

CGMega: Explainable Graph Neural Network Framework with Attention Mechanisms for Cancer Gene Module Dissection

Hao Li (✉ lihao_thu@163.com)

Institute of Health Service and Transfusion Medicine <https://orcid.org/0000-0002-9464-1372>

Zebei Han

Shanghai Jiao Tong University

Yu Sun

Beijing Institute of Radiation Medicine

Fu Wang

Shanghai Jiao Tong University

Pengzhen Hu

School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China.

Chao Ren

Beijing Institute of Radiation Medicine

Xiang Xu

Institute of Health Service and Transfusion Medicine, Beijing 100850, China.

Hebing Chen

Beijing Institute of Radiation Medicine <https://orcid.org/0000-0003-4102-356X>

Yang Yang

Shanghai Jiao Tong University <https://orcid.org/0000-0001-5720-773X>

Xiaochen Bo

Institute of Health Service and Transfusion Medicine <https://orcid.org/0000-0003-3490-5812>

Article

Keywords:

Posted Date: July 21st, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3180743/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

1 CGMega: Explainable Graph Neural Network

2 Framework with Attention Mechanisms for Cancer

3 Gene Module Dissection

4 Hao Li ^{1,3,#}, Zebei Han ^{2,#}, Yu Sun ^{1,#}, Fu Wang ², Pengzhen Hu ⁴, Chao Ren ¹, Xiang Xu
5 ¹, Hebing Chen ^{1,*}, Yang Yang ^{2,*}, Xiaochen Bo ^{1,*}

6 ¹ Institute of Health Service and Transfusion Medicine, Beijing 100850, China

7 ² Department of Computer Science and Engineering, Shanghai Jiao Tong University, and
8 Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive
9 Engineering, Shanghai 200240, China

10 ³ Beijing Institute of Radiation Medicine, Beijing 100850, China

11 ⁴ School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China

12 # These authors contributed equally

13 * These authors jointly supervised this work

14 Abstract

15 Cancer is rarely the straightforward consequence of an abnormality in a single
16 gene, but rather reflects a complex interplay of many genes, represented as gene
17 modules. Here, we leveraged the recent advances of model-agnostic interpretation
18 approach and developed CGMega, an explainable and graph attention-based deep
19 learning framework to perform cancer gene module dissection. CGMega outperforms
20 current approaches in cancer gene prediction, and it provides a promising approach to
21 integrate multi-omics information. We applied CGMega to breast cancer cell line and
22 acute myeloid leukemia (AML) patients, and we uncovered the high-order gene module
23 formed by ErbB family and tumor factors *NRG1*, *PPM1A* and *DLG2*. We identified 396
24 candidate AML genes, and observed the enrichment of either known AML genes or
25 candidate AML genes in a single gene module. We also identified patient-specific AML
26 genes and associated gene modules. Together, these results indicate that CGMega can
27 be used to dissect cancer gene modules, and provide high-order mechanistic insights
28 into cancer development and heterogeneity.

29 Introduction

30 The complex functions of a living cell are conducted through the concerted activity

of many genes and gene products. Much of the activity of a cell is organized into gene modules: sets of genes that are coregulated to respond to different conditions ¹. Gene modules have been widely studied in cell identity dissection ², transcription factor (TF)-gene regulation ^{3,4}, functional genome annotation ⁵, disease progression ⁶, disease origin ⁷, drug repurposing ⁸, and cancer research ^{9,10}. Disease-associated genes are not randomly scattered across biological networks. Instead, they tend to be located in disease gene modules ^{11,12}. Dissecting the gene modules that drive disease progression enables screening for the molecules that correct the network rather than targeting peripheral downstream effectors that may not be disease modifying ¹³⁻¹⁵. Active driver modules can trigger the hallmarks of cancer and confer fitness advantages to cancer cells ^{16,17}. The elucidation of cancer gene modules can substantially further our understanding of cancer development and inform the design of optimal treatments ^{18,19}.

Ever since the development of high-throughput sequencing technologies, gene module detection methods have been a cornerstone for the biological interpretation of large gene compendia. Numerous approaches and algorithms have been proposed for the detection of gene modules through measuring gene expression ^{3,20,21} and across omics information ²²⁻²⁵. However, the methods currently in use suffer from two main drawbacks. First, recent technologies for chromosome conformation capturing have uncovered the three-dimensional (3D) genome architecture and demonstrated its critical role in establishing gene-gene relationships ^{26,27}. Apart from the genome, epigenome, transcriptome, and proteome information, the 3D chromatin structure information is indispensable for detecting gene modules, especially in cancer study. But few related work has yet investigated gene module using Hi-C data. 3D chromatin data such as Hi-C data can either be represented as gene attributes that depict the spatial features of genes, or structural relationships between genes. Thus, how to encode Hi-C data and integrate with other multi-omics data challenge the precise detection of gene modules. Second, gene modules exhibit characteristics of high-order network, and the high-order interactions regulate complex functions in biological systems ²⁸. For example, essential genes are densely connected hubs in gene modules ^{29,30}. Deciphering the high-order relationships that are embedded in gene modules remains challenging. Also, multi-omics features have different impacts on each gene from the same gene module. For example, many genes that play important roles in tumorigenesis are not only altered on the level of their DNA sequences, but are dysregulated through epigenetic effects or other cellular mechanisms ^{16,31,32}. Most current co-expression clustering or correlation-based methods are not able to assign important omics features

to module genes.

Recent studies have shown the utility of deep learning algorithms to data-driven sciences, in particular, to biological data analysis³³. These approaches provide the ability to measure large and complex multi-omics data, and support the discovery of unanticipated relationships, and from which novel hypotheses and models can be derived and used to make predictions³⁴. It is worth mentioning that graph neural network (GNN) constitutes a powerful approach for measuring graph-structured data such as biological network³⁵, and succeed in modeling PPIs³⁶ and Hi-C data³⁷, as well as in discovering gene modules across cellular networks³⁸. GNN is capable of handling different Hi-C representations, that is, either gene attributes as node features or relationships between genes as graph edges. Moreover, we have demonstrated that GNN together with its interpretation techniques are powerful tools for dissecting high-order relationships among genome interactions³⁹. Together, advances in graph deep learning make it possible to dissect cancer gene modules from a multi-omics perspective, further leveraging deep understanding of the underlying data.

Here, we present a new framework (CGMega) for dissecting cancer gene module with explainable graph attention. First, we constructed a multi-omics representation graph in which nodes were genes and edges were defined as protein-protein interactions (PPIs) between genes. Node features are concatenation of condensed Hi-C features, promoter densities of ATAC, CTCF, the histone modifications H3K4me3 and H3K27ac, and frequencies of single nucleotide variants (SNVs), copy number variants (CNVs). Then, CGMega utilized a transformer-based graph attention neural network over the multi-omics representation graph and predicted cancer genes in a semi-supervised manner. The good performance of CGMega (area under the precision recall curve, AUPRC 0.9140) ensured downstream cancer gene module detection. Finally, we adopted the model-agnostic approach GNNExplainer⁴⁰ to interpret the contribution factors to cancer genes under the context of multi-omics, and further detected the cancer gene modules with representative features. We applied CGMega to breast cancer cell line and acute myeloid leukemia (AML) patients, and uncovered the high-order relationships between genes in cancer gene modules. Together, CGMega harnesses the recent advent of GAT over multi-omics data, and gains fundamental discovery and understanding of the hierarchy in cancer gene modules.

98 **Results**

99 **Overview of CGMega framework**

100 We proposed a new framework, CGMega, for studying cancer gene module based
101 on graph attention and graph interpretation technologies (Fig. 1a). CGMega leverages
102 a combination of multi-omics data across genomics, epigenomics, proteome, and
103 especially 3D genome architecture. In CGMega, we first removed the potential effects
104 of structural variation on Hi-C contact map and normalized it with iterative correction
105 and eigenvector decomposition (ICE) ⁴¹, and calculated the spatial distances between
106 genes (see Methods). Then, singular value decomposition (SVD) was applied on the
107 normalized matrix to get condensed Hi-C features (See Methods). Simultaneously, we
108 calculated the SNV and CNV frequencies for each gene and calculated epigenetic
109 densities within each gene promoter (see Methods). Then, we constructed a multi-omics
110 information combination graph, in which the nodes represent genes and the edges are
111 obtained from PPIs. The features of nodes are the concatenation of condensed Hi-C
112 features, SNV and CNV frequencies, and epigenetic densities. Notably, based on the
113 detailed evaluation in the following section, we deployed Hi-C data as node features
114 instead of edge features. Further, we constructed a transformer-based GAT model to
115 predict cancer genes in a semi-supervised manner (see Methods). Finally, CGMega
116 implemented the model-agnostic approach GNNExplainer to detect cancer gene
117 modules. GNNExplainer utilizes a masking approach to detect the compact subgraph
118 structure and a small subset of node features that have a crucial role in GNN prediction
119 ⁴⁰. Applying GNNExplainer, we identified the subset of genes that are most influential
120 for the prediction of cancer genes, together forming the cancer gene modules (Fig. 1b).
121 These genes are one-hop or two-hop neighbors to cancer genes, and GNNExplainer
122 also provides important features for each gene. To examine the robustness of
123 interpretation results in CGMega, we repeated GNNExplainer and obtained high
124 consistent cancer gene modules (Figs. S1a and S1b). In sum, the output of CGMega is
125 the probability of each gene being a cancer gene and their influential genes interpreted
126 from GATs. Gene-specific features are also assigned to these genes and together formed
127 the gene modules.

Figure 1

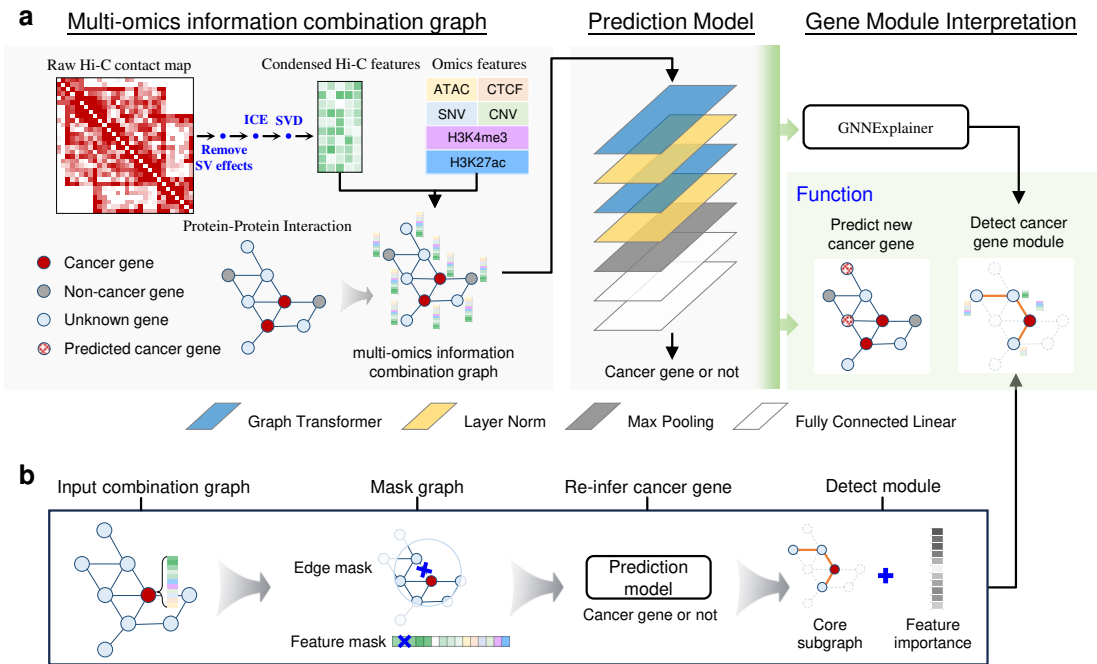


Fig. 1 Overview of CGMega framework. **a** CGMega pipeline. First, condensed Hi-C features were obtained by removing SV effects, ICE normalization, and SVD on raw Hi-C contact map step-by-step. Simultaneously, we calculated omics features including SNVs and CNVs frequencies for each gene as well as epigenetic densities within each gene promoter. To combine multi-omics information, we created a graph, where nodes represent genes and edges are derived from CPDB PPIs. Node features were the concatenations of condensed Hi-C features and omics features. Next, a cancer gene prediction model consisting of two Graph Transformer layers, two Layer-Norm layers, one max-pooling layer, and two fully connected linear layers was constructed. Finally, the model-agnostic approach GNNExplainer was employed to detect cancer gene modules. **b** GNNExplainer interpretation. Given a gene (represented as a node in graph), GNNExplainer identified a subgraph G that contains the relevant features crucial for the prediction. G is a connected subgraph where the gene nodes cover at most a two-hop region with no more than 20 edges.

