metabolites

MDPI

Article

Identification of Cancer Driver Genes by Integrating Multiomics Data with Graph Neural Networks

Hongzhi Song [1,†], Chaoyi Yin [1,†], Zhuopeng Li [2], Ke Feng [1], Yangkun Cao [1], Yujie Gu [1] and Huiyan Sun [1,*]

[1] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
* Correspondence: huiyansun@jlu.edu.cn; Tel.: +86-0431-85167648
† These authors contributed equally to this work.

Abstract: Cancer is a heterogeneous disease that is driven by the accumulation of both genetic and nongenetic alterations, so integrating multiomics data and extracting effective information from them is expected to be an effective way to predict cancer driver genes. In this paper, we first generate comprehensive instructive features for each gene from genomic, epigenomic, transcriptomic levels together with protein–protein interaction (PPI)-networks-derived attributes and then propose a novel semisupervised deep graph learning framework GGraphSAGE to predict cancer driver genes according to the impact of the alterations on a biological system. When applied to eight tumor types, experimental results suggest that GGraphSAGE outperforms several state-of-the-art computational methods for driver genes identification. Moreover, it broadens our current understanding of cancer driver genes from multiomics level and identifies driver genes specific to the tumor type rather than pan-cancer. We expect GGraphSAGE to open new avenues in precision medicine and even further predict drivers for other complex diseases.

Keywords: graph neural network; multiomics data; cancer driver gene; PPI network; biomarker

Citation: Song, H.; Yin, C.; Li, Z.; Feng, K.; Cao, Y.; Gu, Y.; Sun, H. Identification of Cancer Driver Genes by Integrating Multiomics Data with Graph Neural Networks. *Metabolites* **2023**, *13*, 339. https://doi.org/10.3390/metabo13030339

Academic Editor: Rui Wang-Sattler

Received: 5 January 2023
Revised: 20 February 2023
Accepted: 22 February 2023
Published: 24 February 2023

1. Introduction

Cancer is a complex disease with abnormal cellular metabolism. Cancer driver gene alterations always influence the normal-to-cancer cell transformation process and reprogram metabolism to meet the higher demands for nutrition and growth of uncontrolled cell proliferation [1]. Hence, it is critical to accurately identify such genes as the biomarkers for facilitating precision diagnosis and therapeutics [2]. Growing evidence suggests that cancer is driven by both genetic and nongenetic alterations [3], so integrating multiomics data and extracting valuable information from them may be an effective way to predict cancer driver genes.

Cancer progression is usually thought to result from the accumulation of driver genetic mutations which confer a selective growth advantage to the cell [2]. Many existing studies have been proposed to annotate cancer driver genes from genetic mutation data by counting mutation frequency, both alteration and conversion, and performing a series of statistical analyses, such as ccpwModel and xGeneModel [4]. In bladder cancer cells, the cancer driver genes reportedly tend to have a higher conversion frequency of C->G and C->T than other conversion patterns. Besides single nucleotide polymorphisms, a copy number variation (CNV) is also a major, well-studied mutation type. CanDrA [5] is a machine learning model for predicting cancer driver genes based on a set of 95 structural and evolutionary features. However, such methods only considering mutation frequency would overlook some real cancer driver genes with low mutation rates [6]. Moreover, aggregation mutation events in some genes on chromosomes lead to different mutation probabilities per base, so it is one-sided to identify cancer driver genes based on the average mutation rate [7]. On the other hand, although genome sequences have facilitated the identification of many cancer

genes, the number of identified cancer genes specific to each tumor type is still low [8] and many genes that play important roles in tumorigenesis do not alter their DNA sequence [9].

Actually, cancer driver genes are usually dysregulated through various cellular mechanisms and signals, hence these nonmutated cancer-dependent genes, such as transcriptional and epigenetic regulators, are also of great interest. For example, DNA hypermethylation and hypomethylation at CpG islands surrounding gene promoters can inactivate tumor suppressor genes and activate oncogenes to promote tumor growth, respectively [10]. In addition to epigenetic effects, disrupting transcription factor binding sites also alters the regulation and expression of genes in several ways [11]. Therefore, it is necessary to integrate and make full use of the complementary information contained in multiomics data.

When looking at genes from the perspective of biological systems, they always work together. Existing studies have shown that cancer driver genes have the capacity to alter the gene expression of their interacting proteins or genes that share the same biochemical pathways [12], and the disruption of some driver genes may promote the progression of disease and lead to a cancer phenotype. Hence, taking advantage of the information contained in biological networks, such as PPI networks, is highly important for predicting cancer driver genes. Bashashati et al. proposed DriverNet [6] which identified driver mutated genes by analyzing their impact on the expression of genes connecting to them in the PPI network. The 20/20+ tool [9] generated a feature vector composed of mutation clustering, evolutionary conservation, functional impact of variants, type of mutation consequences, gene interaction network connectivity and other relevant factors for predicting cancer driver genes.

