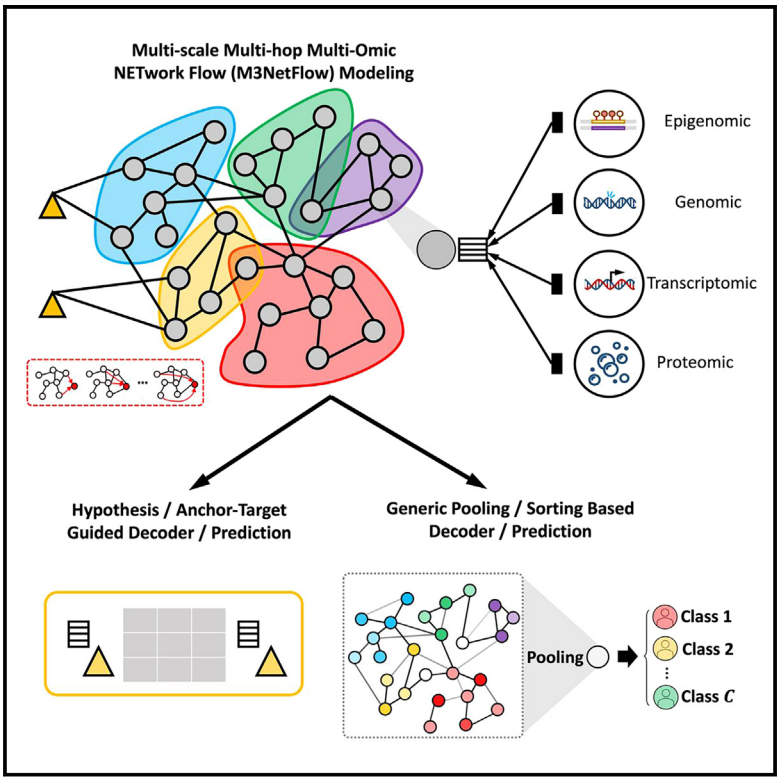
iScience

Article

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

Graphical abstract



Highlights

- M3NetFlow is a graph AI model for integrative and interpretable multi-omic analysis
- M3NetFlow supports target and pathway inference based on given targets of interest
- M3NetFlow supports generic target and pathway inference from multi-omic data
- NetFlowVis can visualize multi-omic features of predicted targets and pathways

Authors

Heming Zhang, S. Peter Goedegebuure, Li Ding, ..., Yixin Chen, Philip Payne, Fuhai Li

Correspondence

fuhai.li@wustl.edu

In brief

Biocomputational method; Complex systems; Omics

Zhang et al., 2025, iScience *28*, 111920
March 21, 2025
https://doi.org/10.1016/j.isci.2025.111920

CellPress

iScience

CellPress
OPEN ACCESS

Article

M3NetFlow: A 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] Michael Province,[5] Yixin Chen,[6] Philip Payne,[1] and Fuhai Li[1,5,7,8,*]

[1]Institute for Informatics, Data Science and Biostatistics (I2DB), Washington University in St. Louis, St. Louis, MO, USA
[2]Department of Surgery, Washington University in St. Louis, St. Louis, MO, USA
[3]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO, USA
[4]Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA
[5]Division of Statistical Genomics, Department of Genetics, Washington University in St. Louis, St. Louis, MO, USA
[6]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA
[7]Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA
[8]Lead contact
*Correspondence: fuhai.li@wustl.edu
https://doi.org/10.1016/j.isci.2025.111920

SUMMARY

Multi-omic data-driven studies are at the forefront of precision medicine by characterizing complex disease signaling systems across multiple views and levels. The integration and interpretation of multi-omic data are critical for identifying disease targets and deciphering disease signaling pathways. However, it remains an open problem due to the complex signaling interactions among many proteins. Herein, we propose a multi-scale multi-hop multi-omic network flow model, M3NetFlow, to facilitate both hypothesis-guided and generic multi-omic data analysis tasks. We evaluated M3NetFlow using two independent case studies: (1) uncovering mechanisms of synergy of drug combinations (hypothesis/anchor-target guided multi-omic analysis) and (2) identifying biomarkers of Alzheimer's disease (generic multi-omic analysis). The evaluation and comparison results showed that M3NetFlow achieved the best prediction accuracy and identified a set of drug combination synergy- and disease-associated targets. The model can be directly applied to other multi-omic data-driven studies.

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 generated to characterize dysfunctional biological processes and signaling pathways from multiple levels/views and to elucidate the panoramic view of the disease pathogenesis.[1–3] For example, The Cancer Genome Atlas (TCGA) program has generated multi-omic datasets of over 20,000 samples spanning 33 cancer types, to understand the key molecular targets and signaling pathways of cancer. 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.[4] In addition, the multi-omic data of Alzheimer's disease (AD) are generated and publicly available in The Religious Orders Study and Memory and Aging Project (ROSMAP[5]) project to uncover the pathogenesis of AD. Also, the exceptional longevity (EL), like the Long-Life Family Study project, has been generating multi-omic data[6,7] 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 effective drugs and drug cocktails to guide the development of precision medicine. However, it remains an open problem and a challenging task to integrate multi-omic data and mine core disease signaling pathways from the complex and intensive signaling interactions among a large number of proteins in the cell signaling system.[8]

The two problems to be tackled in this study are (1) hypothesis/anchor-target guided multi-omic data analysis and (2) generic multi-omic data analysis. The hypothesis/anchor-targets can be known as disease-associated targets, or drug targets; and the expected outcome is the upstream or downstream signaling pathways of given anchor-targets. Specifically, the drug combination synergy score prediction task (as the hypothesis/anchor-target guided analysis example) is defined as follows. In two experimental datasets (National Cancer Institute [NCI] 60 and the O'Neil), the synergy scores (label information to train the model) of a set of pairwise drug combinations were experimentally measured on a set of cancer cell lines. The multi-omic data of these cancer cell lines were also measured and publicly accessible. The biological and computational problems are if a computational model can (1) predict synergy score of drug combinations and (2) investigate the potential mechanism of synergy (MoS) using (model inputs) the