Figure S1

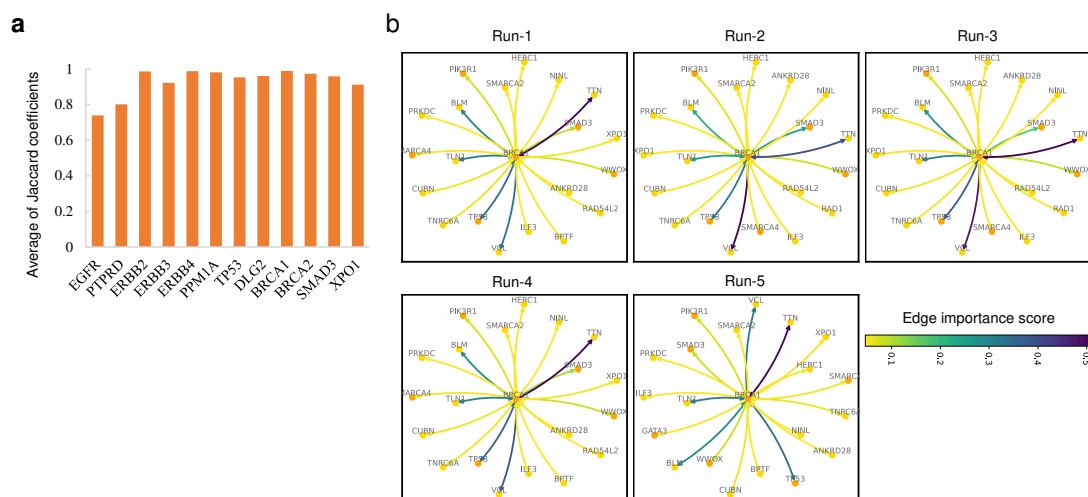


Fig. S1 robustness of GNNExplainer in CGMega. **a** We repeated GNNExplainer five times on 12 randomly selected cancer genes, and calculated the average Jaccard coefficients between any two repeats. **b** Taken BRCA1 gene module as an example to show the robustness of GNNExplainer interpretation.

CGMega is effective in cancer gene prediction

CGMega outperformed current methods. CGMega identified gene modules based on the accurate prediction of cancer genes, and we thus tested the performance of CGMega in cancer gene prediction on the MCF7 cell line (see Methods), a human breast cancer cell line with high-confidence multi-omics data. CGMega achieved 0.9140 AUPRC (Fig. 2a) and 0.9630 area under the receiver operating characteristic curve (AUROC) (Fig. S2a). To demonstrate the advances of CGMega in cancer genes prediction task, we compared CGMega with other methods (see Methods), including GCN, GAT, MLP, SVM, and the state-of-the-art methods MTGCN⁴² and EMOGI²⁵. By computing AUPRC, AUROC, accuracy (ACC), and F1 score, CGMega outperformed all other methods across these four metrics (Fig. 2b).

CGMega supports few-shot transfer learning. CGMega predicts cancer genes in a semi-supervised manner, where sufficient labeled genes are required, and this is limited in rare cancer research. To resolve this, we pre-trained CGMega on the MCF7 cell line, and then performed fine-tuning on other cancers. First, we tested retrained CGMega (training from scratch) and pre-trained CGMega on K562 cell line using all labeled genes (597 positives and 1839 negatives). Pre-trained CGMega achieved a higher accuracy and F1 score and comparable AUPRC and AUROC values (Fig. S2b). Then, we tested retrained CGMega and pre-trained CGMega using down-sampled labeled genes. Here, we also tested CGMega models without Hi-C features. As the number of

labeled genes decreased, the performance of retrained CGMega dropped sharply while the pre-trained CGMega continued to have high performance (Fig. 2c). Moreover, the Hi-C features exhibited powerful improvements in prediction especially when the labeled genes were less than 200. Further, we compared the performance of few-shot transfer learning in CGMega with other methods, and pre-trained CGMega had the highest value (Fig. S2c).

Ablation experiments. CGMega leverages 15-dimensional gene features including 10-dimensional omics features and 5-dimensional condensed Hi-C features derived from dimensionality reduction of the Hi-C data. We performed ablation experiments by removing or shuffling gene features (Fig. S2d), and we observed that both omics and Hi-C features made contributions for model prediction (Fig. 2d). Moreover, CGMega with 5-dimensional condensed Hi-C-only features was not as good as CGMega with 10-dimensional omics features, suggesting that the structural features may have a compensatory effect on the quality of omics features.

Robustness assessment. We tested CGMega on Hi-C data with different resolutions and read depths, CGMega maintained its stable performance using Hi-C data with resolutions from 5-kb resolution to 25-kb and the AUPRC slightly dropped while the Hi-C read depth decreased (Fig. 2e), demonstrating that our approach is robust in its adaptation to scenarios with lower data quality and holds promise for a wide range of application settings. We also tested CGMega on datasets with different ratios of positive to negative. CGMega can still achieve stable well performance with extreme ratios (Fig. S2e). In addition, CGMega is effective for different PPI datasets (Fig. S2f).

Figure 2

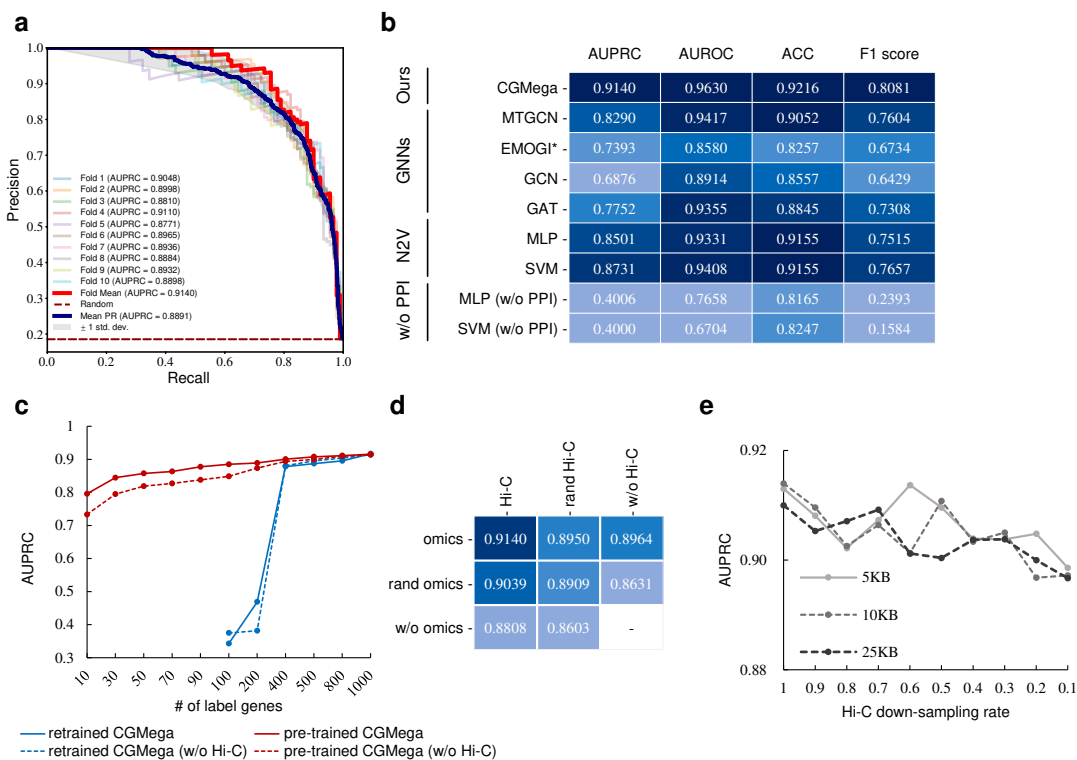


Fig. 2 CGMega performance in cancer gene prediction task. **a** AUPRC on breast cancer cell line MCF7. **b** Methods comparison on MCF7 cell line. EMOGI*: The input dimension used in CGMega (15-dimension) was much lower than that in EMOGI (64-dimension). EMOGI was ran with our 15-dimension and got a relatively lower performance (AUPRC = 0.1856), and thus we presented EMOGI performance as reported in the original paper. N2V represents node2vector. MLP and SVM were tested with and without (w/o) PPIs. **c** AUPRCs of retrained CGMega and pre-trained CGMega on datasets with different numbers of labeled genes. **d** Ablation experiments. AUPRCs of CGMega with random or without omics features and Hi-C features. **e** AUPRC of CGMega on Hi-C data under different resolutions and down-sampling rates.

Figure S2

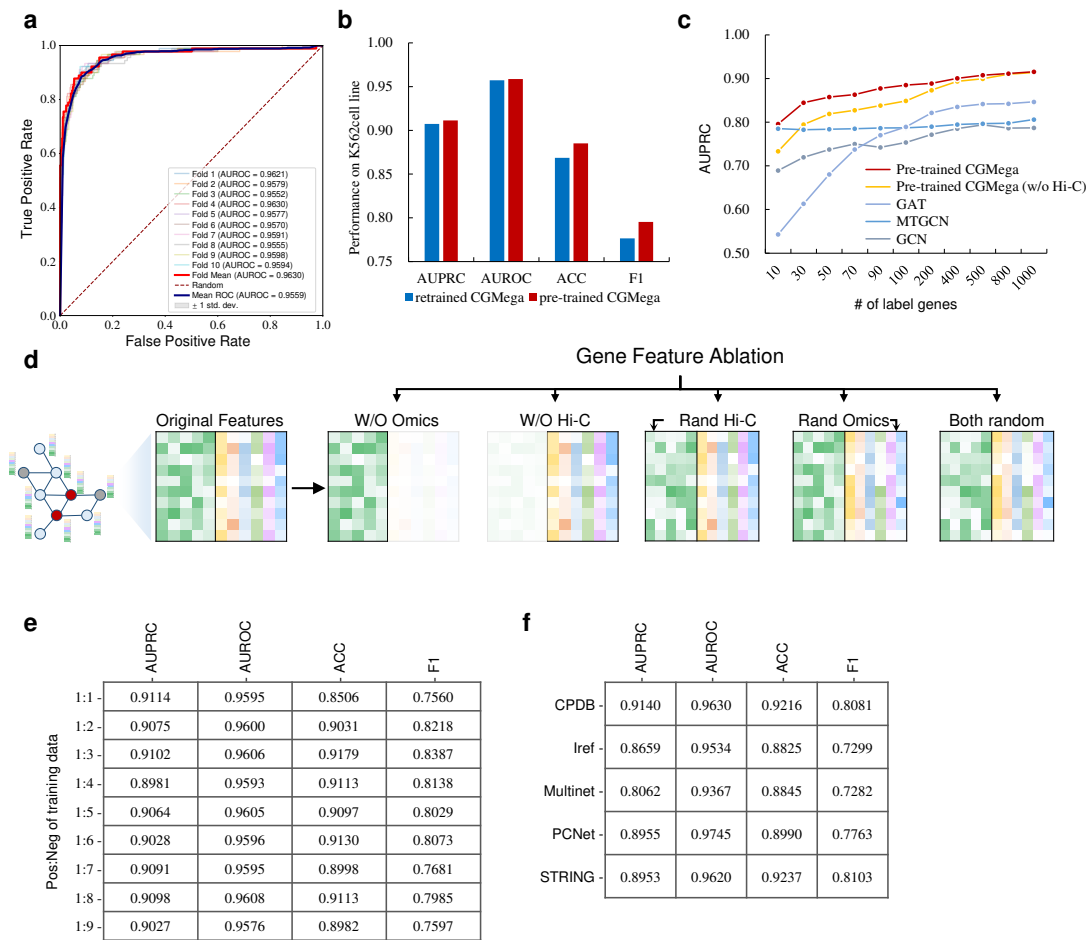


Fig. S2 CGMega performance in cancer gene prediction task. **a** AUROC on breast cancer cell line MCF7. **b** Performance of retrained and pre-trained CGMega on K562 cell line. **c** Methods comparison on datasets with different number of labeled genes. **d** Gene feature ablation illustration. Five types of feature ablation inputs were generated, including the removal of omics feature and Hi-C feature and random generation of Hi-C feature, omics feature, and both omics and Hi-C features. **e** Performance of CGMega on datasets with different positive to negative ratios. **f** Performance of CGMega on different PPI databases.