In combination with the valuable information contained in a graph and a deep learning model's power, graph neural networks are widely used in classifying nodes in a network and perform well [13]. In recent years, some methods for classifying pan-oncogenes based on graph neural networks and multiomics data have been proposed. For example, Schulte-Sasse et al. designed the EMOGI [8] model, which combined CNV data, DNA methylation data, single nucleotide variant (SNV) data, and gene expression data and applied a graph convolution network with PPI networks to classify pan-cancer genes. In reality, due to the heterogeneity of tumors, cancer driver genes should be specific to each cancer type. However, EMOGI cannot distinguish the driver genes of cancer types so far. Furthermore, due to the limitation of graph convolution networks (GCNs) [14] in terms of efficiency, they are difficult to apply to large networks. Moreover, a GCN incorporates noise information in the way of aggregating nodes [15]. Compared with a GCN, GraphSAGE [16] aims to learn an aggregator rather than learning a feature representation for each node. Thus, the nodes and their neighborhood can be well-distinguished to reduce the influence of false-positive protein interactions in the network [17,18]. In each layer of a GraphSAGE model, the multiomics information of each gene and its interacted nodes, such as genes acting together with it in a signaling or regulatory pathway, as well as in protein complexes, are aggregated. In a real biological system, the strength of the interactions between the nodes usually varies, while a GCN assigns the same weight to each edge in the graph, leading to its limitations in practical application. A graph attention network (GAT) [19] incorporates an attention mechanism to assign weights to the edges between nodes for better learning the graph's structural information and nodes' representation.

In this study, we propose a novel framework called GGraphSAGE which integrates multiomics data, network-derived features, and graph neural networks for identifying cancer driver genes of each cancer type (Figure 1). We first generate a new feature vector for each gene in each tumor type, which is basically composed of four categories of features including 3 transcriptomic features, 1 epigenomic feature, 26 genomic features, and 6 network-derived features. Then, we take a total of 36 features along with the PPI network to train a graph neural network model. Considering a large number of genes cannot be definitely determined as cancer driver genes or not, we propose a semisupervised classification method to identify cancer driver genes, by applying one layer of a GAT and two layers of GraphSAGE to improve the performance of the model, where the GAT adds

weights to each interaction of the PPI network and GraphSAGE is utilized to improve the robustness of the model and make it more suitable for semisupervised tasks [20]. In order to test the robustness of the model, we select TCGA-SKCM with a high mutation rate for validation. Experimental results demonstrate that for all eight selected tumor types, the classification performance of GGraphSAGE outperforms most existing cancer driver genes identification methods, and it also has a good potential for discovering newly predicted cancer driver genes (NPCDs).