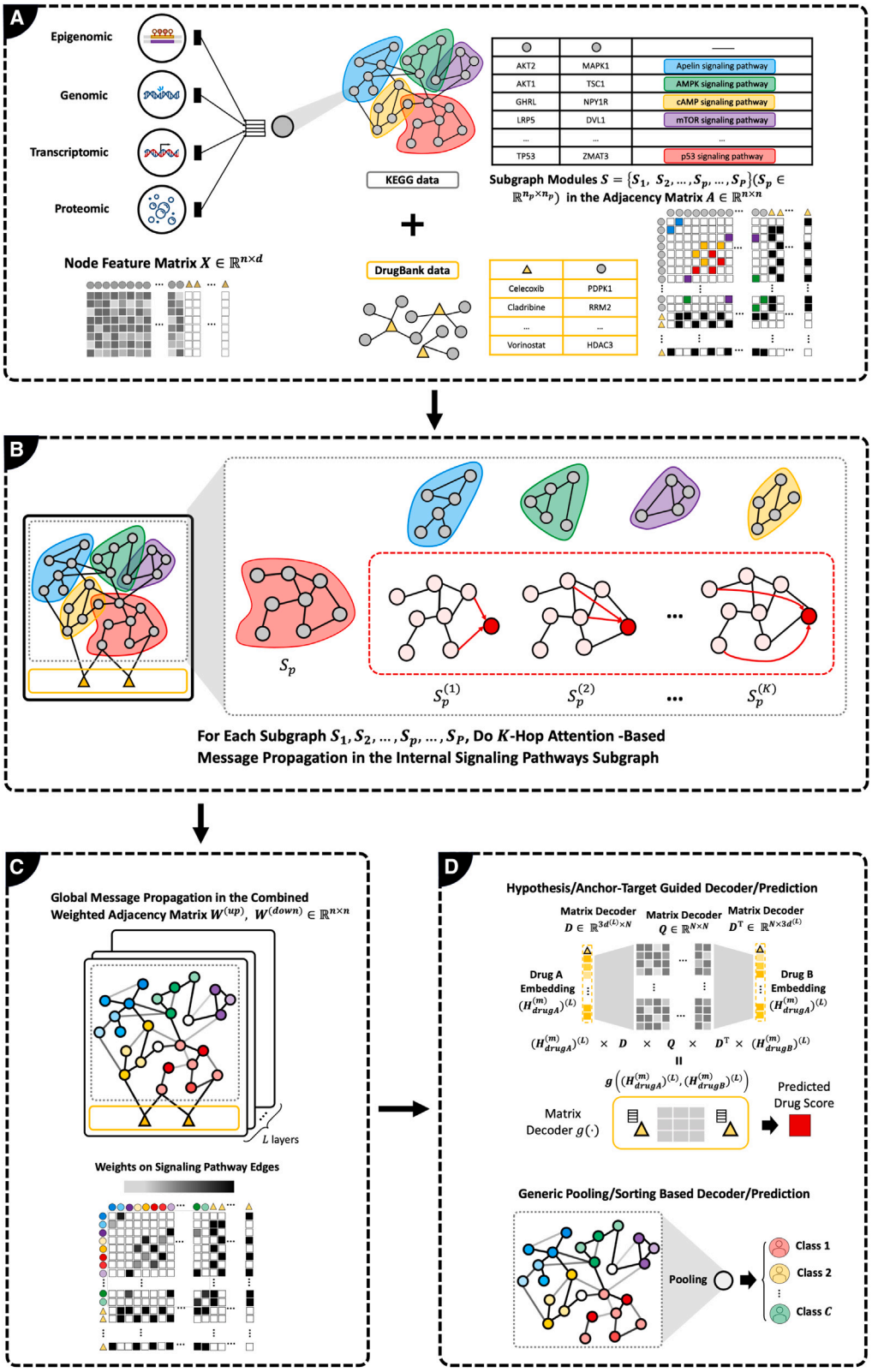
multi-omic data of cancer cell lines, known drug targets (drug-target interactions), and Kyoto Encyclopedia of Genes and Genomes (KEGG[9]) signaling pathways (a graph with signaling/protein-protein interactions). So each data point for the model can be represented as <DrugA (drug targets), DrugB (drug targets), Cell_Line (multi-omic data), SynergyScore (label to train the model)> (in addition to KEGG signaling pathways). The expected model outcomes/output are (1) synergy score prediction capability of the model and (2) MoS of effective drug combination on given cell lines, i.e., important signaling targets and cascades/flows within cell lines that can explain which drug combinations are effective and which are not. For the generic multi-omic data analysis demonstrated using AD sample classification task, the multi-omic datasets of AD and control human samples are publicly accessible. The biological and computational problems are if a computational model can (1) predict sample categories and (2) identify disease-associated targets and pathways. Specifically, the model inputs are the (1) multi-omic data of samples, (2) KEGG signaling graph, and (3) sample categories (label information to train the model): AD vs. control. So each data point for the model can be represented as <Sample (multi-omic data), Category (AD vs. control, label to train the model)> (in addition to KEGG signaling pathways). The expected model outputs are (1) model prediction capability to classify samples into AD vs. control and (2) the potential AD-associated targets.

A comprehensive review of existing multi-omic data integration analysis models was reported.[8] Specifically, these models were clustered into a few categories, like similarity, correlation, Bayesian, multivariate, fusion, and network-based models.[8] The pathway representation and analysis by direct inference on graphical models (PARADIGM[10]) is one of the most widely used methods among these traditional computational methods. The mixOmics[11] and timeOmics[12] models and tools were developed to identify disease-associated genes using multivariate analysis on the multi-omic datasets of one and multiple time points, respectively, in which the signaling network information was not integrated. The mechanism of action generator involving network analysis (MAGINE)[13] is an enrichment-based framework, in which the differentially expressed genes from multi-omic data are identified first, and enriched pathways and network modules are then identified based on the identified genes. The GLUE[14] (graph-linked unified embedding) model was proposed to integrate multi-omic data by mapping raw omic data/features into a new embedding space for the cell clustering and correlation analysis. In the GLUE model, the signaling network information was converted to embedding features to adjust the embedding of raw omic data, which did not directly use the network information to calculate the signaling flow on the network. Recently, graph neural networks (GNNs) have gained prominence due to their capability to model relationships within graph-structured data.[15–17] And numerous studies have applied the GNN with the integration of the multi-omic data. Multi-omics graph convolutional networks (MOGONET[18]) initially creates similarity graphs among samples by leveraging each omic data and then employs a graph convolutional network (GCN[17]) to learn a label distribution from each omic data independently. Subsequently, a cross-omic discovery tensor is implemented to refine the prediction by learning the dependency among multi-omic data. Multi-omics integration model based on graph convolutional network (MoGCN[19]) adopts a similar approach by constructing a patient similarity network using multi-omic data and then using GCN to predict the cancer subtype of patients. The universal framework for the integration of single-cell multi-omics data based on graph convolutional network (GCN-SC[20]) utilizes a GCN to combine single-cell multi-omic data derived from varying sequencing methodologies. Nevertheless, none of these models 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 hamper their ability to incorporate many layers. The over-smoothing problem was firstly mentioned by extending the propagation layers in GCN.[21] Moreover, theoretical papers using Dirichlet energy showed diminished discriminative power by increasing the propagation layers. And multiple attempts were made to compare the expressive power of the GCNs,[22] and it is shown that Weisfeiler-Lehman (WL) subtree kernel[23] 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.[24–28] 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, M3NetFlow (multi-scale, multi-hop, multi-omic network flow) to address the challenges mentioned earlier. The major and unique contributions are as follows. Compared with existing models, we first mapped multi-omic data and drug-target information onto KEGG signaling pathways,[9] which can support both anchor-target (drug targets) guided analysis and generic multi-omic analysis. To improve the model expression power, then we conducted the message propagation or signaling flowing with attention, multi-hop, on both local signaling modules and then updated the node embedding on the global signaling graph. This model design and model architecture are novel for multi-omic analysis tasks. Then the important signaling targets and interactions can be identified using the attention-based scores, learned in the model on the training dataset. To assess and demonstrate the effectiveness of the proposed model, M3NetFlow, it was applied in two independent multi-omic case studies: (1) uncovering mechanisms of synergy of effective drug combinations (hypothesis/anchor-target guided multi-omic analysis) and (2) identifying biomarkers of AD (generic multi-omic analysis). The evaluation and comparison results showed that M3NetFlow achieved the best prediction accuracy and identified a set of essential drug combination synergy- and disease-associated targets. To facilitate investigating the analysis results, a visualization tool, NetFlowVis, was developed to visualize the top-ranked signaling targets and interactions based on the attention-based target and interaction scores. The details of the studies are introduced in the following sections.

RESULTS

Figure 1 shows the schematic architecture of the proposed M3NetFlow model. The model input parameters are $\mathcal{X} = \{(X^{(1)}, T^{(1)}), (X^{(2)}, T^{(2)}), \ldots, (X^{(m)}, T^{(m)}), \ldots, (X^{(M)}, T^{(M)})\}$ $(X^{(m)} \in \mathbb{R}^{n\times d}, T \in \mathbb{R}^{n\times 2})$, $A \in \mathbb{R}^{n\times n}$, $S = \{S_1, S_2, \ldots, S_p, \ldots, S_P\}(S_p \in \mathbb{R}^{n_p\times n_p})$, $D_{in} \in \mathbb{R}^{n\times n}, D_{out} \in \mathbb{R}^{n\times n}$, where M represents the number of data



(legend on next page)

points, $\mathcal{X}$ 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 two drugs (drug combinations) targeted on those n nodes (in the general multi-omic data analysis, the variable T will not be used). The matrix A is the adjacency matrix that demonstrates the node-node interactions. The element in adjacency matrix A such as a_{ij} indicates an edge from i to j. 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 node 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. The model is to build up a model $f(\cdot)$ to predict the labels of samples: $f(\mathcal{X}, A, S, D_{in}, D_{out}) = Y$, where Y is the sample labels, $Y \in \mathbb{R}^{M \times 1}$ (e.g., the drug combination synergy scores or AD vs. control).