CGMega provides a new strategy for multi-omics data integration

The outperformance of CGMega benefits from the effective integration of multi-omics information, including genome, epigenome, proteome, and especially the 3D genome architecture. Hi-C is currently the most widely used assay for investigating the 3D genome organization. However, measuring Hi-C data together with other omics data is often limited by its noise, sparsity, and variable resolution. To obtain the best performance on the cancer gene prediction task, we tested integration approaches with different Hi-C data embeddings (Fig. 3a).

218 *Regarding Hi-C data as gene linkages.* Molecular networks are important issues
219 in biological studies ^{2,11,29}. For example, EMOGI has demonstrated the utility of PPIs
220 in cancer gene prediction ²⁵. Hi-C data measures the interactions that connect different
221 genomic loci and thus enables the construction of gene interaction networks. Using Hi-
222 C contact maps, we constructed unweighted and weighted networks, respectively. In
223 the unweighted network, interactions between genes were determined by the existence
224 or nonexistence of contacts. For weighted networks, interaction values were log10 or
225 tenth root of contact strength. Then, epigenetic information was assigned as gene
226 features. Finally, we combined gene interaction network with the PPI network and
227 constructed three types of graphs: a Hi-C only graph, a Hi-C/PPI independent graph,
228 and a Hi-C/PPI combined graph. In these, nodes represent genes and the node features
229 are epigenetic information. We trained GAT-based neural networks on these graphs.
230 Among these methods, the Hi-C only graph was ineffective for predicting cancer genes
231 (AUPRC < 0.5). The Hi-C/PPI independent graph exhibits only a marginal
232 improvement over the PPI-only strategy. It is solely when Hi-C is combined with PPI
233 that a modest increase, of roughly half a point, is observed in the two-edge construction
234 methods (Fig. 3b). This result does not offer compelling support for the inclusion of Hi-
235 C as graph structure information within the model.

236 *Regarding Hi-C data as gene features.* Hi-C data are intuitively used for measuring
237 gene interactions. However, due to the noise and sparsity of Hi-C data, gene interaction
238 networks based on Hi-C data tend to be incomplete and flawed. For this reason, we
239 tested different methods of obtaining condensed Hi-C features, including Node2Vec,
240 SVD, locally linear embedding (LLE), isometric feature mapping (ISOMAP), non-
241 negative matrix factorization (NMF), and *t*-SNE. The condensed Hi-C features were
242 concatenated with epigenetic information as gene features. PPI networks were still used
243 to measure the interactions between genes. We also trained GAT-based neural networks
244 on these graphs, and the situation improved significantly. Generally, incorporating Hi-
245 C features using dimensionality reduction methods improved the prediction of cancer
246 genes. The best-performing method, SVD, achieved an AUPRC of 0.9140, while
247 Node2Vec, NMF, and *t*-SNE also demonstrated promising results (Fig. 3c).
248 Additionally, we compared the impact of different dimensions of condensed Hi-C
249 features for model prediction (Fig. 3d). Combining the four metrics, all methods with
250 the Hi-C feature received a performance improvement compared to those without the
251 Hi-C feature (Fig. S3a). SVD-based reduction of Hi-C data to a condensed five-
252 dimensional feature set was found to be the optimal solution based on both results.

Taken together, by systematically comparing different integration approaches with Hi-C data embedding, we showed that, in cancer gene prediction task, using Hi-C latent features as gene features outperforms measuring Hi-C data as the gene interactions directly. SVD is an effective dimensionality reduction method for combining Hi-C data with other omics data.

Figure 3

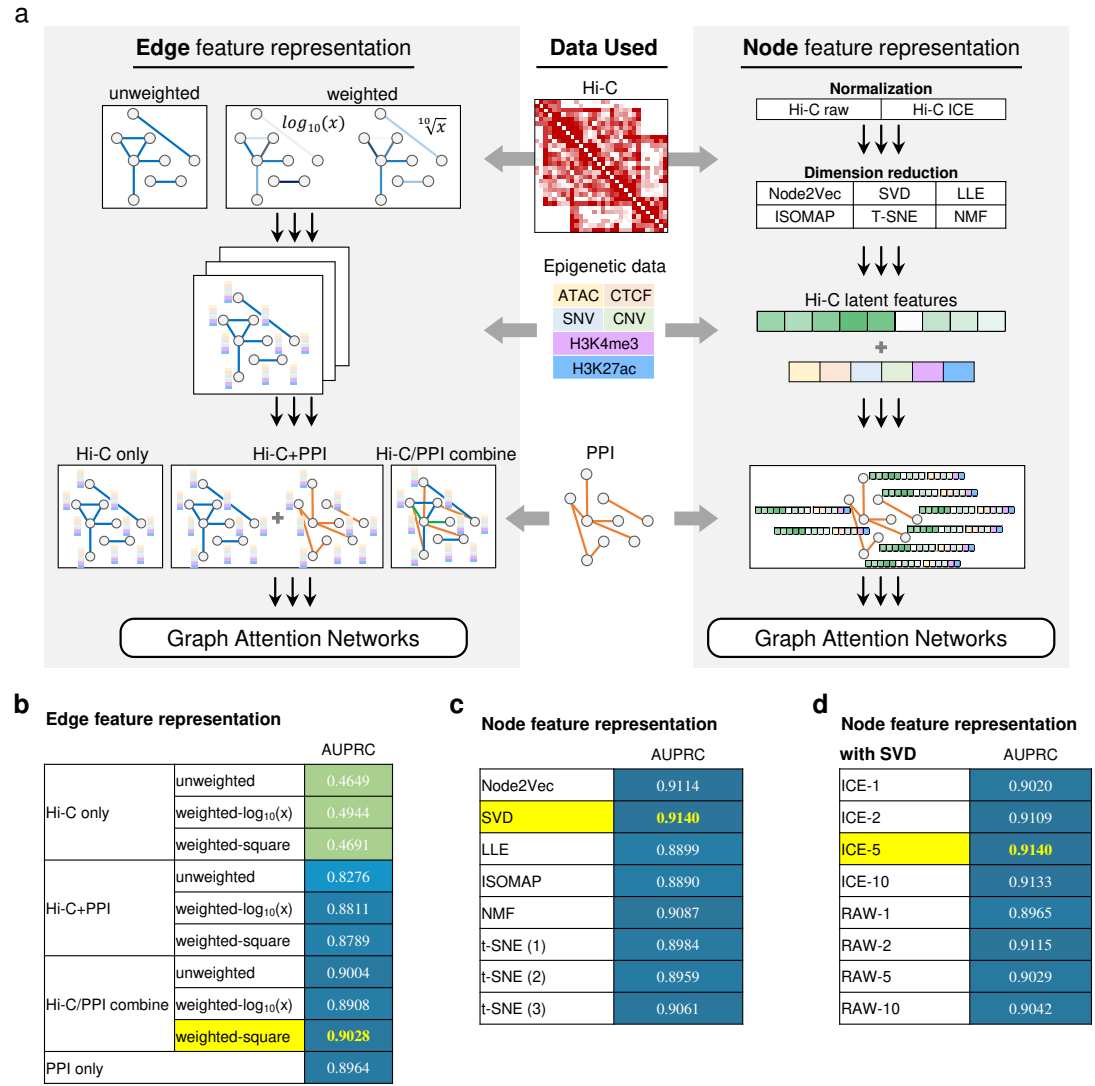


Fig. 3 Performance evaluation of multi-omics data integration approaches. **a** Design of multi-omics data integration. (Left) Hi-C data were regarded as graph edges with two types. In the unweighted graph, edges were determined by the existence of Hi-C contacts or not. In the weighted graph, edge weights were interaction values calculated as $\log_{10}$ or tenth root Hi-C contact maps. In either graph, node features were omics features. To combine Hi-C with PPIs, we performed GAT-based networks on three graph inputs, including (1) GAT on Hi-C graph alone, (2) two GATs on Hi-C and PPI respectively and then combined embeddings, and (3) first combined Hi-C graph and PPIs then performed GAT. (Right) Hi-C data were regarded as node features. Raw and normalized Hi-C data

as well as different dimensionality reduction methods were tested. Then, condensed Hi-C data were concatenated with omics features, and complete node features were formed. Graph edges were determined by PPIs. For either graph structure (left or right), a GAT-based cancer gene prediction model was made, as described in Fig. 1. **b** AUPRC of cancer gene prediction model with Hi-C input as edge features. **c** AUPRC of cancer gene prediction model with Hi-C input as node features. **d** AUPRC of cancer gene prediction model with raw and normalized Hi-C.

Figure S3

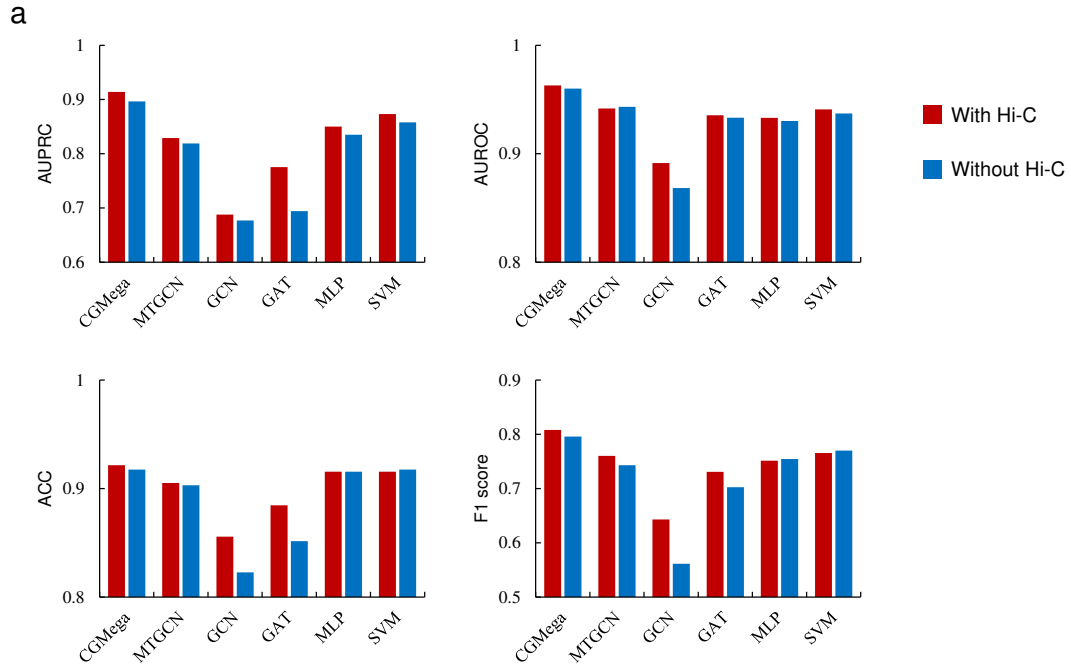


Fig. S3 a Methods comparison on input data with or without Hi-C.