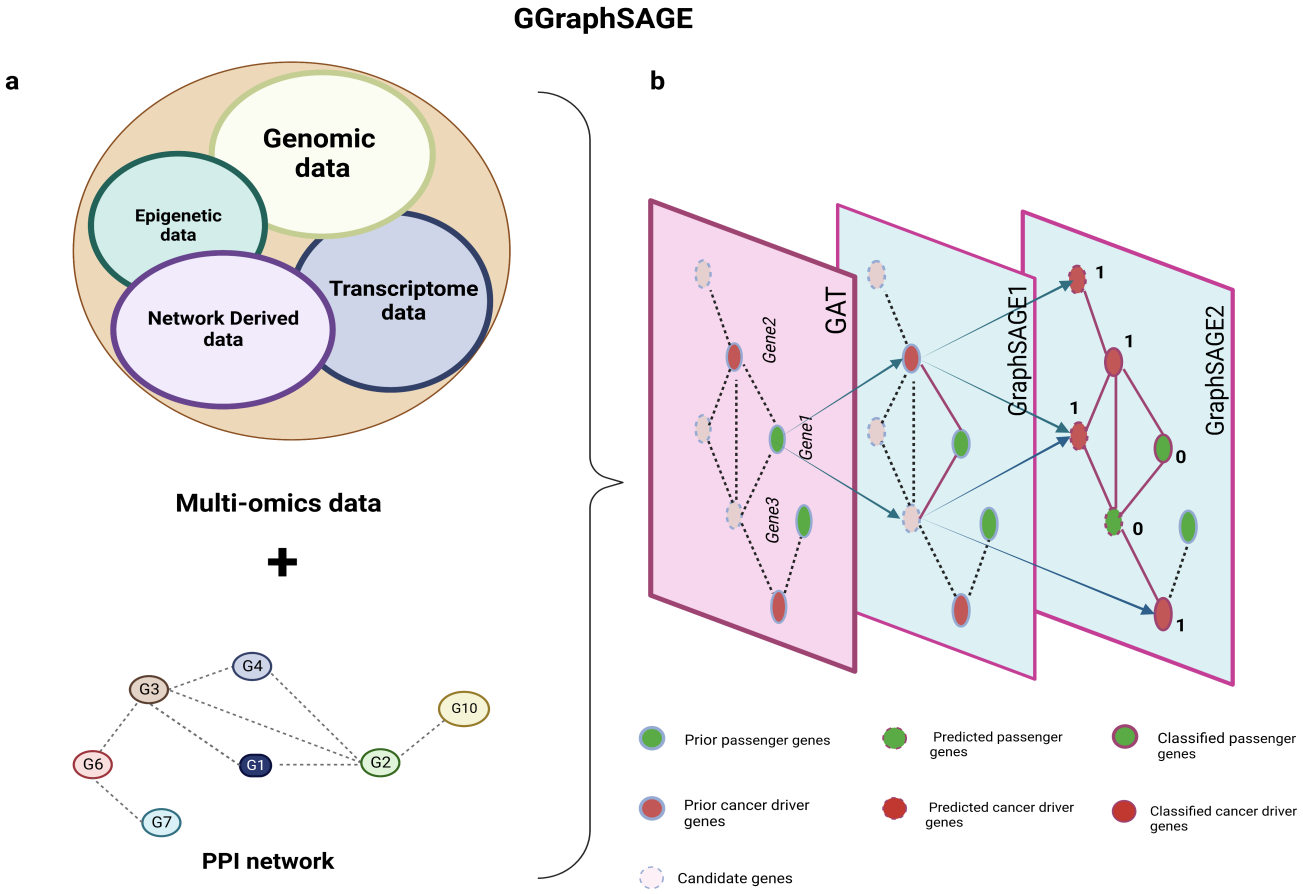

Figure 1. Schematic diagram of GGraphSAGE framework. (**a**) Inputs of GGraphSAGE framework include multiomics data (transcriptomic, genomic, and epigenetic data) and network and network-derived data; (**b**) all features and the PPI network are used for the semisupervised training of cancer driver genes. During the training of the GGraphSAGE model, the features are compressed by a three-layer graph aggregate operation and gradually include an increasingly wide range of neighboring nodes. The final output is determined by a probabilistic model (of whether it is a cancer driver gene).

2. Materials and Methods

2.1. Data Collection

2.1.1. Multiomics Data

We collected RNA-seq data, SNV data, DNA methylation data, and CNV data of cancer samples and control samples from TCGA [21]. In our study, we used 3778 tumor samples and 258 control samples across eight tumor types for analyses as each of them had complete multiomics data (Table 1). Table 1 gives the detailed sample size of each tumor type.

Table 1. Number of tumor samples and control samples in each tumor type.

Tumor Type	Number of Tumor Samples	Number of Control Samples
BLCA	408	19
BRCA	1095	19
LUSC	501	51
LUAD	515	59
ESCA	184	13
LIHC	371	59
STAD	238	33
SKCM	466	5

2.1.2. Network Data

We built the PPI network from the STRING database (https://cn.string-db.org/ accessed on 1 May 2021) and only considered the interaction whose score was greater than 0.9. We calculated the degree and betweenness centrality of each gene from the BioGrid network.

2.1.3. Assignment of Positive Labels and Negative Labels for Genes

Aiming at classifying the genes, we first generated a positive or negative label for each gene. The positive label referred to high-confidence cancer driver genes of different tumor types from the intOGene library [22]. The negative label referred to genes that were most likely not related to cancer. To obtain a list of negative genes, we excluded the following five gene types from the known nonpositives and assigned negative labels to the remaining genes:

1. Genes associated with the expression level of cancer driver genes.
2. Genes related to existing cancer pathways.
3. Cancer genes in the OMIM dataset [23] (https://omim.org//downloads accessed on 1 May 2021).
4. Known cancer driver genes in the DriverDBV3 dataset (http://driverdb.tms.cmu.edu.tw/ accessed on 1 May 2021).
5. Cancer driver genes in the NCG dataset [24] (http://ncg.kcl.ac.uk/ accessed on 1 May 2021).

Random pseudolabels were allocated to the above five types' genes, and eventually, the model learned and predicted the labels by semisupervised learning. Table 2 shows the number of positive and negative labels in each tumor type.

Table 2. Number of driver genes (positive labels) and nondriver genes (negative labels).

