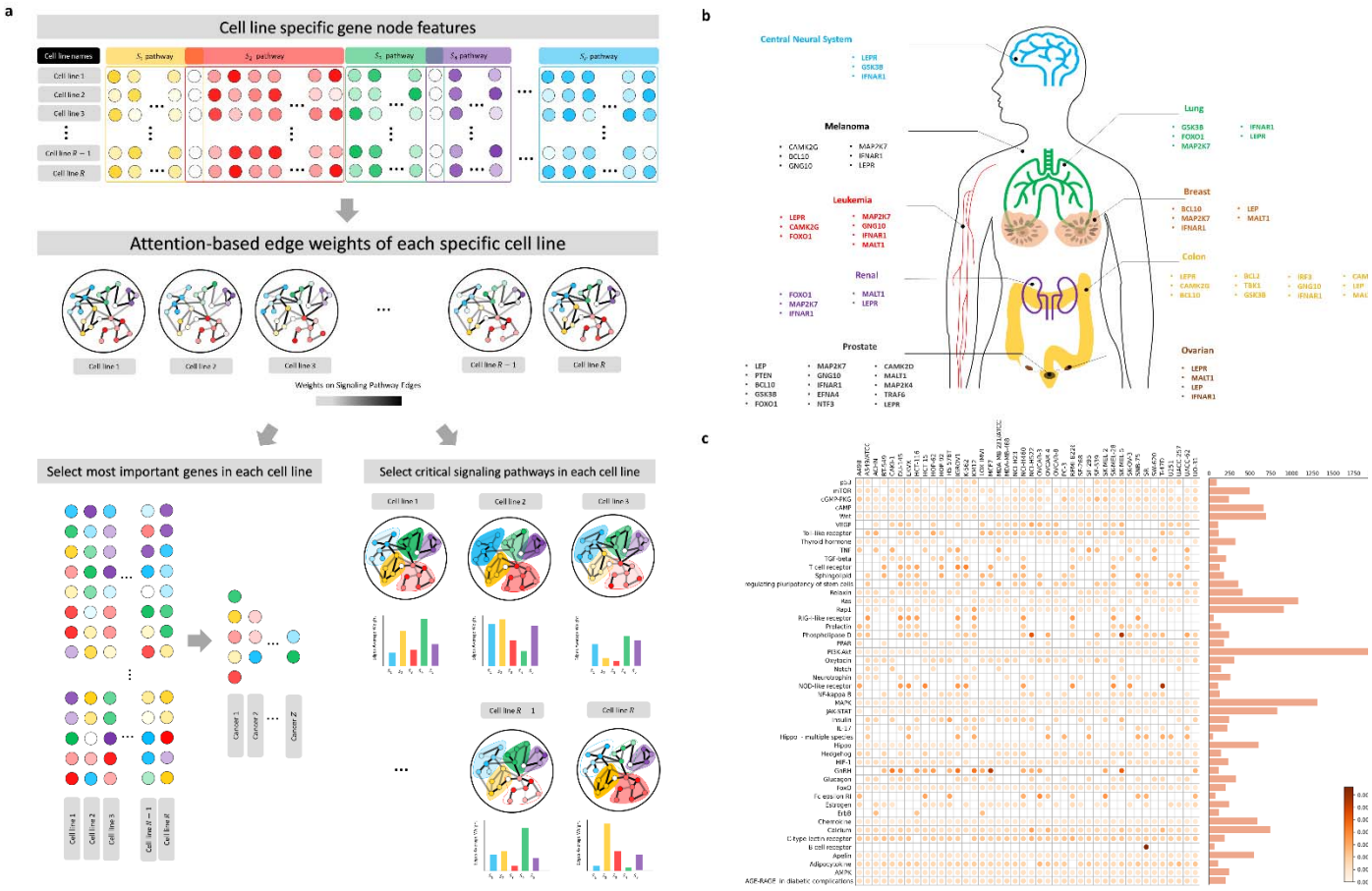

Figure S1. Downstream analysis of important genes and signaling pathways. **a.** Procedures of identifying the important genes in different cell lines and cancers, and identifying the important signaling pathways in different cell lines. **b.** Overlapped genes for the cell lines in each type of cancer. **c.** Heatmap of sum of edge weights in different signaling pathways in whole NCI ALMANAC dataset.