Gene modules with multi-omics features in human breast cancer cell line

CGMega detects gene modules based on model-agnostic neural network interpretation approach (Fig. 1b), and these gene modules consist of two parts: i) a core subgraph consisting of the most influential pairwise relationships for the prediction of cancer gene, and ii) 15-dimensional importance scores that quantify the contributions of each gene feature to cancer gene prediction. We applied CGMega to the human breast cancer MCF7 cell line and examined the modules of 358 known cancer genes (Supplementary Data 1). These cancer genes were not randomly scattered throughout gene modules; they tended to be co-located in same modules (Fig. S4a). This is consistent with previous reported as so-called disease modules⁸. Among these gene modules, *TP53* showed the highest enrichment and participated in 139 cancer gene modules, followed by *ESR1* (63 participations) and *AKT1* (61 participations) (Fig. 4a).

In addition to these well-known cancer genes, we observed another 12 highly module-participating genes such as *XPO1*, *NCOR2*, and *PPM1A*. These genes may be the collaborators of well-known cancer genes. We also examined the structural features of gene modules with respect to their graphical metrics including transitivity, clustering coefficients, degree centrality, and betweenness centrality, and we found that topological structure of cancer gene modules were significantly more consistent than that of non-cancer gene modules ($p < 2.47\text{e-}5$, paired t -test) (Fig. S4b).

Beyond the topology of gene modules, we next investigated the feature importance scores. CGMega utilized 15-dimensional multi-omics features as inputs and generated an importance score for each feature. It is necessary to examine whether the importance scores were just related to the corresponding input. We thus examined the distributions of these two values, and found that feature importance is irrelevant to input data (Pearson correlation coefficients $r < 0.26$, Fig. S4c), suggesting that the importance score is the interpretation of neural networks instead of simple determination due to its input. Besides, the feature importance scores were not evenly distributed; instead, one or several features were dominant (Fig. S4d). We defined the representative features (RFs) as features that have top-ranked importance scores (Fig. 4b, see Methods for details). For example, the RF of the gene *TP53* is SNV while gene *PIK3RI*'s RF is Hi-C. Generally, 1158 genes had only one RF and 149 genes had multiple RFs (Fig. 4c).

We next focus on the gene modules of *BRCA1* and *BRCA2*, which are the most commonly encountered genes in breast cancer. As previously reported, these two cancer genes play different roles in the common pathway of genome protection⁴³. We also observed topological differences between their gene modules. In brief, *BRCA1*, which is a pleiotropic DNA damage response (DDR) protein working in several stages of DDR, was also found to be widely connected with another 20 genes (Fig. 4d). By contrast, *BRCA2*, as a mediator of the core mechanism of homologous recombination (HR), was connected with other genes via *ROCK2*, an important gene that directly mediates HR repair⁴⁴. Moreover, SNV was the RF for both *BRCA1* and *BRCA2*. We also observed a high-order gene module combined from the *BRCA1* gene module and the *BRCA2* gene module through three shared genes including *TP53*, *SMAD3*, and *XPO1* (Fig. 4e). Taken together, these indications mean that CGMega is capable of detecting the interpretable and high-order gene modules with multi-omics features.

Figure 4

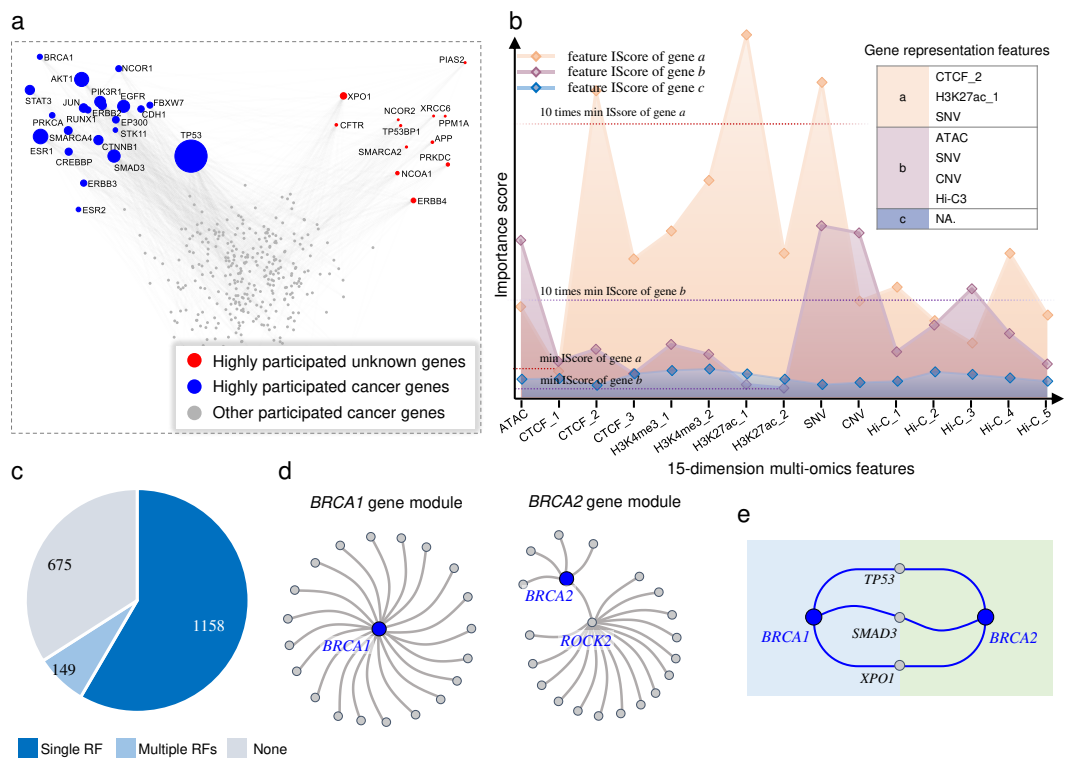


Fig. 4 Gene modules in breast cancer cell line. **a** Scatter plot shows the gene participation in cancer gene modules. In gene modules of 358 well-known breast cancer genes, 22 known cancer genes (blue dots) and another 12 genes (red dots) which were not known as breast cancer genes were highly involved (participated in over 20 cancer gene modules). Gray dots donated known cancer genes which were not highly involved in cancer gene modules. **b** An example for illustrating gene representation features (RFs). For a given gene, if a feature is assigned with an importance score (calculated by GNNExplainer) 10 times higher than the minimum score, it will be referred to as the RF of this gene. **c** Distribution of genes with different numbers of RFs. **d** Illustrations of *BRCA1* and *BRCA2* gene modules. **e** *BRCA1* and *BRCA2* gene modules formed a high-order gene module through three shared genes including *TP53*, *SMAD3*, and *XPO1*.

Figure S4

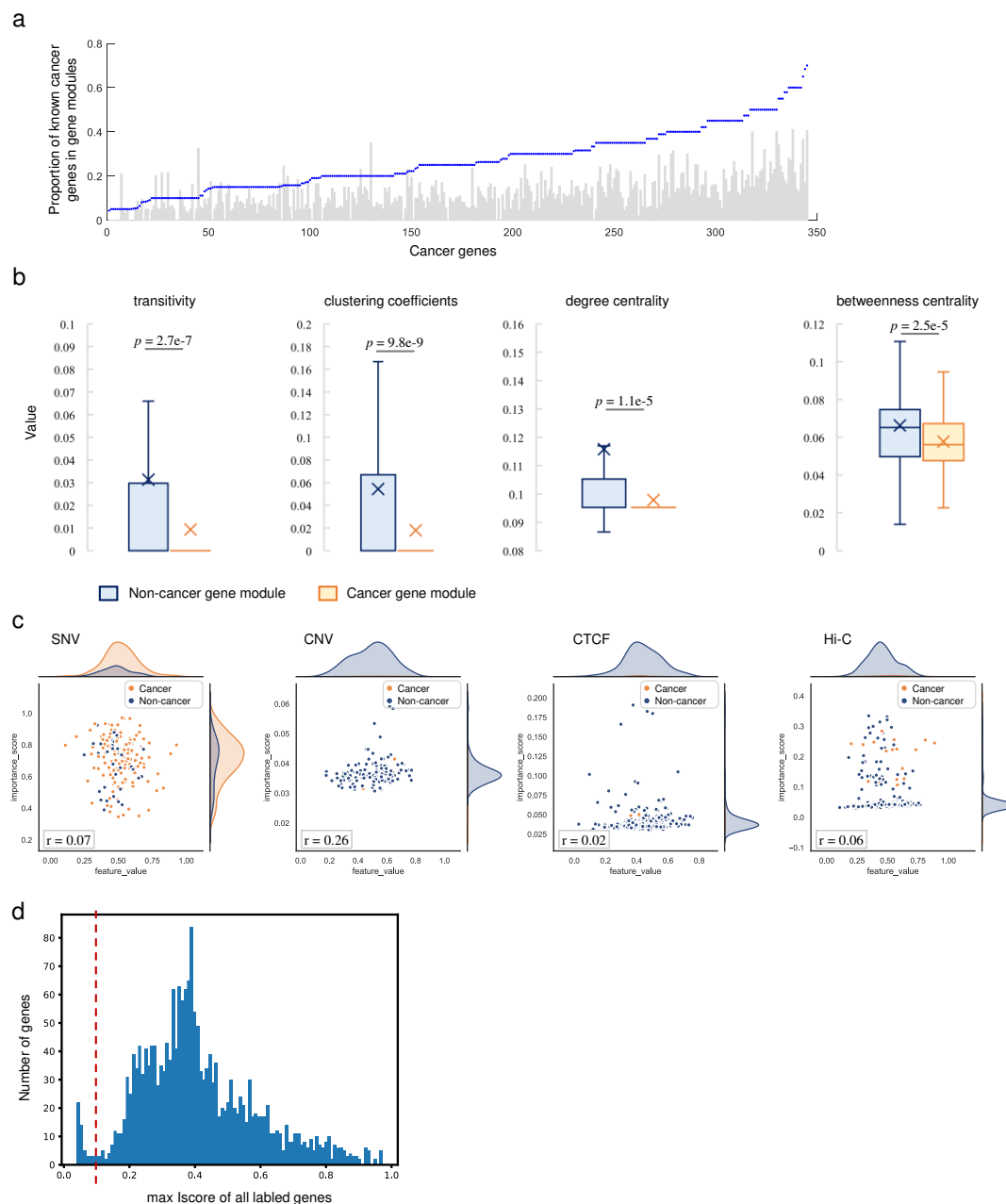


Fig. S4 Gene modules in breast cancer cell line. **a** For each cancer gene module, the proportion of known cancer genes was calculated (blue dots). Random gene modules with the same number of genes was shuffled 100 times, and the average proportion of known cancer genes in the module was calculated (gray bars). **b** Graphical metrics of gene modules, p values were calculated by paired t-test. **c** Scatter plots show the distributions of raw feature inputs and importance scores calculated by GNNExplainer. **d** Distribution of the highest importance score of genes showed that only dozens of genes with importance score lower than 0.1 (red line).