Tumor Type	Driver Genes	Nondriver Genes
BLCA	78	3586
BRCA	65	3646
LUSC	78	3586
LUAD	44	3690
ESCA	73	3659
LIHC	33	3636
STAD	35	3707
SKCM	14	3801

2.2. Feature Generation

A gene's features consist of its original multiomics information and network features derived from the PPI and BioGrid network. For each gene, for the transcriptome level, we computed the genes' average expression and outlier ratio; for the genome level, we counted the mutation frequency, CNV, base transition frequency, variant allele fraction, and mutational heterogeneity signatures obtained from MutSigCV; for the epigenome level,

we calculated the genes' average methylation. For biomolecular networks, we combined the biological network and the genomic and transcriptomic data in a bipartite graph. Four bipartite graph features were extracted for each candidate gene (Figure 2). Two network features were calculated by the BioGrid network. The values of each feature of different scales were normalized by row before being joined into a feature matrix (Figure 3).

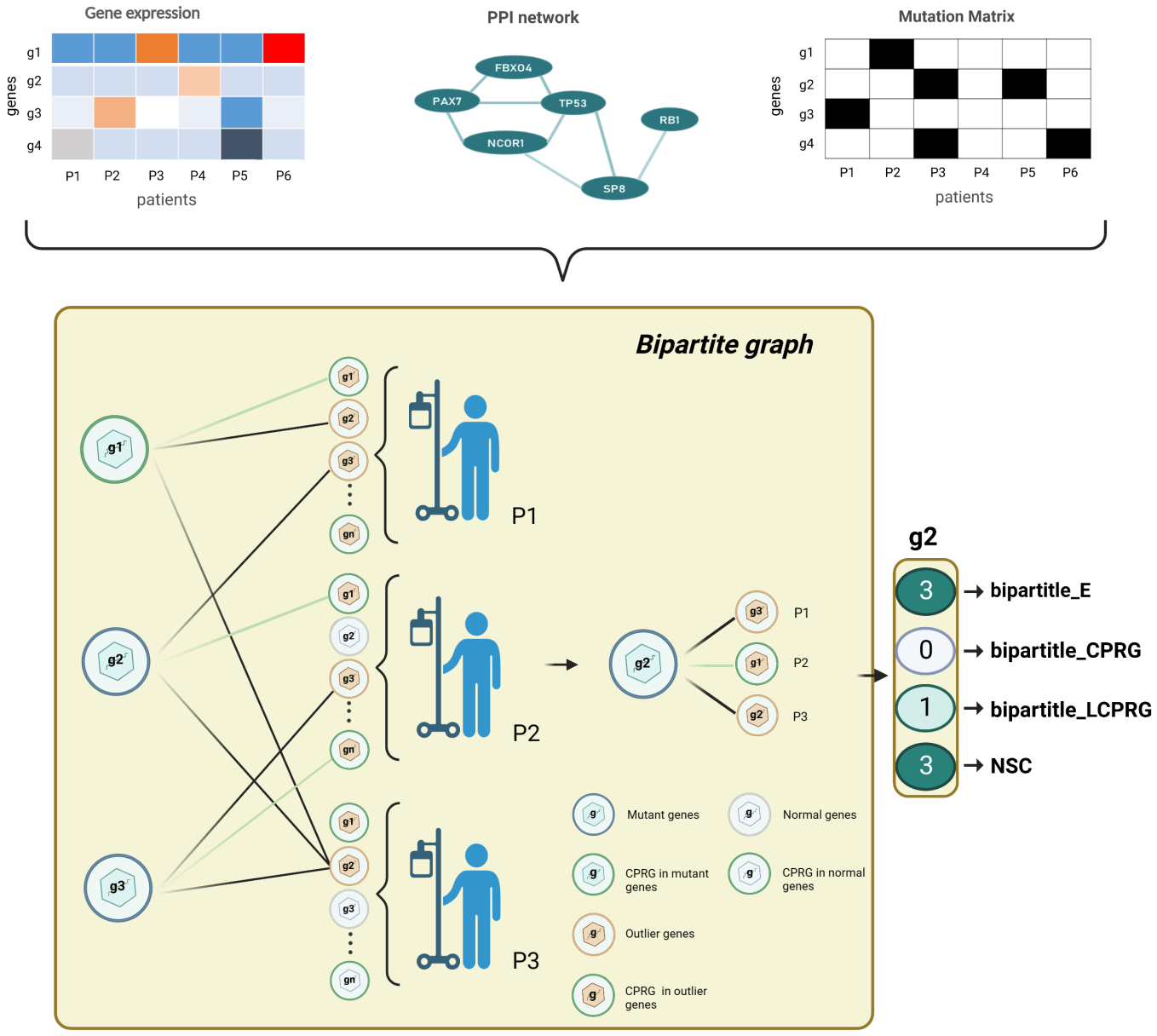