Experimental setup

For drug combination synergy score prediction task, the 5-fold cross-validation ($\mathcal{F} = 5$) was employed. For each sample, it has four elements: <D_A, D_B, C_C, S_{ABC}> representing that drugs D_A and D_B are used on cell line C_C and S_{ABC} is the drug synergy score. The drug targets of D_A and D_B and the multi-omic data of cell line C_C were used as the model input to predict S_{ABC} (see STAR Methods). Specifically, there are 2,788 samples for the NCI ALMANAC[29] (A Large Matrix of Anti-Neoplastic Agent Combinations) drug combination dataset and 1,008 samples for the O'Neil dataset.[30] For this task, 1,489 proteins on KEGG signaling graph were selected (overlapping with the omic data), and each protein (graph node) is characterized by 6 omic features: gene expression (RNA sequencing), copy-number variation (CNV), gene amplification, gene deletion, and gene methylation maximum and minimum values, as well as it is a target of D_A or D_B from Cell Model Passports,[31] CCLE, and DrugBank[32] databases. For AD sample classification task, multi-omic data of 138 (74 AD, 64 control [non-AD]) samples were derived from the ROSMAP[5] database. We randomly selected 64 AD and 64 control samples as a balanced dataset and used the 5-fold cross-validation ($\mathcal{F} = 5$) to evaluate the model performance. For this task, 2,099 proteins on KEGG signaling graph were selected (overlapping with the omic data), and each protein (graph node) is characterized by 10 omic features: methylation values on upstream, distal promoters, proximal promoters, core promoters and downstream, genetic duplication, deletions, CNV, and gene expression and protein expression from ROSMAP and GEO[33] platform.

Hyperparameters

The models were implemented using pytorch and torch geometric. For both tasks, the learning rate started at 0.002 and was reduced equally within each batch for a certain epoch stage. After 60 epochs, the learning rate was set at 0.0001. Adam optimizer was used for optimization with eps = 1e−7 and weight_decay = 1e−20. We empirically set the K-hop subgraph message propagation with $K = 3$ (3 hops) and the global bi-directional message propagation with $L = 3$ (3 *layers*). Afterward, the feature dimensions will vary at the different layers and will be denoted by $(d^{(1)}, d^{(2)}, \ldots, d^{(l)}, \ldots, d^{(L)})$. At the global message propagation step, the layer will concatenate new node embedding features and output 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's (input dims, output dims), $(d^{(0)}, 3d^{(1)})$; (2) second layer's (input dims, output dims), $(3d^{(1)}, 3d^{(2)})$; and (3) third layer's (input dims, output dims), $(3d^{(2)}, 3d^{(3)})$. The final embedded node dims were $3d^{(3)}$ ($L = 3$). For drug combination synergy score 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; $E = 150$ was used. In AD sample classification task, the max pooling was employed to predict the sample outcome as described in Equation 6. As for the LeakyReLU function, the parameter was set as 0.1 for both tasks.

M3NetFlow improves 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 1, 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 prediction accuracy was about 66% using the test data in the ROSMAP dataset. These prediction results are comparable with existing deep learning models.[34,35] Moreover, we also compared our proposed model M3NetFlow with other deep learning models, which included the GCN,[17] graph attention network[16] (GAT), UniMP,[36] MixHop,[25] principal neighborhood aggregation[37] (PNA), and graph isomorphism network (GIN). By checking the p values over 5-fold cross-validation, the performances of the M3NetFlow have significant improvement over most of the GNN-based methods (see Table 1 and Figure 2A).

M3NetFlow ranks important targets and interactions via attention scores

Targets with higher importance scores predicting drug combination synergy

Based on the node importance score calculated from edge attention scores (see target/node importance score calculation

Figure 1. Model architecture of M3NetFlow

(A) Mapping multi-omic data and drug targets onto KEGG signaling pathways.

(B) Multi-hop attention-based signaling propagation on subgraphs.

(C) Global signaling propagation.

(D) Downstream tasks: (D.1) hypothesis/anchor-target guided decoder/prediction; (D.2) generic pooling/sorting based decoder/prediction.

Table 1. Model comparisons using average Pearson correlation and prediction accuracy (mean ± standard deviation) of 5-fold cross-validation using NCI ALMANAC, O'Neil, and ROSMAP datasets

Dataset	NCI ALMANAC	O'Neil	ROSMAP
GCN	51.93% ± 3.55%	44.47% ± 6.60%	59.43% ± 4.53%
GAT	49.16% ± 2.26%	57.06% ± 3.72%	62.80% ± 7.11%
UniMP	49.02% ± 4.62%	55.84% ± 10.93%	61.83% ± 3.78%
MixHop	57.78% ± 3.66%	27.15% ± 10.01%	57.20% ± 3.92%
PNA	55.63% ± 2.47%	62.20% ± 2.27%	57.83% ± 1.82%
GIN	53.76% ± 2.47%	33.12% ± 9.89%	49.83% ± 5.71%
M3NetFlow	60.72% ± 0.77%	64.36% ± 2.53%	67.34% ± 5.12%

in STAR Methods section for details), the important targets for each cell line can be selected. For example, to investigate the MoS, we compared the importance scores of drug targets of the top 5 drug combinations (with highest synergy scores) and bottom 5 drug combinations (lowest drug synergy scores) (Figure 3A). In another word, it is expected that the targets of synergistic drug combinations have higher importance scores than the targets of the non-synergistic drug combinations. Our results confirmed this pattern in 37 out of 41 (∼90%) cell lines, which have higher mean target importance scores (light green in Table S1). Specifically, in 27 out of 41 (∼65%) cell lines, the targets of synergistic drug combinations have higher importance scores with p value $\leq$ 0.1 (deep orange color in Table S1), and, in 32 out of 41 (∼78%) cell lines, the targets of synergistic drug combinations have higher importance scores with p value $\leq$ 0.3 (light orange color in Table S1). Figure 4 shows the pattern of target/node importance scores of drug combinations. Specifically, the left boxplots compared the node/target importance score distribution of individual top 5 and bottom (low) 5 drug combinations, respectively (each drug combination has multiple targets). The right boxplots show the node/target importance score distribution of all top 5 and bottom 5 drug combinations, respectively. The scores of individual drug targets and proteins in each cell line were provided in Table S2, and the top 20 protein lists for each cell line were provided in Table S3. Moreover, we identified the commonly important proteins across cell lines of the same cancer type (Figure S2).

Top-ranked targets are associated with AD

By setting the filters (edge threshold as 0.106 and a small component threshold as 15), we identified 100 potential important genes for AD. Among those genes, 28 genes are filtered out by setting the threshold of attention-based node weight as 2.0 and 15 of them with p values smaller than 0.1 in at least one of the 10 multi-omic features (see Figures 5A and 5B). To evaluate the top targets ranked by attention-based node weight, the top-ranked 28 AD-associated biomarkers were further analyzed via pathway enrichment analysis. Interestingly, a set of AD-associated signaling pathways are identified. As shown in Figure 5C, the top-ranked targets are involved in a set of signaling transduction pathways, which indicates the importance of these targets. Among them, several signaling pathways play critical roles in AD particularly through mechanisms involving inflammation, immune responses, and cellular growth and death. For instances, the B cell receptor (BCR) and T cell receptor (TCR) signaling pathways are vital for B cell and T cell activation, which plays a crucial role in immune surveillance and inflammation. Shared molecular components between the BCR and TCR pathways, including Mitogen-Activated Protein Kinase Kinase 1 (MAP2K1), Jun Proto-Oncogene, AP-1 Transcription Factor Subunit (JUN), Component Of Inhibitor Of Nuclear Factor Kappa B Kinase Complex (CHUK), RELA Proto-Oncogene, NF-KB Subunit (RELA), Nuclear Factor Kappa B Subunit 1 (NFKB1), Inhibitor Of Nuclear Factor Kappa B Kinase Subunit Beta (IKBKB), and protein kinase B (AKT) genes, suggest significant crosstalk that contributes to neuroinflammatory processes in AD. Prolonged activation of these immune pathways may exacerbate chronic inflammation and contribute to AD pathology by sustaining harmful neuroinflammatory responses.[38] The nuclear factor κBNF-κB signaling axis, a critical component in both BCR and TCR pathways, is particularly relevant in AD due to its role in regulating inflammatory cytokine production. Chronic activation of NF-κB has been linked to increased amyloid-beta (Aβ) deposition and neurofibrillary tangle formation, both of which are hallmark features of AD pathology.[39,40] Additionally, activation of NF-κB, mitogen-activated protein kinase (MAPK), and AKT signaling cascades can induce neuronal apoptosis, leading to cognitive decline associated with AD. The MAPK, rat sarcoma (RAS), Phosphatidylinositol 3-Kinase / Protein Kinase B (PI3K/Akt), and forkhead box O (FoxO) signaling pathways also play critical roles in modulating immune responses and neuronal survival. The p38 MAPK pathway, for example, drives neuroinflammation by activating microglia and promoting the release of pro-inflammatory cytokines, which can result in neuronal death.[41] Overactivation of RAS signaling increases oxidative stress, contributing to Aβ accumulation and tau pathology, both of which are central to neurodegeneration in AD.[42] While Aβ accumulation is an early event in AD, tau pathology is more closely associated with cognitive decline, and, together, these processes are considered the primary drivers of neuronal death in AD.[43,44] The Akt signaling pathway, which promotes cell survival by inhibiting apoptosis through downstream effectors such as B-cell CLL/lymphoma 2 (BCL2), is also impaired in AD. Reduced Akt activity has been linked to increased neuronal apoptosis and the accumulation of Aβ and hyperphosphorylated tau. Notably, the PI3K-Akt pathway is intimately connected to insulin signaling, which is disrupted in AD and is characterized by reduced Akt signaling, impaired glucose metabolism, and an increased vulnerability to neurodegeneration.[45] Furthermore, neurotrophin signaling, which regulates axonal growth and regeneration via MAPK and PI3K-Akt pathways, is another pathway that becomes disrupted in AD. This disruption impairs axonal repair mechanisms and contributes to synaptic loss and neuronal degeneration.[46] Additionally, endocrine-related pathways, including insulin and relaxin signaling, play crucial roles in maintaining cellular homeostasis and survival. Dysregulation of these pathways in AD contributes