The complex gene module formed by ErbB family

The ErbB family, including *ERBB1* (also known as *EGFR*), *ERBB2* (also known

as *HER2*), *ERBB3* (also known as *HER3*) and *ERBB4*, plays a central role in the tumorigenesis of many types of solid tumor. The members of the ERBB family are receptor tyrosine kinases (RTKs), which have an analogous structure⁴⁵. However, their gene modules exhibit heterologous structures and none of these four ErbB genes were hubs in their gene modules (Fig. 5a). In the *ERBB1* gene module, *PTPRD* located at the center and connected *ERBB1* and other genes. The *ERBB2* and *ERBB4* gene modules shared the same center gene *DLG2*. *NRG1* located at the center of the *ERBB3* gene module and *ERBB4* was also present in this gene module. We examined the representative features of the ErbB family. Hi-C and SNV were major RFs for *ERBB2*, *ERBB3*, and *ERBB4*. The mechanisms of genetic alteration such as SNV in cancer development have been demonstrated previously⁴⁶⁻⁴⁸. The Hi-C features uncovered by CGMega suggest new insights regarding the signal and crosstalk between the ErbB family genes in the context of the chromatin structure in tumor progression.

In spite of the difference among the gene modules of the ErbB family, we observed several shared genes connecting the ErbB members and forming a complex module (Fig. 5b). *NRG1*, *PPM1A*, and *DLG2* were key connectors in this high-order module. Previous studies have demonstrated the importance of these three genes for cancer development. *NRG1* is a main physiological ligand to ErbB family and, together with *ERBB2* and *ERBB3*, can form a potent pro-oncogenic heterocomplex⁴⁹. *DLG2* is a member of a family of membrane-associated guanylate kinase (MAGUK), and *DLG2* over expression will affect the level of protein phosphorylation^{50,51}. The protein serine/threonine phosphatase *PPM1A* is a crucial regulator of cell cycle progression in triple-negative breast cancer⁵², and *PPM1A* is also an important factor in protein dephosphorylation⁵³⁻⁵⁵. By combining these isolated evidences with the high-order gene module, we proposed a hypothetical model of the gene module in maintaining protein phosphorylation homeostasis (Fig. 5c). The *NRG1* ligand binds to homo- or hetero-dimers of ErbB proteins, leading to the activation of ErbB-mediated downstream signaling pathways that mediate the activity of serine/threonine (Ser/Thr) protein kinases. Ser/Thr protein kinases and proteins encoded by *DLG2* modulate the phosphorylation of Ser/Thr proteins, while *PPM1A* mediates their dephosphorylation, together maintaining protein phosphorylation homeostasis.

Figure 5

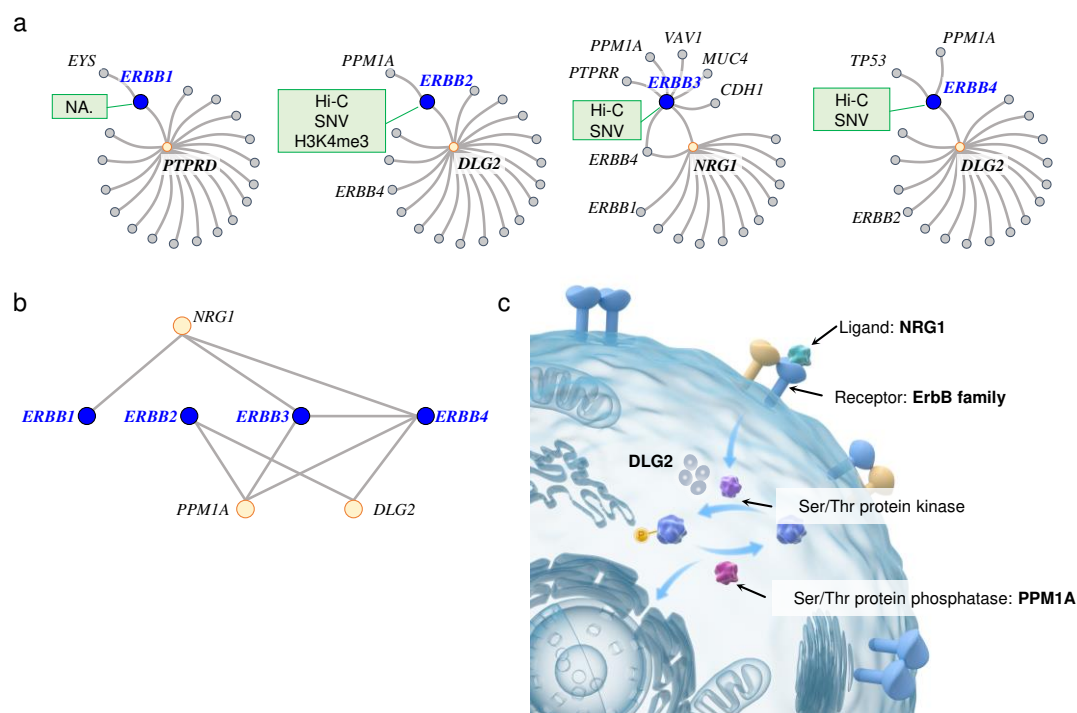


Fig. 5 Gene modules of ErbB family. **a** Illustrations show the gene modules of the ErbB family. Blue dots indicate query genes, namely *ERBB1*, *ERBB2*, *ERBB3* and *ERBB4*. Yellow dots indicate genes located at the center of gene modules. Green boxes show the RFs of query genes. **b** The high-order gene module formed by the ErbB family gene modules. **c** Hypothetical model of the high-order gene module formed by the ErbB family gene modules in maintaining protein phosphorylation homeostasis. Details were described in the main text.

Gene module dissection in acute myeloid leukemia patients

We applied CGMega to acute myeloid leukemia (AML), a myeloid neoplasm that is characterized by differentiation blockade and clonal proliferation of abnormal myeloblasts in the bone marrow⁵⁶. We collected multi-omics data for eight AML patients from a previous study⁵⁶. Unlike the case of those cell lines, gene modules are heterogeneous across different patients⁵⁷, and the clinical course of AML is also highly heterogeneous⁵⁸. Thus, we studied both patient-common and patient-specific cancer gene modules (Fig. 6a). First, we used CGMega to predict cancer genes and identified 2746 new genes in total (Supplementary Data 2). Among these, 396 were predicted to be cancer genes in all AML patients (referred as “candidate AML genes”, Supplementary Data 3). We next investigated gene functions and found that these candidate AML genes contained many essential genes and TFs, and the pan-cancer genes were significantly enriched in these 396 genes ($p = 1.32e-22$, hypergeometric test) (Fig. 6b). Moreover, Gene Ontology (GO) analysis showed that candidate AML genes

397 together with known AML genes participated in 15 hematopoietic and blood diseases
398 biology processes such as leukocyte migration and T cell receptor signaling pathway
399 (Fig. 6c). This enrichment could not be retrieved using known AML genes alone.

400 We then examined the AML gene modules. As with MCF7 cell line, cancer genes
401 were also enriched in same AML gene module. This enrichment was observed not only
402 in known AML genes but also in candidate AML genes (Fig. S5a). In addition, 10.5%
403 of these pairwise relationships in cancer gene modules were conserved over half of total
404 patients. For example, in the *DLX4* gene module, connections among *DLX4*, the known
405 cancer gene *ABL1*, and four candidate AML genes (*SPI1*, *FYN*, *GRB2*, and *SMAD2*) co-
406 occurred in multiple patients (Fig. 6d). Beyond the enrichment and co-occurrence of
407 AML gene modules, we observed that some candidate AML genes were shared by
408 dozens of known AML gene modules (Supplementary Data 4). For example, *ESR1* was
409 predicted to be candidate AML gene and it existed in modules of various known AML
410 genes, such as *EGFR*, *PIK3CA*, and *FOS* (Fig. S5b). This hub location implies a high-
411 order pattern of cancer gene modules.

412 The AML patients used in this study come from seven different mutation types and
413 CGMega achieved good performance (AUPRC = 0.8528 on average, Fig. S5c). We next
414 identified patient-specific candidate AML genes for each patient (Supplementary Data
415 5). Examining the modules of these genes, we also observed patient-specific patterns.
416 For example, in the *KLF4* gene module drawn from patient 168, the candidate AML
417 gene *SETD7* connected *KLF4* with other known AML genes including *TP53*, *STAT3*,
418 *DNMT1*, *PCNA*, and *MDM2*. However, this two-hop gene module did not appear in
419 other patients (Fig. 6e), suggesting that the dissection of cancer gene modules can help
420 provide a high-order perspective for understanding cancer heterogeneity.

Figure 6

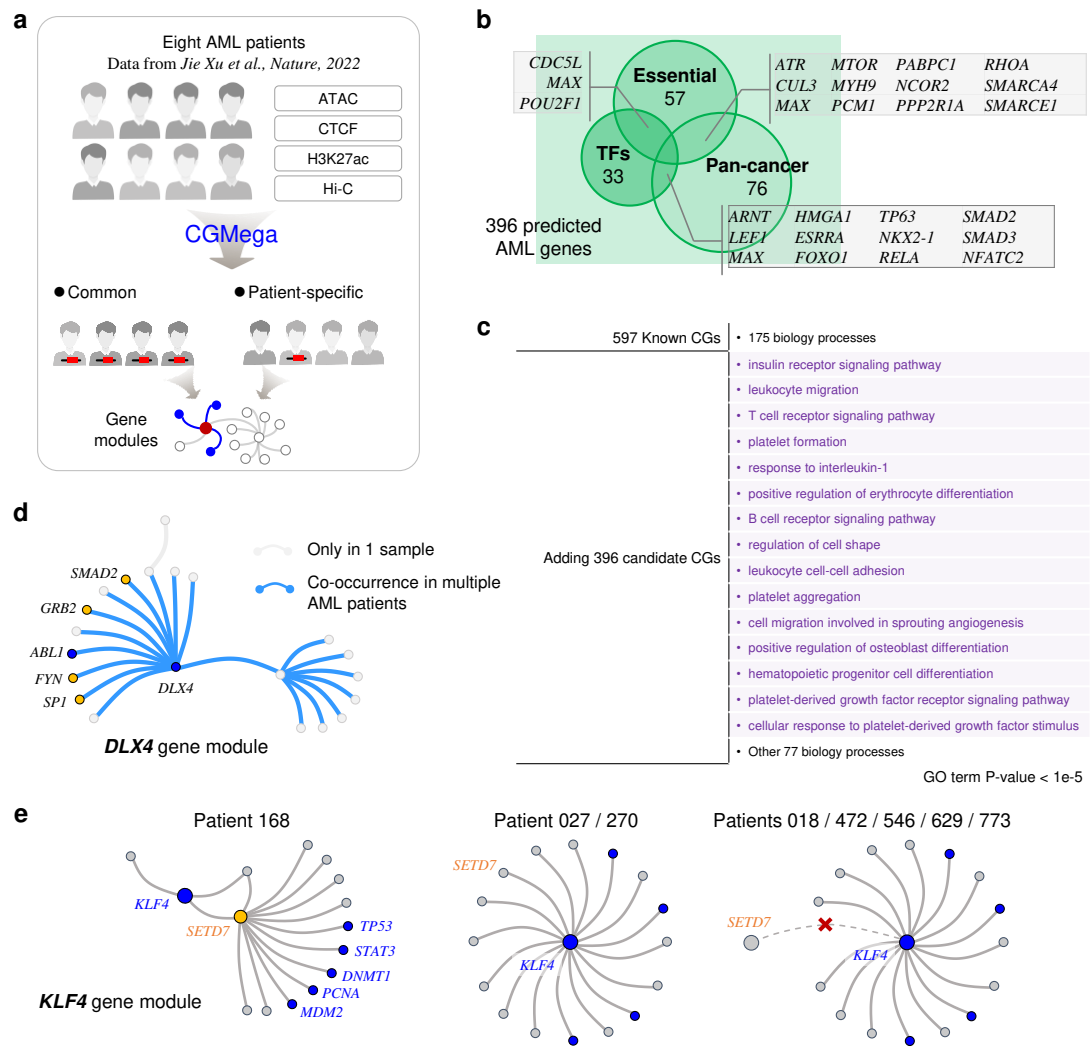


Fig. 6 Gene modules in AML patients. **a** Application of CGMega on AML. Multi-omics data of eight AML patients were obtained from a previous study. **b** 396 candidate AML genes contained essential genes, transcription factors, and pan-cancer genes. Gray boxes show genes in two categories. **c** Gene ontology (GO) enrichments. We performed GO analysis on 597 known AML genes (first line), and on 993 genes (597 known AML genes and 396 candidate AML genes), respectively. GO analysis was conducted using DAVID and GO terms with p value lower than 1e-5 were shown. **d** Illustration of *DLX4* gene module. Blue dots indicate known AML genes, while yellow dots indicate candidate AML genes. **e** Illustration of *KLF4* gene modules in separate patients. *SETD7* was located at the center in Patient 168 while it was just a participant in Patient 027 and Patient 270. In other patients, *SETD7* did not appear in *KLF4* gene modules.