Figure 2. Bipartite graph architecture. For each tumor type, we built a bipartite graph by gene expression data, mutation data, and PPI network. We finally output a 4 × n dimensional feature matrix (n is the number of genes to be predicted). The left partition of the bipartite graph is the mutated genes, and the right partition of the bipartite graph is the outlying expression status for each tumor sample. *bipartite_E* represents the number of edges of a mutated gene; *bipartite_LCPRG* denotes the number of edges of a mutated gene linked to CPRGs; *bipartite_CPRG* is a binary number: if a gene was a CPRG, we set the *bipartite_CPRG* feature of the gene as 1, otherwise 0; NSC is the number of patients covered by a mutated gene.

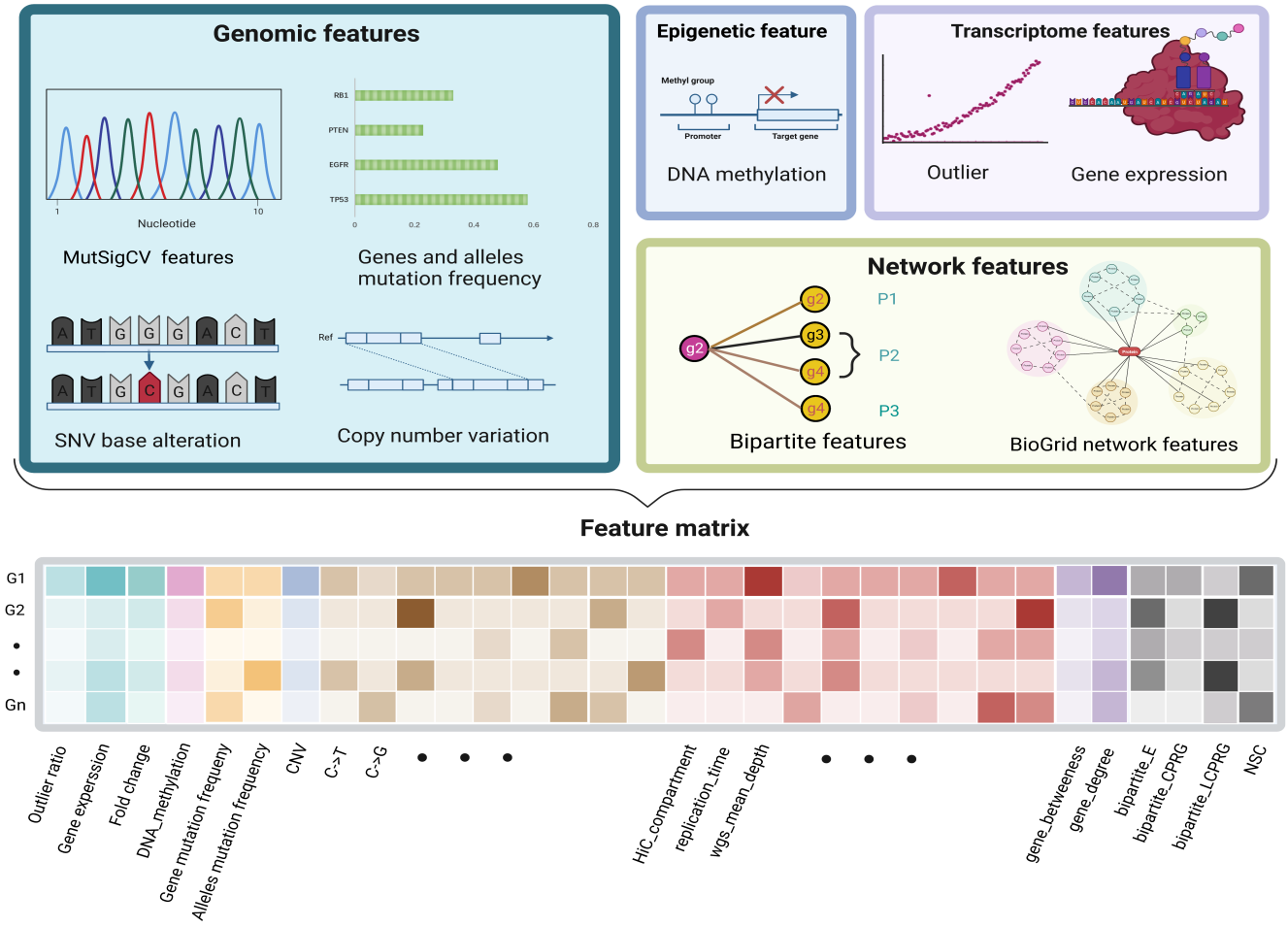

Figure 3. The characteristics of each gene. The attributes of a gene included the following four parts, 26 genomic features, 3 transcriptomic features, 1 epigenomic feature, and 6 network derived features from the bipartite graph and BioGrid network, which together formed a 1× 36-dimension feature vector for each gene.