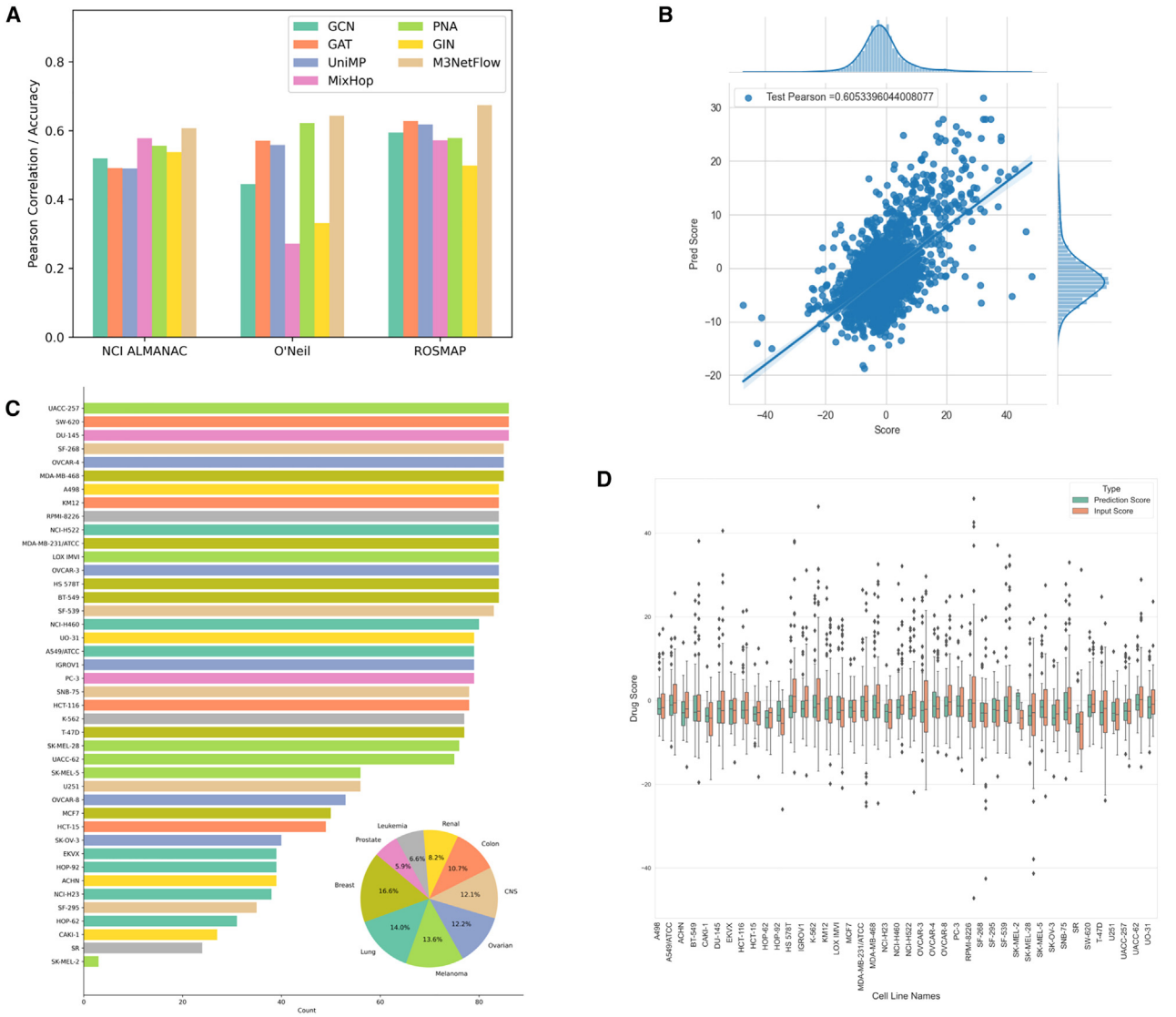


Figure 2. Model performance and overview of input datasets NCI ALMANAC, O'Neil (drug combination multi-omic data), and ROSMAP (AD multi-omic data)

(A) Averaged Pearson correlation across 5-fold validation comparisons for GCN, GAT, UniMP, MixHop, PNA, GIN, and M3NetFlow models for NCI ALMANAC, O'Neil, and ROSMAP dataset (data are represented as mean).

(B) Scatterplot of the model with data points in the whole NCI ALMANAC dataset.

(C) Distributions of all cell lines in the whole NCI ALMANAC dataset.

(D) Boxplots across all cell lines in the whole NCI ALMANAC dataset.

to neuronal apoptosis, neuroinflammation, and impaired protein clearance, further exacerbating disease pathology.[45,47,48]

DISCUSSION

Multi-omic data-driven studies are the forefront of biomedical research. Large-scale multi-omic datasets have been generated to characterize the dysfunctional 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. However, it remains an open problem to integrate and interpret the multi-omic datasets to identify key molecular targets and core signaling pathways. In this study, we clearly formulated and defined the two types of need of multi-omic data analysis, i.e., hypothesis/anchor-target guided multi-omic data analysis and generic multi-omic data analysis. To tackle the multi-omic data analysis tasks, we proposed a graph AI model, M3NetFlow, which is specifically designed for integrating and analyzing multi-omic datasets and can deal with both of analysis tasks. Compared with existing models, the proposed model design and model architecture, mapping multi-omic data into signaling networks, combing local and global signaling, and