Figure S5

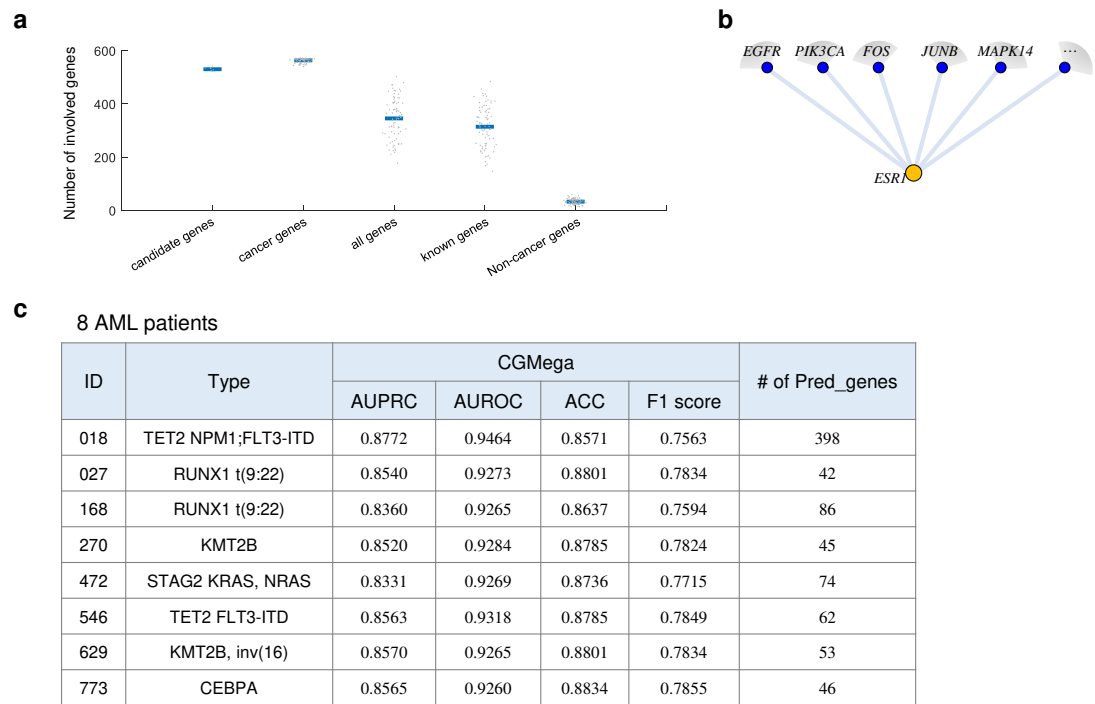


Fig. S5 Gene modules in AML patients. **a** Application. **b** illustration of *ESR1* in various cancer gene modules. Gray background indicates other parts of gene modules expect *ESR1*. **c** Details of CGMega application on each AML patient. Mutation types were obtained from a previous study.

Discussion

With the recent accumulation of multi-omics cancer data, different phenotypic manifestations of cancer hallmarks have seen intensive exploration⁵⁹. However, beyond these data, the molecular mechanisms of cancer are far from transparent. The challenge of finding candidate drivers is considerable: tumors are heterogeneous, the data are noisy and highly correlated, and there are a large number of possible combinations of drivers and genes in modules⁶⁰.

Therefore, in this work, we introduced a general framework, CGMega, for predicting cancer genes and dissecting gene modules. CGMega differs from other methods in three main ways: (1) It leverages a GAT-based neural network across multiple molecular features, rather than single molecular feature, such as gene co-expression (extensively evaluated elsewhere⁶¹). Compared to EMOGI²⁵, a different multi-omics integration method, for cancer gene prediction, CGMega is advanced in its ability to capture the 3D genome architecture, which has been widely demonstrated as a new perspective for the study of cancer⁶². Multi-omics data integration is effective

for avoiding noisy and sparse data, and neural networks are sufficiently powerful to capture the intrinsic relationships among complex data. (2) CGMega utilizes the model-agnostic approach GNNExplainer⁴⁰ to interpret contributing factors to cancer gene prediction. This approach provides a comprehensive evaluation for all gene neighbors (genes that connect with predicted gene via PPIs) and molecular features (including SNVs, CNVs, histone modifications, chromatin accessibility, and 3D genome architectures) in a masking-based manner. This is critical for detecting complex and high-order gene modules. For example, *TP53* is a two-hop neighbor in the *KLF4* gene module and *SETD7* serves as the hub connecting *KLF4* and other genes (Fig. 6e). (3) CGMega supports pre-training and fine-tuning, which helps improve the prediction performance especially for rare cancer studies. Together, as a result, CGMega has improved performance in cancer gene prediction and cancer gene module dissection.

In addition to these advantages of CGMega, we also provided a comprehensive evaluation of approaches for Hi-C data integration with other omics data, and we demonstrated that (1) the graph structure is advanced in integrating multi-omics information, especially for molecular signals and gene relationships combination, and (2) using SVD to encode Hi-C data as gene features is better than measuring Hi-C data as gene linkages. This may be due to the high sparsity and noise of Hi-C data.

Incorporating multi-omics information enables us to consider any class of genes in a functional module and to relate the malignant phenotype to the molecular features from which it is likely to have originated. The application of CGMega on the breast cancer cell line and AML patients helped uncover that (1) cancer gene modules are widespread and well-organized including cancer gene-centered patterns (such as the *BRCA1* module in Fig. 4d) and non-cancer gene-centered patterns (such as modules of the ErbB family in Fig. 5a). (2) cancer genes (both known cancer genes and predicted ones) tend to be enriched in a single module (Figs. S4a and S6a), suggesting the complex interplay of cancer genes in tumorigenesis. (3) beyond these well-known cancer genes, there are hub genes that are either located at the center in cancer gene modules (such as *SETD7* in the *KLF4* gene module, Fig. 6e) or are present in dozens of cancer gene modules (such as *ESR1* present in 95 cancer gene modules, Fig. S6a). Not only can this help distinguish between driver and passenger genes, but the genes in the associated module can also provide insight into the role of the driver. In addition, the good performance of CGMega in the breast cell line (AUPRC = 0.9140) and AML patients (average AUPRC = 0.8528) demonstrates that (1) CGMega exhibits efficacy

for both cell line and donor samples, as well as for both solid tumor and liquid tumor studies. (2) CGMega is flexible for inputs with few missing molecular features (for example, a lack of SNV and CNV information in AML application), suggesting that our framework may be applicable to other similar tasks.

Methods

Data collection and preprocessing. In this study, we collected well-known cancer genes and obtained mutations (SNVs), CNVs, chromatin accessibility (ATAC-seq), histone modification (H3K4me3 and H3K27ac), CTCF ChIP-seq, PPIs and 3D chromatin (Hi-C) data from public databases. The details are as follows:

Collection of positive and negative training samples. The collection of positive and negative samples was performed as previously reported²⁵. In brief, known cancer genes were extracted from the Network of Cancer Genes (NCG)⁶³, Cancer Gene Census (CGC)⁶⁴ along with high-confidence (level ≥ 0.95) cancer genes using DigSEE⁶⁵. Negative samples were selected using a rigorous exclusionary process, considering all coding genes that were not marked as positive and removing genes labeled as cancer-related or cancer candidates in NCG and KEGG (<https://www.kegg.jp/kegg/kegg1.html>), along with all genes from the OMIM database (<https://omim.org/>). Additionally, genes predicted to be involved in cancer-related expressions and those with regulatory effects on oncogenes by MSigdb⁶⁶ were excluded. The remaining genes were completely unrelated to cancer (where possible pan-oncogenes were also excluded) and were negative samples. The final dataset consisted of positive samples and negative samples obtained through these steps. Finally, 358 positive and 1581 negative samples were collected for breast cancer cell line MCF7, and 598 positive and 1838 negative samples were collected for leukemia cell line K562 and AML patients.

Omics data. The SNVs and CNVs information were retrieved from The Cancer Genome Atlas (TCGA) project using the corresponding data generated from the closest tumour types of breast cancer and myeloid leukemia. As for SNVs, we first filtered out the silent somatic mutations and then we calculated the mutation frequency for a gene in a sample by dividing the number of single nucleotide mutations of this gene by its genomic length. We next averaged the mutation frequency of each gene over all samples. The mean value of the obtained gene mutation frequency was magnified by 1000 times as a dimension of gene features. As for CNVs, we used the number of copy

numbers of all samples of the cumulative gene and divided them by the total number of samples as the average CNV value as a dimension of gene features. For chromatin accessibility, histone modification and transcription factor binding, we downloaded ATAC-seq, H3K4me3, H3K27ac, and CTCF ChIP-seq data from the ENCODE project⁶⁷ for MCF7 and K562 cell lines, and from Gene Expression Omnibus (GEO) with accession number GSE152136 for AML patients. We then calculated the signal densities at gene promoter (TSS $\pm$ 1 kb) for each marker as gene features. Human genome annotation (GRCh38) was used in this study.

Hi-C data. We collected Hi-C data from GEO with accession number GSE66733 for the MCF7 cell line, GSE63525 for the K562 cell line, and GSE152136 for AML patients. To remove the complex structural variants and improve accuracy, we first performed NeoLoopFinder⁶⁸ on raw Hi-C contact maps to reconstruct local Hi-C data. Then, we normalized Hi-C data using iterative correction and eigenvector decomposition (ICE)⁴¹. Finally, we decomposed normalized Hi-C data into five dimensions as gene features using SVD.

Protein-protein interaction network. We collected the PPI network from the ConsensusPathDB (CPDB)⁶⁹ database, which incorporates the interaction information from 32 widely used and important biological databases. We removed the interactions with scores (indicating experimental evidence of the interaction) lower than 0.5 and finally included approximately 270,000 interactions as graph edges. To validate the suitability of CGMega on other PPIs (Fig. S3a), we also tested CGMega on other four PPIs including IRef, Multinet, PCNet, and STRING. These PPIs were downloaded and pre-processed as previously described²⁵.