2.2.1. Epigenetic Feature

For each gene i in tumor type T_j, we defined its epigenome feature as the fold change of mean methylation signal level (β) between the cancer and control samples:

$$Methylation_{T_j,i} = \frac{\bar{\beta}^C_{T_{j,i}}}{\bar{\beta}^N_{T_{j,i}}} \quad (1)$$

where $\bar{\beta}^C_{T_{j,i}}$ and $\bar{\beta}^N_{T_{j,i}}$ represent the average methylation level of gene i in cancer and control samples of tumor type T_j, respectively.

2.2.2. Transcriptome Features

Expression Fold Change

To quantify the differential expression level of each gene, we defined the log2 fold-change between the average expression level of cancer samples and the average expression level of normal samples as the expression fold-change feature as follows:

$$FoldChange_{T_j,i} = \frac{E\bar{X}P^C_{T_{j,i}}}{E\bar{X}P^N_{T_{j,i}}} \quad (2)$$

where $E\bar{X}P^{C}_{T_{j,i}}$ and $E\bar{X}P^{N}_{T_{j,i}}$ represent the average expression of gene i in cancer and control samples of tumor type T_j, respectively. For the genes which were not measured in cancer or normal samples, we set them to 0.

Significance of Differential Expression

In addition to the average expression ratio, we conducted a Wilcoxon test and took the differences in the distribution of the gene expression between cancer and control samples as an attribute. If there was no significant difference, the attribute value of the gene was 0; if there was a significant difference and the mean value of the normal group was smaller than that of the cancer group, the attribute value of the gene was 1; otherwise, the attribute value was -1.

Gene Outlier Feature

The outlier feature of a gene was defined based on the Grubbs test. First, we generated a patient-outlier matrix to represent the specificity of these genes. For each gene i, we used the difference between the expression value of each sample and the mean of the expression values of all samples to obtain G_{ij}, which was calculated as follows:

$$G_{ij} = \frac{EXP_{ij} - mean(EXP_i)}{S_i} \tag{3}$$

where EXP_{ij} is the expression level of gene i in tumor sample j, and S_i is the standard deviation of the expression level of gene i in all samples.

$Gp(N)$ was defined as follows:

$$G_p(N) = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}} \tag{4}$$

where N represents the number of measurements, and $t^2_{\frac{\alpha}{2N},N-2}$ denotes the critical value of the t-distribution. α is the significance level and we set it to 0.05. $\alpha/2N$ is the significance level of the t-distribution with degrees of freedom for $N-2$. If there existed $G_{ij} > G_p(N)$, then the value of the ith row and jth column in the outlier matrix was 1, otherwise it was 0. Then, we calculated the sum of the outliers of all cancer samples corresponding to each gene from the patient-outlier matrix as the outlier feature for the gene.

2.2.3. Genomic Features

Gene Mutation Frequency

For each tumor type, we generated a mutation matrix consisting of "1" and "0" with rows as genes and columns as samples. A "1" indicated that the gene had a single nucleotide variation in the sample, otherwise it was "0". We computed the mutation frequency of each gene in all cancer samples.

$$MF_{T_j,i} = \frac{m_{T_j,i}}{n_{T_j,i}} \tag{5}$$

where $MF_{T_j,i}$ represents the mutation frequency of gene i in the tumor type T_j's samples, $m_{T_j,i}$ and $n_{T_j,i}$ denote the number of samples with mutations in gene i and the number of total samples of cancer type j, respectively.

Variant Allele Fraction

To analyze the tumor heterogeneity and tumor purity, we generated the variant allele fraction feature by the ratio of the coverage depth of mutant loci in tumor samples to the coverage depth of sequencing as follows:

$$Variant\ allele\ fraction = \frac{Allele\ Depth}{Total\ Depth} \quad (6)$$

where *Allele Depth* denotes the depth of coverage of reads supporting mutant loci, and *Total Depth* denotes the depth of coverage of the total reads of that locus.

MutSigCV Derived Attribute

Mutation rate is an important indicator for identifying cancer driver genes. However, due to the difference in length and CG base content between genes, the average mutation frequency across samples did not completely reflect the real mutation level. To avoid this situation, we selected 11 quantified features from MutSigCV [9], aiming to reveal the regional heterogeneity present in nucleotide mutations and to evaluate the impact of these features on the identification of cancer driver genes.