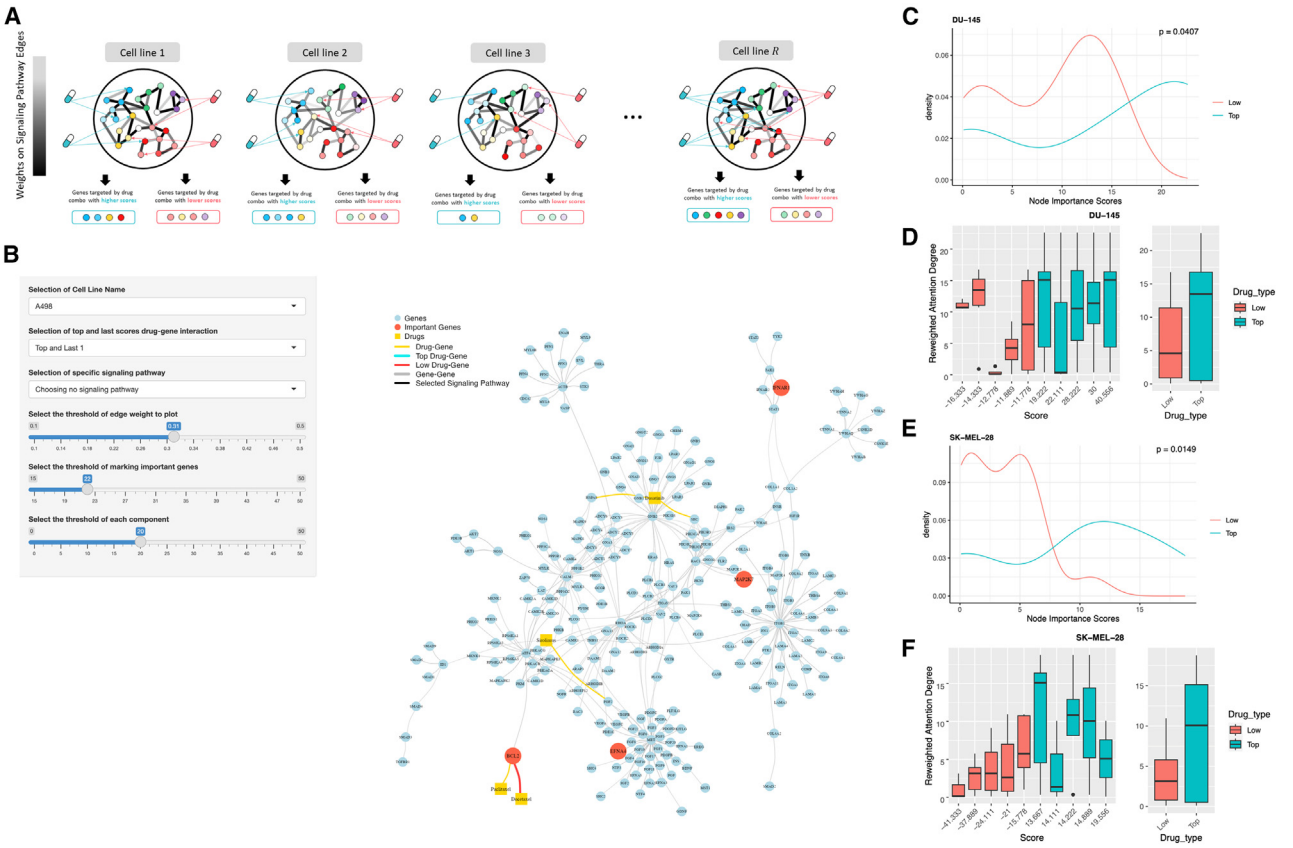


Figure 3. Target importance score patterns of synergistic and non-synergistic drug combinations

(A) Illustration of analyzing the important targets of top 5 synergistic and bottom 5 non-synergistic drug combinations.

(B) Visualization tool NetFlowVis for core signaling network interactions of cell line DU-145.

(C–F) Target importance score distribution and boxplots of the top 5 synergistic and bottom 5 non-synergistic drug combinations in DU-145 and SK-MEL-28 cell lines, and t-test p-values were used to statistically compare the two distributions in (C) and (E).

multi-hop flowing/propagation on the signaling network, are unique. The model prediction and interpretation (attention-based target and interaction scores) capabilities are demonstrated and evaluated using two case studies: drug combination synergy score prediction task (hypothesis/anchor-target guided multi-omic data analysis) and AD sample classification (generic multi-omic data analysis). The source code of M3NetFlow is publicly accessible, which enables users to modify and improve the model for their own studies. This method can be an alternative option and can be combined with other analysis methods, like the MAGINE for pathway and network module enrichment analysis.

Limitations of the study

This study is an exploratory effort in multi-omic data analysis, with several areas requiring further investigation. For example, more signaling pathways and larger protein-protein interactions should be evaluated. Moreover, dividing large signaling graphs into sub-networks or network modules can be achieved by using biologically meaningful annotations, such as gene ontology (GO) terms. In the current analysis, the generic pre-analyzed multi-omic features were used. It is worth investigating and selecting additional and biological meaningful features that can be derived from the multi-omic datasets in the future work. Additionally, more and 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. Aside from this, incorporating large language models (LLMs) into graph-based frameworks offers further possibilities, such as interpreting textual annotations/descriptions and associated GO terms or pathways to enrich the informative features, which can improve prediction accuracy and identify the accurate and interpretable biomarkers and core signaling pathways. Developing a hybrid graph-language framework that integrates structured data with unstructured text could enable contextual learning, while pre-trained LLMs like Generative Pre-training Transformer (GPT) or Bidirectional Encoder Representations from Transformer (BERT) may assist in mapping textual annotations to graph components, improving pathway discovery. Therefore, sophisticated and improved graph AI models are needed to integrate and interpret multi-omic datasets, identify and infer key

Figure 4. Boxplots of target importance scores of cell lines 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, HCI-H460, NCI-H522, OVCAR-3, OVCAR-4, OVCAR-8, PC-3, RPMI-8226, SF-268, SF-295, SF-539, SK-MEL-28, SK-MEL-5, SK-OV-3, SNB-75, SR, SW-620, T-47D, U251, UACC-257, UACC-62, and UO-31

molecular targets and signaling pathways of complex diseases, and guide the development of precision medicine.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Prof. Fuhai Li (fuhai.li@wustl.edu).

Materials availability

This study did not generate unique reagents.

Data and code availability

- The multi-omic data derived from cancer cell lines and AD samples, along with the drug synergy dataset and the dataset detailing AD sample conditions, are accessible through the key resources table.
- The data processing code and associated details, along with the M3NetFlow code, are available in the GitHub repository links given in the key resources table.

ACKNOWLEDGMENTS

This study was partially supported by NIA R56AG065352, NIA 1R21AG078799-01A1, NINDS 1RM1NS132962-01, Children's Discovery Institute M-II-2019-802, and NLM 1R01LM013902-01A1. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org).

AUTHOR CONTRIBUTIONS

F.L. conceived this study. F.L. and H.Z. designed the model and prepared the manuscript. H.Z. implemented the model and analyzed the data and results. F.L., H.Z., P.G., L.D., D.D., R.C.F., M.P., Y.C., and P.P. developed methods and discussed results.

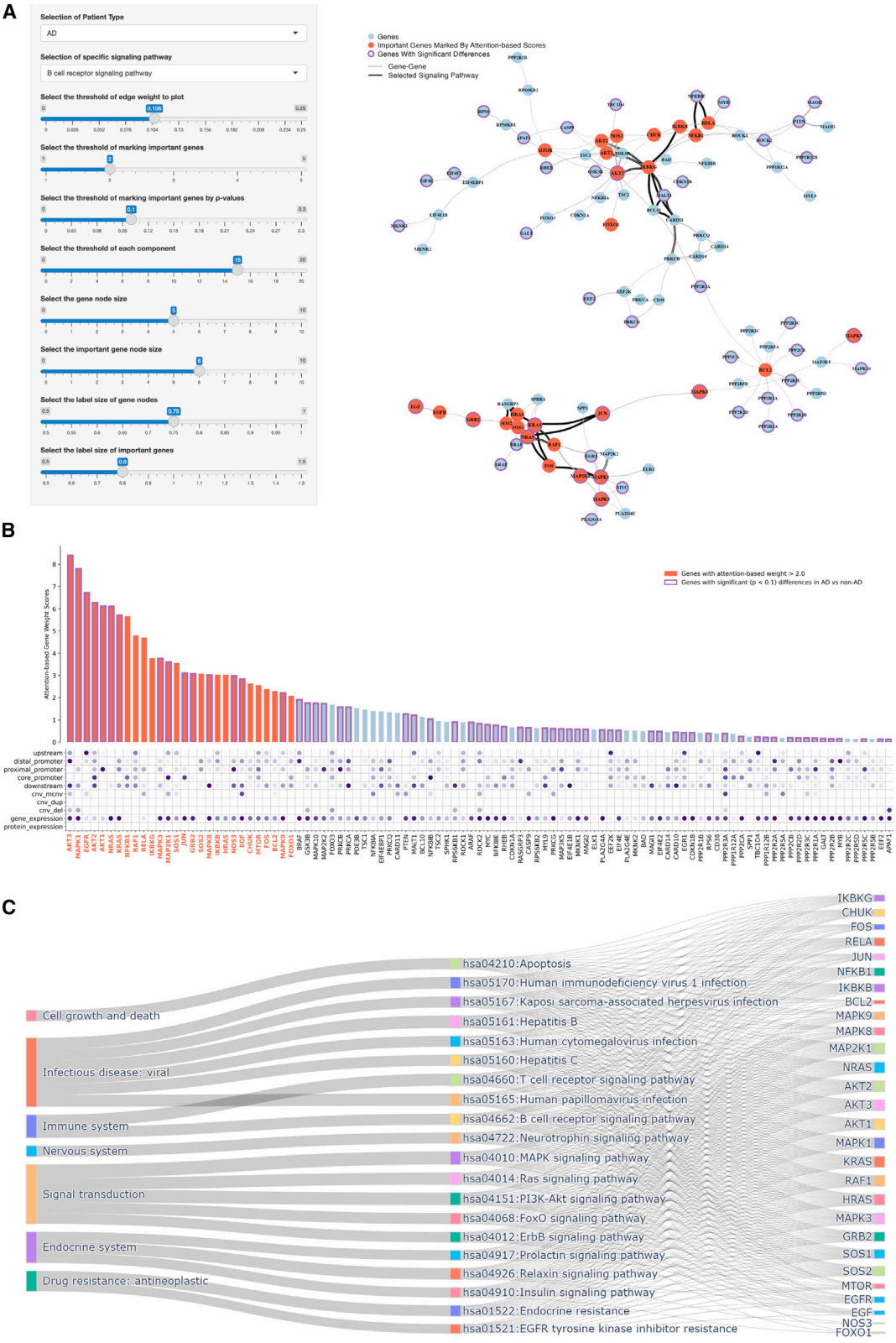
DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Datasets introduction for two case studies
  - Subgraph and multi-hop message propagation