Dataset organization. According to above data collection and preprocessing, we removed genes with no corresponding information and finally obtained 16,165 protein coding genes (Supplementary Data 6). Then we constructed the multi-omics information combination graph, an undirected graph, in which nodes represent genes and edges are obtained from CPDB PPI network. For the MCF7 and K562 cell lines, the node features include a total of 15 dimensions including ATAC-seq (1), CTCF (3), H3K4me3 (2), H3K27ac (2), CNV (1), SNV (1), and Hi-C (5). For AML patients, it is unavailable to get patient genomic information (SNV and CNV), and node features comprise a total of eight dimensions including ATAC (1), CTCF (1), H3K27ac (1), and Hi-C (5). For experiments on different PPI networks, graph edges were alternatively obtained from IRef, Multinet, PCNet, and STRING. In this study, CGMega predicts

cancer genes in a semi-supervised manner. Genes were labeled as positive, negative, and unknown according to the above section “*Collection of positive and negative training samples*” section. To conduct evaluation, 25% of the positive and negative genes were assigned to the test set while the remaining 75% were divided into 10 equal parts. Each part was consecutively used as a validation set in the 10-fold cross-validation process. The combination graph for multi-omics information is constructed and then converted into a graphical structure dataset using PyTorch Geometric ⁷⁰ for further processing and computation.

GAT-based cancer gene prediction model. The model consists of two graph transformer layers, two fully connected linear layers, two layer-norm layers, one max-pooling layer, and one dropout layer. The two graph transformer layers are used to perform GAT mechanism, as is discussed in following section. The output of each graph transformer layer is passed through a corresponding layer-norm layer to normalize the activation:

$$x'_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i) + \epsilon}} \odot \gamma + \beta,$$

where x_i denotes the embedding of node i , ϵ denotes a small constant, and γ and β denote learnable affine parameters.

The max pooling layer is used to reduce the dimensionality of the features extracted by using the graph transformer layers. The two fully connected linear layers are used to make the final prediction, with the first linear layer transforming the data into a 32-dimensional representation and the second linear layer producing the final output. The dropout layer is used to prevent overfitting.

Graph transformer layer. CGMega employs two graph transformer layers ⁷¹ to enable graph message passing. In the graph $G = (V, E)$, the nodes are associated with feature matrix $X \in \mathbb{R}^{n \times m}$. In a multi-layer graph neural network, let the set of outputs of the l^{th} hidden layer be $H^{(l)} = \{h_1^{(l)}, h_2^{(l)}, \dots, h_m^{(l)}\}$, where $h_i^{(l)}$ denotes the i^{th} node's representation in the l^{th} layer. For the a^{th} head attention, the attention score of an edge from j to i ($\alpha_{a,ij}^{(l)}$) can be calculated as follows:

$$\alpha_{a,ij}^{(l)} = \frac{\langle q_{a,i}^{(l)}, k_{a,j}^{(l)} + e_{a,ij} \rangle}{\sum_{u \in \mathcal{N}(i)} \langle q_{a,i}^{(l)}, k_{a,u}^{(l)} + e_{a,iu} \rangle},$$

$$q_{a,i}^{(l)} = W_{a,q}^{(l)} \mathbf{h}_i^{(l)} + b_{a,q}^{(l)},$$

$$k_{a,j}^{(l)} = W_{a,k}^{(l)} \mathbf{h}_j^{(l)} + b_{a,k}^{(l)},$$

$$e_{a,ij} = W_{a,e} \mathbf{e}_{ij} + b_{a,e},$$

where e_{ij} denotes the edge feature; $W_{a,q}^{(l)}$, $W_{a,k}^{(l)}$, $W_{a,e}$, $b_{a,q}^{(l)}$, $b_{a,k}^{(l)}$, and $b_{a,e}$ denote different learnable parameter matrices; $\langle q, k \rangle = \exp\left(\frac{q^\top k}{\sqrt{d}}\right)$ is the exponential scale dot-product function; and d is the hidden size of each head.

For the a^{th} head attention of the l^{th} hidden layer, the source feature $\mathbf{h}_i^{(l)}$ and distant feature $\mathbf{h}_j^{(l)}$ are first transformed into the query vector $q_{a,i}^{(l)} \in \mathbb{R}^d$ and key vector $k_{a,j}^{(l)} \in \mathbb{R}^d$. Then the encoded edge feature $e_{a,ij} \in \mathbb{R}^d$ is added to the key vector for each layer.

After obtaining the graph multi-head attention, a message aggregation is made from j to i :

$$\hat{\mathbf{h}}_i^{(l+1)} = \frac{1}{A} \sum_{a=1}^A \sum_{j \in \mathcal{N}(i)} \alpha_{a,ij}^{(l)} (v_{a,j}^{(l)} + e_{a,ij}),$$

$$v_{a,j}^{(l)} = W_{a,v}^{(l)} \mathbf{h}_j^{(l)} + b_{a,v}^{(l)},$$

where A denotes the number of heads. According to experiments in the field of NLP⁷², averaging the attention head vectors has been shown to have a minimal impact on results relative to stitching them together. Moreover, this approach does not significantly affect the scale of the hidden layer vectors nor reduce computational overhead.

To mitigate the issue of over-smoothing that often occurs in GNNs, CGMega employs a gated residual connection between the graph transformer layers. The final output of each layer can be expressed as follows:

$$\mathbf{r}_i^{(l)} = W_{\mathbf{r}}^{(l)} \mathbf{h}_i^{(l)} + b_{\mathbf{r}}^{(l)},$$

$$\beta_i^{(l)} = \text{sigmoid}\left(W_g^{(l)} \left[\hat{\mathbf{h}}_i^{(l+1)}; \mathbf{r}_i^{(l)}; \hat{\mathbf{h}}_i^{(l+1)} - \mathbf{r}_i^{(l)}\right]\right),$$

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}\left(\text{LayerNorm}\left(\left(1 - \beta_i^{(l)}\right) \hat{\mathbf{h}}_i^{(l+1)} + \beta_i^{(l)} \mathbf{r}_i^{(l)}\right)\right).$$

608 To further mitigate the over-smoothing and effectively leverage the node features,
 609 CGMega also includes a residual connection for the nodes' initial features. The
 610 resulting representation learning model can be represented using the following equation:

$$\begin{aligned} 611 \quad \mathbf{h}_i^{(l+1)} &= \text{GraphTransformerConv}(G, \mathbf{X})^{(l)} \\ 612 \quad &= \text{ReLU} \left(\text{LayerNorm} \left(\left((1 - \beta_i^{(l)}) \hat{\mathbf{h}}_i^{(l+1)} + \beta_i^{(l)} \mathbf{r}_i^{(l)} + x_i \right) \right) \right). \end{aligned}$$

613 Especially, for the output layer L :

$$614 \quad \mathbf{h}_i^{(L)} = (1 - \beta_i^{(L-1)}) \hat{\mathbf{h}}_i^{(L)} + \beta_i^{(L-1)} \mathbf{r}_i^{(L-1)}.$$

615 *Training settings.* In the experiments, a warm-up strategy for the learning rate was
 616 employed to improve stability and avoid falling into local minima during the initial
 617 training phase. The learning rate increased linearly from 0 to 0.005 for the first 20% of
 618 the total iterations. To prevent overfitting and over smoothing, an early stop strategy
 619 was implemented that would stop the training process if the model's performance on
 620 the validation set did not improve over a consecutive 100 epochs. This strategy helps
 621 prevent the waste of computational resources and ensures that the model can achieve
 622 its optimal performance within a reasonable number of training iterations. To improve
 623 model robustness and reduce overfitting, the dropout rate was set to 0.1 for the attention
 624 mechanism and 0.4 for the other modules. The max pooling step size was set to 2,
 625 resulting in a 32-dimensional representation. To reduce the number of parameters and
 626 ensure that training is feasible within the time and resource constraints, input graphs
 627 were sampled using neighbor sampler⁷³. The subgraphs included all first and second
 628 order neighbors for each node and training was performed on these subgraphs. Because
 629 the positive and negative samples in the dataset were highly imbalanced, especially in
 630 the breast cancer dataset, which had an approximate ratio of 1:4, 50% of the negative
 631 samples were removed from training to prevent bias towards negative samples. These
 632 hyper-parameters were determined via grid search.

633 **Gene module interpretation.** We implemented GNNExplainer to interpret the
 634 important gene features and interactions for cancer gene prediction. Given node i ,
 635 GNNExplainer identified a subgraph $G_{Si} \subseteq G$ that contains the relevant features X_{Si}
 636 that are crucial for the predicting $\hat{y}_i$. G_{Si} is a connected subgraph for which the gene
 637 nodes cover at maximum a two-hop region with no more than 20 edges. The task can
 638 be formulated with the following optimization framework:

$$639 \quad \max_{G_S} \mathbf{MI}(\mathbf{Y}, (G_S, X_S)) = \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{Y} | G = G_S, X = X_S).$$

640 A computationally efficient version is as follows:

$$641 \quad \min_{M_i, f_i} -\log P_{\Phi} (Y = \hat{y}_i | G = A_i \odot \sigma(M_i), X = X_i \odot \sigma(f_i)),$$

642 where A_i and X_i denote the adjacency matrix and node features, $M_i \in \mathbb{R}^{n \times n}$ and
643 $f_i \in \mathbb{R}^m$ denote the mask matrix or vector for adjacency matrix and node features
644 needed to be learned for i , n is the number of nodes, m is the dimension of node
645 feature, Φ denotes the trained GNN model, $\odot$ denotes element-wise multiplication
646 and σ denotes the sigmoid that maps the masks to $[0,1]$.

647 For undirected graphs such as PPI in this work, a symmetric M_i is maintained during
648 optimization. The values for M_i and f_i represent the importance of corresponding
649 edges and features, and the explanation G_{Si} and X_{Si} for the prediction $\hat{y}_i$ at node i
650 are computed as follows:

$$651 \quad G_{Si} = A_i \odot 1\{M_i \geq \text{edge threshold}\},$$

$$652 \quad X_{Si} = \mathbf{x}_{j'} | j' \in G_{Si},$$

653 where $x_j' = x_j \odot 1\{f_i \geq \text{feature threshold}\}$.

654 **Methods comparison.** We compared the performance of CGMega with other relevant
655 methods. Here, we provided detailed descriptions of the model architectures and
656 hyperparameter settings for these methods. MTGCN and EMOGI maintained the same
657 model architectures as described in their original papers. For MTGCN, the dimensions
658 of the two hidden layers were set to 300 and 100, respectively, with a linear layer
659 dimension of 100. The dropout rate was set to 0.5. As for EMOGI, the dimensions of
660 the two hidden layers were also 300 and 100, with a dropout rate of 0.5. The loss
661 multiplier was set to 45.0, and the weight decay was set to 0.0005. Both GCN and GAT
662 adopted a two-layer architecture, with parameters kept consistent with CGMega. The
663 MLP consisted of two linear layers, with the hidden layer and dropout settings matching
664 those of CGMega. The optimal combination of SVM was determined through grid
665 search. The polynomial kernel was selected, with C set to 1.0 and gamma set to scale.

666 **Representative features calculation.** According to the feature importance scores
667 calculated by GNNExplainer, we defined representative features (RFs) for each gene
668 as features that have relatively prominent importance scores. In detail, for a given gene,
669 among its features from ATAC, CTCF, H3K4me3 and H3K27ac, SNVs, CNVs and Hi-
670 C, if a feature is assigned with importance score as 10 times higher than the minimum
671 score, it is referred to as the RF for this gene.

Identification of candidate AML genes. For each patient, unknown genes with prediction scores ranking in the top 10% were reserved as the predicted cancer genes. Then we examined all of the predicted cancer genes across the eight patients and defined the candidate AML genes as those genes predicted as cancer genes for every patient. The corresponding modules for candidate AML genes were calculated by GNNExplainer in a manner similar to the sections of interpretation mentioned above.

Data availability

In this study, we collected data from public databases. ATAC-seq, H3K4me3, H3K27ac, and CTCF ChIP-seq data were obtained from the ENCODE project for MCF7 and K562 cell lines, and from Gene Expression Omnibus (GEO) with accession number GSE152136 for AML patients. PPI network was obtained from the ConsensusPathDB (CPDB) database. Hi-C data were obtained from GEO with accession number GSE66733 for the MCF7 cell line, GSE63525 for the K562 cell line, and GSE152136 for AML patients.