1. DNA replication time;
2. Noncoding mutation rate;
3. Local GC content;
4. HiC compartment;
5. Local gene density;
6. Wgs mean depth;
7. Wgs percent 20x;
8. Capture on target rate;
9. Capture mean depth;
10. Capture pct200;
11. Capture mean percentGC.

Mutant Base Conversion

Existing studies have shown that there are significant differences in the frequency of base interchange in cancer. For example, the frequency of the CpG dinucleotide transition mutation in gastrointestinal cancer (esophagus, colon, rectum, and stomach) is high [25]. The frequency of the C->A and C->G type mutagenesis in bladder cancer cells is higher than that in other types of tumor cells [9]. Herein, we counted the number of base conversion corresponding to each gene in each sample and generated 12 features, namely C->A, C->T, C->G, A->T, A->C, A->G, G->C, G->A, G->T, T->A, T->G, and T->C.

Copy Number Variation Rate

We defined the copy number variation rate of a particular gene i in tumor type T_j as:

$$CNV_{T_j,i} = \frac{\sum_{k=1}^{m} tumorS_{i,T_j}^{k} \times n}{\sum_{l=1}^{n} normalS_{i,T_j}^{l} \times m} \quad (7)$$

where $tumorS_{i,T_j}^{k}$ and $normalS_{i,T_j}^{l}$ represent the number of amplifications or deletions of gene i in cancer sample k and normal sample l from the tumor type T_j, respectively. m and n are the number of cancer and normal samples.

2.2.4. Biological Network Derived Features

Bipartite Graph Attributes

For each tumor type, we used the gene expression matrix, mutation matrix and PPI network to construct a bipartite graph (Figure 2). Nodes in the left partition of the bipartite graph corresponded to mutated genes, and nodes in the right partition represented the outlier expression status of each patient. We drew an edge between two nodes in different partitions under the following conditions: in the same patient, mutations of the left-partition node gene i and the abnormal expression of the right-partition node gene j were found, and there was an interaction between gene i and gene j in the biological network. To estimate

the effect of each mutated gene on DNA replication, we added cell growth attribute to each gene according to whether it belonged to cell-proliferation-related biological functions or pathways in the bipartite graph and gave four features to each mutated gene: the number of connected edges, the number of edges connecting to cell-proliferation-related genes (CPRG), the gene's cell growth function attribute, and the number of patients covered by the gene (Figure 2).

BioGrid Network Features

We separately calculated the degree and betweenness centrality of the genes based on the BioGrid network. The degree centrality score is the number of connected edges of each gene in the network; the betweenness centrality is the ratio of the shortest paths in the network that pass through a gene and connect two genes to the total number of shortest path lines between these two genes. These two features reflected the importance and the topology characteristics of each gene in the network.

2.3. Model Construction

The proposed semisupervised deep graph learning framework GGraphSAGE was constructed using a GAT layer and two GraphSAGE layers, where it took the PPI network as the graph and each gene on the PPI network as a node. The advantage of GGraphSAGE is that it aggregates the neighborhood information in the graph more reasonably. The GAT quantifies the differences in interactions between homologous partners in biological large-scale networks by aggregating the weights of adjacent edges. At the same time, by sampling from the inside out and aggregating from the outside of GraphSAGE, using "concat" instead of "average", it can better distinguish the information of itself from its neighbors. We used the generated feature matrix and PPI networks as inputs, which allowed GGraphSAGE to receive both multiomics information and molecular interactions in biological systems. It also greatly helped to improve the overall performance of the model (Figure 4).