(legend on next page)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111920.

Received: March 21, 2024
Revised: October 17, 2024
Accepted: January 27, 2025
Published: February 6, 2025

REFERENCES

1. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607. https://doi.org/10.1038/nature11003.
2. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, 955–961. https://doi.org/10.1093/nar/gks1111.
3. Li, F., Eteleeb, A.M., Buchser, W., Sohn, C., Wang, G., Xiong, C., Payne, P.R., McDade, E., Karch, C.M., Harari, O., and Cruchaga, C. (2022). Weakly activated core neuroinflammation pathways were identified as a central signaling mechanism contributing to the chronic neurodegeneration in Alzheimer's disease. Front. Aging Neurosci. *14*, 935279.
4. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.
5. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious orders study and rush memory and aging project. J. Alzheimers Dis. *64*, S161–S189.
6. Raghavachari, N., Wilmot, B., and Dutta, C. (2022). Optimizing Translational Research for Exceptional Health and Life Span: A Systematic Narrative of Studies to Identify Translatable Therapeutic Target(s) for Exceptional Health Span in Humans. J. Gerontol. A Biol. Sci. Med. Sci. *77*, 2272–2280. https://doi.org/10.1093/gerona/glac065.
7. Deelen, J., Evans, D.S., Arking, D.E., Tesi, N., Nygaard, M., Liu, X., Wojczynski, M.K., Biggs, M.L., van der Spek, A., Atzmon, G., et al. (2019). A meta-analysis of genome-wide association studies identifies multiple longevity genes. Nat. Commun. *10*, 3669. https://doi.org/10.1038/s41467-019-11558-2.
8. Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinf. Biol. Insights *14*, 1177932219899051.
9. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.
10. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237–i245.
11. Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput. Biol. *13*, e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
12. Bodein, A., Scott-Boyer, M.P., Perin, O., Lê Cao, K.A., and Droit, A. (2022). TimeOmics: an R package for longitudinal multi-omics data integration. Bioinformatics *38*, 577–579. https://doi.org/10.1093/bioinformatics/btab664.
13. Pino, J.C., Lubbock, A.L.R., Harris, L.A., Gutierrez, D.B., Farrow, M.A., Muszynski, N., Tsui, T., Sherrod, S.D., Norris, J.L., McLean, J.A., et al. (2022). Processes in DNA damage response from a whole-cell multi-omics perspective. iScience *25*, 105341. https://doi.org/10.1016/j.isci.2022.105341.
14. Cao, Z.J., and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat. Biotechnol. *40*, 1458–1466. https://doi.org/10.1038/s41587-022-01284-4.
15. Hamilton, W.L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. *2017*, 1025–1035.
16. Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., and Bengio, Y. (2018). Graph attention networks. In 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings (OpenReview.net), pp. 1–12.
17. Kipf, T.N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (OpenReview.net), pp. 1–14.
18. Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., and Huang, K. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat. Commun. *12*, 3445.
19. Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., and Zhu, Y. (2022). MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. Front. Genet. *13*, 806842.
20. Gao, H., Zhang, B., Liu, L., Li, S., Gao, X., and Yu, B. (2023). A universal framework for single-cell multi-omics data integration with graph convolutional networks. Briefings Bioinf. *24*, bbad081.
21. Li, G., Muller, M., Thabet, A., and Ghanem, B. (2019). DeepGCNs: Can GCNs go as deep as CNNs? In Proceedings of the IEEE International Conference on Computer Vision, *2019*, pp. 9266–9275. https://doi.org/10.1109/ICCV.2019.00936.
22. Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., and Grohe, M. (2019). Weisfeiler and leman go neural: Higher-order graph neural networks. Proc. AAAI Conf. Artif. Intell. *33*, 4602–4609.
23. Weisfeiler, Y.B., and Leman, A. (1968). The reduction of a graph to canonical form and the algebra which appears therein. Nauchno-Technicheskaya Informatsia (NTI Series) *2*, 12–16.
24. Nikolentzos, G., Dasoulas, G., and Vazirgiannis, M. (2020). k-hop graph neural networks. Neural Netw. *130*, 195–205.

Figure 5. Top-ranked proteins and associated pathways related to Alzheimer's disease

(A) Visualization tool NetFlowVis-AD for core signaling network interactions of AD. Red color indicates nodes with node score > 2.0. The purple circle represents nodes with at least one feature's p value < 0.1 between AD and control samples.

(B) Barplot of the node scores, where the red bars and purple borders have the same meaning as in (A). The color indicates the p value of individual omic data between AD and control samples.

(C) Sankey plot of top-enriched signaling pathways of the top-selected proteins.

25. Abu-El-Haija, S., Perozzi, B., Kapoor, A., Alipourfard, N., Lerman, K., Harutyunyan, H., Ver Steeg, G., and Galstyan, A. (2019). Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In International Conference on Machine Learning (PMLR), pp. 21–29.

26. Feng, J., Chen, Y., Li, F., Sarkar, A., and Zhang, M. (2022). How powerful are k-hop message passing graph neural networks. Adv. Neural Inf. Process. Syst. *35*, 4776–4790.

27. Li, R., Zhang, F., Li, T., Zhang, N., and Zhang, T. (2023). DMGAN: Dynamic multi-hop graph attention network for traffic forecasting. IEEE Trans. Knowl. Data Eng. *35*, 9088–9101.

28. Wang, L., Li, Z.W., You, Z.H., Huang, D.S., and Wong, L. (2023). MAGCDA: a multi-hop attention graph neural networks method for CircRNA-disease association prediction. IEEE J. Biomed. Health Inform. *28*, 1752–1761.

29. Holbeck, S.L., Camalier, R., Crowell, J.A., Govindharajulu, J.P., Hollingshead, M., Anderson, L.W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., et al. (2017). The National Cancer Institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. Cancer Res. *77*, 3564–3576. https://doi.org/10.1158/0008-5472.CAN-17-0489.

30. Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., and Tang, J. (2019). DrugComb: An integrative cancer drug combination data portal. Nucleic Acids Res. *47*, W43–W51. https://doi.org/10.1093/nar/gkz337.

31. Van Der Meer, D., Barthorpe, S., Yang, W., Lightfoot, H., Hall, C., Gilbert, J., Francies, H.E., and Garnett, M.J. (2019). Cell Model Passports - a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Res. *47*, D923–D929. https://doi.org/10.1093/nar/gky872.

32. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

33. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995.

34. Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: Predicting anti-cancer drug synergy with Deep Learning. Bioinformatics *34*, 1538–1546. https://doi.org/10.1093/bioinformatics/btx806.

35. Zhang, T., Zhang, L., Payne, P.R.O., and Li, F. (2021). Synergistic drug combination prediction by integrating multiomics data in deep learning models. Methods Mol. Biol. *2194*, 223–238.

36. Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. (2020). Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. Preprint at arXiv. https://arxiv.org/abs/2009.03509.

37. Corso, G., Cavalleri, L., Beaini, D., Liò, P., and Veličković, P. (2020). Principal neighbourhood aggregation for graph nets. Adv. Neural Inf. Process. Syst. *33*, 13260–13271.

38. Sweatt, J.D. (2004). Mitogen-activated protein kinases in synaptic plasticity and memory. Curr. Opin. Neurobiol. *14*, 311–317. https://doi.org/10.1016/j.conb.2004.04.001.

39. Glass, C.K., Saijo, K., Winner, B., Marchetto, M.C., and Gage, F.H. (2010). Mechanisms Underlying Inflammation in Neurodegeneration. Cell *140*, 918–934. https://doi.org/10.1016/j.cell.2010.02.016.

40. Heneka, M.T., Carson, M.J., El Khoury, J., Landreth, G.E., Brosseron, F., Feinstein, D.L., Jacobs, A.H., Wyss-Coray, T., Vitorica, J., Ransohoff, R.M., et al. (2015). Neuroinflammation in Alzheimer's disease. Lancet Neurol. *14*, 388–405.

41. Munoz, L., and Ammit, A.J. (2010). Targeting p38 MAPK pathway for the treatment of Alzheimer's disease. Neuropharmacology *58*, 561–568. https://doi.org/10.1016/j.neuropharm.2009.11.010.

42. Malumbres, M., and Barbacid, M. (2003). RAS oncogenes: the first 30 years. Nat. Rev. Cancer *3*, 459–465.

43. Bloom, G.S. (2014). Amyloid-β and tau: The trigger and bullet in Alzheimer disease pathogenesis. JAMA Neurol. *71*, 505–508. https://doi.org/10.1001/jamaneurol.2013.5847.

44. Braak, H., and Del Tredici, K. (2015). The preclinical phase of the pathological process underlying sporadic Alzheimer's disease. Brain *138*, 2814–2833. https://doi.org/10.1093/brain/awv236.

45. Talbot, K. (2014). Brain insulin resistance in Alzheimer's disease and its potential treatment with GLP-1 analogs. Neurodegener. Dis. Manag. *4*, 31–40. https://doi.org/10.2217/nmt.13.73.

46. Poo, M.M. (2001). Neurotrophins as synaptic modulators. Nat. Rev. Neurosci. *2*, 24–32.

47. Bomfim, T.R., Forny-Germano, L., Sathler, L.B., Brito-Moreira, J., Houzel, J.C., Decker, H., Silverman, M.A., Kazi, H., Melo, H.M., McClean, P.L., et al. (2012). An anti-diabetes agent protects the mouse brain from defective insulin signaling caused by Alzheimer's disease–associated Aβ oligomers. J. Clin. Investig. *122*, 1339–1353.

48. Nistri, S., and Bani, D. (2005). Relaxin in vascular physiology and pathophysiology: possible implications in ischemic brain disease. Curr. Neurovasc. Res. *2*, 225–233.

49. Zhang, H., Chen, Y., Payne, P., and Li, F. (2024). Using Signaling to mine signaling flows interpreting mechanism of synergy of cocktails. NPJ Syst. Biol. Appl. *10*, 92.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
NCI Drug Combination Synergy Score	NCI ALMANAC	https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-ALMANAC
O'Neil Drug Combination Synergy Score	DrugComb	https://drugcomb.fimm.fi/
Cell Model Passports RNA-Seq	Cell Model Passports	https://cog.sanger.ac.uk/cmp/download/rnaseq_20191101.zip
Cell Model Passports CNV	Cell Model Passports	https://cog.sanger.ac.uk/cmp/download/cnv_20191101.zip
CCLE Methylation	CCLE	https://data.broadinstitute.org/ccle/CCLE_RRBS_TSS1kb_20181022.txt.gz
CCLE Gene Amplification	CCLE	https://data.broadinstitute.org/ccle_legacy_data/binary_calls_for_copy_number_and_mutation_data/CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct
CCLE Gene Deletion	CCLE	https://data.broadinstitute.org/ccle_legacy_data/binary_calls_for_copy_number_and_mutation_data/CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct
ROSMAP_clinical	ROSMAP	https://www.synapse.org/#!Synapse:syn3191087
ROSMAP_arrayMethylation_imputed	ROSMAP	https://www.synapse.org/#!Synapse:syn3168763
ROSMAP.CNV.Matrix(Mutation)	ROSMAP	https://www.synapse.org/#!Synapse:syn26263118
ROSMAP_RNAseq_FPKM_gene	ROSMAP	https://www.synapse.org/#!Synapse:syn3505720
C2.median_polish_corrected_log2(Proteomic)	ROSMAP	https://www.synapse.org/#!Synapse:syn21266454
GEO GPL16304 Platform	Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16304
KEGG Signaling Pathways	KEGG	https://www.genome.jp/kegg/
Drug-Target Interaction Data	DrugBank	https://go.drugbank.com/
Software and algorithms		
Python version 3.10	Python Foundation Software	https://www.python.org/
PyTorch	PyTorch Official Website	https://pytorch.org/
Torch Geometric	PyG Official Website	https://pytorch-geometric.readthedocs.io/en/latest/
R Software version 4.2.2	R Foundation for Statistical Computing	https://www.r-project.org/
M3NetFlow	GitHub repository	https://github.com/FuhaiLiAiLab/M3NetFlow
NetFlowVis	shinyapps	https://m3netflow.shinyapps.io/NetFlowVis/
NetFlowVis-AD	shinyapps	https://m3netflow.shinyapps.io/NetFlowVis-AD/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This paper analyzes existing, publicly available data. The study does not use experimental models typical in life sciences.

METHOD DETAILS

Datasets introduction for two case studies

As aforementioned, we will apply the proposed model on two case studies: 1) hypothesis guided multi-omic data analysis: mechanism of synergy study using pairwise-drug combination synergy score (as the label data to train the model), multi-omic data of cancer cell lines (input features), drug targets (input anchor-targets), and KEGG signaling pathways (input graph); and 2) general multi-omic data analysis: AD biomarker discovery using sample diagnosis (AD or control as the label data), multi-omic data of samples (input features) and KEGG signaling pathways (input graph). The models will be trained using the label data to identify the key biomarkers and signaling flows based on the attention scores of signaling interactions (edges of the input graph). These datasets are publicly accessible. In key resources table, it shows the links to download the drug combination synergy scores, the pre-analyzed multi-omic data of cancer cell lines, KEGG signaling pathways and drug target information. It also shows the accessible information of the pre-analyzed multi-omic data and label information of AD.