Code availability

Source codes for running CGMega are available on Github (<https://github.com/NBStarry/CGMega>).

Acknowledgements

This work was supported by the Beijing Natural Science Foundation (<http://kw.beijing.gov.cn/>; no.5232025 to HL), Independent Research Project of Medical Engineering Laboratory of Chinese PLA General Hospital, the National Natural Science Foundation of China (<http://www.nsfc.gov.cn>; nos. 61972251 and 62272300 to Y.Y, nos. 31801112 and 61873276 to HC and XB, respectively), and the Beijing Nova Program of Science and Technology (<https://mis.kw.beijing.gov.cn>; no. Z191100001119064 to HC).

Author contributions

X.B., Y.Y, H.C., and H.L. conceived the idea and H.L. designed the study. Y.S. and P.H. performed the acquisition, integration of datasets. Z.H. and F.W. developed and adapted the GAT-based cancer gene prediction and gene module detection model. Y.S., and P.H. performed model application and biological discovery. Y.S. and Z.H. implemented the Github deployment. C.R. and X.X. performed data preprocessing. H.L.

wrote the manuscript. All authors provided critical feedback and discussion, assisted in method development and application.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Data 1: Genes with known label for MCF7 cell line.

Supplementary Data 2: New predicted cancer genes for AML.

Supplementary Data 3: 396 candidate AML genes.

Supplementary Data 4: Candidate AML genes that are present in the module of other known AML cancer genes.

Supplementary Data 5: Patient-specific candidate AML genes and their relations with known AML genes.

Supplementary Data 6: 16,165 protein coding genes.

References

- 1 Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-176, doi:10.1038/ng1165 (2003).
- 2 Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742-751, doi:10.1038/s41586-022-05688-9 (2023).
- 3 Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461-468, doi:10.1038/nature11981 (2013).
- 4 Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663-1677, doi:10.1016/j.cell.2015.11.013 (2015).
- 5 Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat Methods* **12**, 211-214, 213 p following 214, doi:10.1038/nmeth.3249 (2015).
- 6 Chaussabel, D. *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150-164, doi:10.1016/j.immuni.2008.05.012 (2008).
- 7 Alsina, L. *et al.* A narrow repertoire of transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-function mutations in MYD88 or IRAK4. *Nat Immunol* **15**, 1134-1142, doi:10.1038/ni.3028 (2014).
- 8 Sadegh, S. *et al.* Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat Commun* **12**, 6848, doi:10.1038/s41467-021-27138-2 (2021).
- 9 Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**, 1090-1098, doi:10.1038/ng1434 (2004).
- 10 Wouters, J. *et al.* Robust gene expression programs underlie recurrent cell states and

- phenotype switching in melanoma. *Nat Cell Biol* **22**, 986-998, doi:10.1038/s41556-020-0547-3 (2020).
- 11 Goh, K. I. *et al.* The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).
- 12 Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601, doi:10.1126/science.1257601 (2015).
- 13 Theodoris, C. V. *et al.* Network-based screen in iPSC-derived cells reveals therapeutic candidate for heart valve disease. *Science* **371**, doi:10.1126/science.abd0724 (2021).
- 14 Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616-624, doi:10.1038/s41586-023-06139-9 (2023).
- 15 Li, H. *et al.* Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks. *Nature Machine Intelligence* **4**, 389-400, doi:10.1038/s42256-022-00469-5 (2022).
- 16 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).
- 17 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70, doi:10.1016/s0092-8674(00)81683-9 (2000).
- 18 Singer, M. *et al.* A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell* **166**, 1500-1511.e1509, doi:10.1016/j.cell.2016.08.052 (2016).
- 19 Chihara, N. *et al.* Induction and transcriptional regulation of the co-inhibitory gene module in T cells. *Nature* **558**, 454-459, doi:10.1038/s41586-018-0206-z (2018).
- 20 Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat Commun* **11**, 856, doi:10.1038/s41467-020-14666-6 (2020).
- 21 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086, doi:10.1038/nmeth.4463 (2017).
- 22 Silverbush, D. *et al.* Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules. *Cell Syst* **8**, 456-466.e455, doi:10.1016/j.cels.2019.04.005 (2019).
- 23 Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* **46**, 10546-10562, doi:10.1093/nar/gky889 (2018).
- 24 Collier, O., Stoven, V. & Vert, J. P. LOTUS: A single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput Biol* **15**, e1007381, doi:10.1371/journal.pcbi.1007381 (2019).
- 25 Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence* **3**, 513-526, doi:10.1038/s42256-021-00325-y (2021).
- 26 Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 661-678, doi:10.1038/nrg.2016.112 (2016).
- 27 Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345-354, doi:10.1038/s41586-019-1182-7

(2019).

28 Wong, A. S., Choi, G. C. & Lu, T. K. Deciphering Combinatorial Genetics. *Annu Rev Genet* **50**, 515-538, doi:10.1146/annurev-genet-120215-034902 (2016).

29 Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, doi:10.1126/science.aaf1420 (2016).

30 Chen, H. *et al.* New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief Bioinform* **21**, 1397-1410, doi:10.1093/bib/bbz072 (2020).

31 Bell, C. C. & Gilan, O. Principles and mechanisms of non-genetic resistance in cancer. *Br J Cancer* **122**, 465-472, doi:10.1038/s41416-019-0648-6 (2020).

32 Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629-643, doi:10.1016/j.cell.2016.12.013 (2017).

33 Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-332, doi:10.1038/nrg3920 (2015).

34 Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389-403, doi:10.1038/s41576-019-0122-6 (2019).

35 Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Brief Bioinform* **22**, 1515-1530, doi:10.1093/bib/bbaa257 (2021).

36 Liu, L. *et al.* Combining sequence and network information to enhance protein-protein interaction prediction. *BMC Bioinformatics* **21**, 537, doi:10.1186/s12859-020-03896-6 (2020).

37 Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat Biotechnol* **40**, 254-261, doi:10.1038/s41587-021-01034-y (2022).

38 Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat Rev Genet* **14**, 333-346, doi:10.1038/nrg3433 (2013).

39 Sun, Y. *et al.* A graph neural network-based interpretable framework reveals a novel DNA fragility-associated chromatin structural unit. *Genome Biol* **24**, 90, doi:10.1186/s13059-023-02916-x (2023).

40 Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *Adv Neural Inf Process Syst* **32**, 9240-9251 (2019).

41 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

42 Wu, Z., Zhan, M., Zhang, H., Luo, Q. & Tang, K. MTGCN: A multi-task approach for node classification and link prediction in graph data. *Information Processing & Management* **59**, 102902, doi:<https://doi.org/10.1016/j.ipm.2022.102902> (2022).

43 Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* **12**, 68-78, doi:10.1038/nrc3181 (2011).

44 Zhang, X. *et al.* Acquired temozolomide resistance in MGMT(low) gliomas is associated with regulation of homologous recombination repair by ROCK2. *Cell Death Dis* **13**, 138, doi:10.1038/s41419-022-04590-6 (2022).

45 Tebbutt, N., Pedersen, M. W. & Johns, T. G. Targeting the ERBB family in cancer: couples therapy. *Nat Rev Cancer* **13**, 663-673, doi:10.1038/nrc3559 (2013).

- 46 Li, M. *et al.* Whole-exome and targeted gene sequencing of gallbladder carcinoma identifies recurrent mutations in the ErbB pathway. *Nat Genet* **46**, 872-876, doi:10.1038/ng.3030 (2014).
- 47 Morgensztern, D., Politi, K. & Herbst, R. S. EGFR Mutations in Non-Small-Cell Lung Cancer: Find, Divide, and Conquer. *JAMA Oncol* **1**, 146-148, doi:10.1001/jamaoncol.2014.278 (2015).
- 48 Yan, T. *et al.* Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. *Nat Commun* **10**, 1670, doi:10.1038/s41467-019-09255-1 (2019).
- 49 Diwanji, D. *et al.* Structures of the HER2-HER3-NRG1 β complex reveal a dynamic dimer interface. *Nature* **600**, 339-343, doi:10.1038/s41586-021-04084-z (2021).
- 50 Keane, S., Herring, M., Rolny, P., Wettergren, Y. & Ejeskär, K. Inflammation suppresses DLG2 expression decreasing inflammasome formation. *Journal of Cancer Research and Clinical Oncology* **148**, 2295-2311, doi:10.1007/s00432-022-04029-7 (2022).
- 51 Keane, S., de Weerd, H. A. & Ejeskär, K. DLG2 impairs dsDNA break repair and maintains genome integrity in neuroblastoma. *DNA Repair* **112**, 103302, doi:<https://doi.org/10.1016/j.dnarep.2022.103302> (2022).
- 52 Mazumdar, A. *et al.* The phosphatase PPM1A inhibits triple negative breast cancer growth by blocking cell cycle progression. *NPJ Breast Cancer* **5**, 22, doi:10.1038/s41523-019-0118-6 (2019).
- 53 Takekawa, M., Maeda, T. & Saito, H. Protein phosphatase 2C α inhibits the human stress-responsive p38 and JNK MAPK pathways. *Embo j* **17**, 4744-4752, doi:10.1093/emboj/17.16.4744 (1998).
- 54 Cheng, A., Ross, K. E., Kaldis, P. & Solomon, M. J. Dephosphorylation of cyclin-dependent kinases by type 2C protein phosphatases. *Genes Dev* **13**, 2946-2957, doi:10.1101/gad.13.22.2946 (1999).
- 55 Dai, F., Shen, T., Li, Z., Lin, X. & Feng, X. H. PPM1A dephosphorylates RanBP3 to enable efficient nuclear export of Smad2 and Smad3. *EMBO Rep* **12**, 1175-1181, doi:10.1038/embo.2011.174 (2011).
- 56 Xu, J. *et al.* Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature* **611**, 387-398, doi:10.1038/s41586-022-05365-x (2022).
- 57 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).
- 58 Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221, doi:10.1056/NEJMoa1516192 (2016).
- 59 Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31-46, doi:10.1158/2159-8290.Cd-21-1059 (2022).
- 60 Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005-1017, doi:10.1016/j.cell.2010.11.013 (2010).
- 61 Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* **9**, 1090, doi:10.1038/s41467-018-03424-4 (2018).

- 62 Deng, S., Feng, Y. & Pauklin, S. 3D chromatin architecture and transcription regulation in cancer. *J Hematol Oncol* **15**, 49, doi:10.1186/s13045-022-01271-x (2022).
- 63 Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**, 1, doi:10.1186/s13059-018-1612-0 (2019).
- 64 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).
- 65 Kim, J. *et al.* DigSee: Disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Res* **41**, W510-517, doi:10.1093/nar/gkt531 (2013).
- 66 Liberzon, A. A description of the Molecular Signatures Database (MSigDB) Web site. *Methods Mol Biol* **1150**, 153-160, doi:10.1007/978-1-4939-0512-6_9 (2014).
- 67 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).
- 68 Wang, X. *et al.* Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat Methods* **18**, 661-668, doi:10.1038/s41592-021-01164-w (2021).
- 69 Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* **39**, D712-717, doi:10.1093/nar/gkq1156 (2011).
- 70 Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. (2019).
- 71 Shi, Y. *et al.* Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. (2020).
- 72 Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. (2021).
- 73 Hamilton, W. L., Ying, R. & Leskovec, J. Inductive Representation Learning on Large Graphs. (2017).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData1.xlsx](#)
- [SupplementaryData2.xlsx](#)
- [SupplementaryData3.xlsx](#)
- [SupplementaryData4.xlsx](#)
- [SupplementaryData5.xlsx](#)
- [SupplementaryData6.xlsx](#)