Subgraph and multi-hop message propagation

In the graph message passing stages of our architecture (Figures 1B and 1C), 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 was designed. The multi-scale design will ensure the message fully interacts in the internal subgraph. We further made use of the K-hop attention-based design because it allows for the consideration of longer distance information flow from indirect neighboring nodes. Specifically, with those initial embedding features $X^{(m)} \in \mathbb{R}^{n\times d}$, $(m = 1,2,\ldots,M)$, the new embedding/features of the nodes are calculated as follows

$$\left(\alpha_{ij}^{(m)}\right)_p^{(k)} = \mathrm{ATT}_{hop}\left[X_p^{(m)}\right] = \frac{\exp\left(\mathrm{LeakyReLU}\left(a\left[W\left(X_p^{(m)}\right)_i \| W\left(X_p^{(m)}\right)_j\right]\right)\right)}{\sum_{q\in\mathcal{N}_i^{(k)}}\exp\left(\mathrm{LeakyReLU}\left(a\left[W\left(X_p^{(m)}\right)_i \| W\left(X_p^{(m)}\right)_q\right]\right)\right)}, \quad \text{(Equation 1)}$$

where $(\alpha_{ij}^{(m)})_p^{(k)}$ is the attention score between node i and node j in the k-th kop of the subgraph S_p for sample m using the K-hop attention function (ATT_{hop}). 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 $(X_p^{(m)})_i$ represent the feature feature of node i ($i = 1,2,\ldots,n$). Then the updated node embedding $(H_p^{(m)})_i$ for node i will be generated via K-hop message (MSG) propagation function by

$$\left(H_p^{(m)}\right)_i = \mathrm{MSG}_{hop}\left[X_p^{(m)}, \left(\alpha_{ij}^{(m)}\right)_p^{(k)}\right] = \frac{1}{KQ_i}\sum_{k}^{K}\sum_{q\in\mathcal{N}_i^{(k)}}\left[\left(\alpha_{iq}^{(m)}\right)_p^{(k)} W'\left(X_p^{(m)}\right)_q\right], \quad \text{(Equation 2)}$$

where Q_i represents the number of signaling pathway the node i ($i = 1,2,\ldots,n$) belongs to 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 Figure S1 for the algorithm of k-th hop adjacency matrix). Above formula shows the calculation for the 1-head attention and the linear transformation for features of each node will be defined as $W' \in \mathbb{R}^{d'\times d'}$. 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.

Global Bi-directional message propagation

Following the message propagation on the subgraphs, the global weighted bi-directional message propagation will be performed on the global integrated graph, 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) (Figure 1D). 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 $H^{(m)}$ as the initial node features with

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

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

$$\left(H^{(m)}\right)^{(L)} = \mathrm{MPN}\left(\left(H^{(m)}\right)^{(0)}\right), \quad \text{(Equation 4)}$$

where $(H^{(m)})^{(L)} \in \mathbb{R}^{n\times 3d^{(L)}}$ is the final embeddings after L-th layer message propagation and message passing network (MPN) is the layers of global weighted bi-directional graph neural network (WeB-GNN).[49]

Downstream tasks

Anchor-target guided multi-omic analysis

The drug combination synergy score prediction is used to demonstrate the hypothesis/anchor-target guided multi-omic data analysis. 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, the prediction of synergy score will be calculated as follows.

$$g\left(\left(H^{(m)}_{drugA}\right)^{(L)}, \left(H^{(m)}_{drugB}\right)^{(L)}\right) = \left(H^{(m)}_{drugA}\right)^{(L)} DUD^{\mathsf{T}}\left(\left(H^{(m)}_{drugB}\right)^{(L)}\right)^{\mathsf{T}}, \quad \text{(Equation 5)}$$

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

Generic multi-omic analysis

With the embedding of nodes, the global max pooling strategy was applied to predict the patient outcome with

$$\hat{y}^{(m)} = \text{argmax}\left(\text{MLP}\left(\text{MAX}\left[\left(H^{(m)}\right)^{(L)}\right]\right)\right), \quad \text{(Equation 6)}$$

where $\hat{y}^{(m)} \in \mathbb{R}^C$ and C is the number of sample types (e.g., AD vs control); multi-layer perceptron (MLP) is the linear function in artificial neural network; MAX is the maximum pooling function by extract the maximum value from embedded features (Figure 1D).

Signaling flow and target scores generation

Signaling interaction importance score

By extracting the edge weight from K-hop attention function ATT_{hop}, averaged edge importance score between node i and node j in k-th hop for specific sample type C_r ($r = 1, 2, \ldots, R$ and R is total number of sample types) will be generated by aggregating multiple signaling pathways and data points belong to sample type C_r

$$\left(\alpha^{(C_r)}_{ij}\right)^{(k)} = \frac{1}{|C_r|}\sum_{m\in C_r}\sum_{p=1}^{P}\left(\alpha^{(m)}_{ij}\right)_p^{(k)}, \quad \text{(Equation 7)}$$

where $(\alpha^{(m)}_{ij})_p^{(k)}$ will be aggregated from all signaling pathways under the specific sample type C_r and $(\alpha^{(C_r)}_{ij})^{(k)}$ represents the element of i-th row and j-th column in the k-th hop edge importance score matrix $(\mathcal{A}^{(C_r)})^{(k)}$ ($(\mathcal{A}^{(C_r)})^{(k)} \in \mathbb{R}^{n\times n}, k = 1,2,\ldots,K$). In this study, 1st hop edge importance score matrix of fold $\mathfrak{f}$, $(\mathcal{A}^{(C_r)})^{(1)}_{\mathfrak{f}}$, will be utilized to generate final edge importance score matrix by

$$\overline{\mathcal{A}}^{(C_r)} = \frac{1}{|\mathcal{F}|}\sum_{\mathfrak{f}=1}^{\mathcal{F}}\left(\mathcal{A}^{(C_r)}\right)^{(1)}_{\mathfrak{f}}, \quad \text{(Equation 8)}$$

where $\mathcal{F}$ is the number of fold used in cross validation.

Target/node importance score calculation

The weighted importance score of each node (gene) for specific sample type (e.g., cell line in cancer dataset or patient type in AD dataset) will be calculated based on the attention with

$$\mathcal{D}^{(C_r)}_g = \sum_{i}^{n}\overline{\mathcal{A}}_{ig}{}^{(C_r)} + \sum_{j}^{n}\overline{\mathcal{A}}_{gj}{}^{(C_r)}, \quad \text{(Equation 9)}$$

where $\overline{\mathcal{A}}^{(C_r)}_{ig}$ is the element of averaged fold attention matrix $\overline{\mathcal{A}}^{(C_r)}$ from a sample type C_r in the i-th row and g-th column and $\mathcal{D}^{(C_r)}_g$ is the importance score for node g for specific sample type C_r. Therefore, $\mathcal{D}^{(C_r)}_g$ will be used to construct the node importance score vector $\mathcal{D}^{(C_r)}$ ($\mathcal{D}^{(C_r)} \in \mathbb{R}^n$). The overall node importance score matrix, $\mathcal{D}$ ($D \in \mathbb{R}^{n\times R}$), will be generated by concatenating the node importance score vector with $\mathcal{D} = [\mathcal{D}^{(C_1)}, \mathcal{D}^{(C_2)}, \ldots, \mathcal{D}^{(C_r)}, \ldots, \mathcal{D}^{(C_R)}]$. Specifically, for calculating the node importance score for cancer dataset, some of genes may show the importance of relatively higher scores in each sample, which weakens the analysis of the cell-line-specific analysis. In this way, the idea of reweighting the gene importance score was created based on the Term Frequency – Inverse Document Frequency (TF-IDF) is employed to redistribute the importance scores of genes across different cell lines to reweight the node importance in each sample with

$$\mathcal{W}_g = Log\left(\frac{R+1}{S_g+0.01}\right) \quad \text{(Equation 10)}$$

$$\mathcal{D}'^{(C_r)}_g = \mathcal{W}_g\mathcal{D}^{(C_r)}_g, \quad \text{(Equation 11)}$$

where S_g is the number of cell lines with a higher node importance score than the threshold $\mathcal{S}$ in all cell lines and $\mathcal{W}_g$ is the weight coefficient to adjust the importance score in each sample type. In this study, the 95 percentile of node importance scores in the whole matrix $\mathcal{D}$ was used for setting $\mathcal{S}$. Finally, the reweighted node importance score matrix $\mathcal{D}'$ will be generated by concatenating the reweighted node importance score vector with $\mathcal{D}' = [\mathcal{D}'^{(C_1)}, \mathcal{D}'^{(C_2)}, \ldots, \mathcal{D}'^{(C_r)}, \ldots, \mathcal{D}'^{(C_R)}]$ for cancer targets identification.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation metrics

The evaluation metrics of predictions were Pearson correlation for drug synergy scores and accuracy for AD status classification.

Statistical analysis

We performed a t-test to assess the differences in importance scores of drug targets between the top drug combinations (with the highest synergy scores) and the bottom drug combinations (with the lowest synergy scores).

ADDITIONAL RESOURCES

We developed ***NetFlowVis*** application is provided for visualizing the results. The cancer results can be accessed through the following link: https://m3netflow.shinyapps.io/NetFlowVis/, while the visualization tool of ROSMAP AD results is available at https://m3netflow.shinyapps.io/NetFlowVis-AD/.