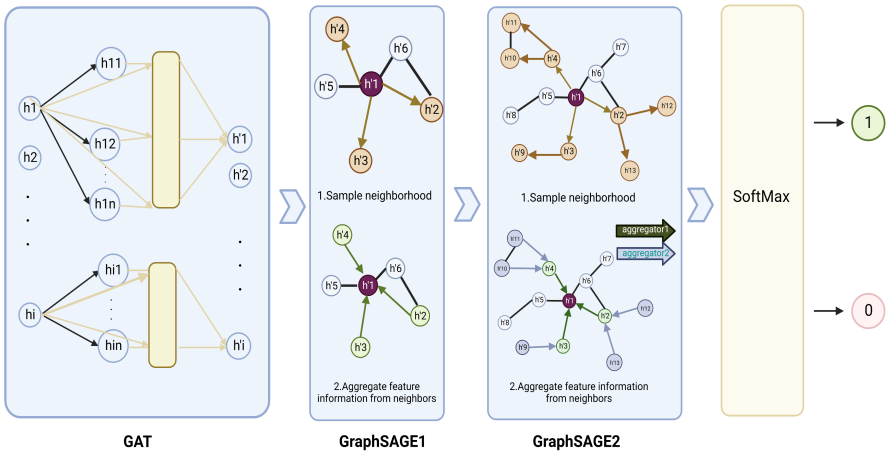

Figure 4. Graph neural network structure of GGraphSAGE framework. In the GGraphSAGE model, the feature matrix is aggregated around nodes through a GAT layer and two GraphSAGE layers for aggregating node, and each node contains the 3rd-order neighborhood information of the node. The final output is that each node (gene) is assigned a two-dimensional vector by the softmax layer, which consists of the probability that the node is a driver cancer gene and the probability that the node is a passenger gene. A label (1 or 0) is set to each node by comparing the magnitude of the two probabilities.

2.3.1. GraphSAGE Layer

GraphSAGE aims to improve the efficiency of a GCN and reduce noise. It learns an aggregator rather than the representation of each node, which enables one to accurately distinguish a node from its neighborhood information. In addition, it can be trained in

batches to improve the polymerization speed. Compared with a GCN, GraphSAGE was more flexible and suitable for our semisupervised task. The GraphSAGE algorithm is formulated as follows:

$$h_v^0 \leftarrow x_v, \forall v \in V \tag{8}$$

$$h_{N(v)}^k \leftarrow AGGREGATE_k(\left\{h_u^{k-1}, \forall u \in N(v)\right\}) \tag{9}$$

$$h_v^k \leftarrow \sigma(W^K \cdot CONCAT(h_u^{k-1}, h_{N(v)}^k)) \tag{10}$$

where x_v is the input feature, k is the depth of the feature matrix, $h_{N(v)}^k$ is defined as the aggregated representation of the neighbors of node v, h_v^k denotes the representation of the v node after aggregating its neighbors, and W is a learnable parameter.

2.3.2. GAT Layer

Since there are significant differences in the level of signal transduction between genes in biological systems, it is reasonable to set different weights for each edge in the PPI network. A GAT can set weights to each edge in the graph and was used to distinguish edges with various weights in the PPI network. A GAT computes the weight of each edge as follows:

$$a_{ij} = \frac{exp(LeakyReLU(a^T\left[Wh_i||Wh_j\right]))}{\sum_{k \in N_i} exp(LeakyReLU(a^T[Wh_i||Wh_k])} \tag{11}$$

where a_{ij} represents the weight relation coefficients of node i and node j, a^T is a learnable vector ($a^T \in R^{1\times n}$), and n is feature dimension of each node. W is a learnable weight matrix ($W \in R^{n\times n}$). h_i denotes the embedding of node i. N_i are all the neighbor nodes of i.

$$h_i^{'} = \sigma(\sum_{j \in N_i} a_{ij} Whj) \tag{12}$$

where σ is defined as the nonlinear activation function. $h_i^{'}$ denotes the embedding of node i after aggregating all the neighboring nodes.

2.4. Model Training

We randomly divided all the samples into a training set and test set with the proportion of 70% and 30%, respectively, and took the generated feature matrix with partial node labels and the PPI network as inputs. The cross-entropy loss L for our training nodes was defined as:

$$L = \frac{1}{N}\sum_g L_g = \frac{1}{N}\sum_g -\left[y_g \cdot log(p_g) + (1 - y_g) \cdot log(1 - p_g)\right] \tag{13}$$

where y_g is the label of gene g, with 1 representing the positive class and 0 representing the negative class. p_g is the probability of gene g being predicted as positive. We implemented our framework using Torch [26] and optimized the training process with the Adam optimizer. The parameters for each layer are listed in Table 3.

Table 3. Parameters of GGraphSAGE framework.

Tumor Type	GAT (Input/Output Size)	GraphSAGE1 (Input/Output Size)	GraphSAGE2 (Input/Output Size)
BRCA	36/64	64/128	128/2
BLCA	36/64	64/128	128/2
ESCA	36/64	64/32	32/2
LUAD	36/256	256/64	64/2
LIHC	36/256	256/128	128/2
LUSC	36/128	128/32	32/2
STAD	36/128	128/64	64/2
SKCM	36/64	64/256	256/2

3. Results

3.1. Evaluation Metrics

We evaluated the performance of the model across eight tumor types and compared it with eight methods for identifying cancer driver genes, including state-of-the-art methods and classic machine learning methods. We calculated the average precision (AP) for each tumor type, which is the area under the precision–recall curve (P-R curve) for each method to quantify the performance of the models. We drew the P-R curve according to the following:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

where TP is the number of genes correctly classified as positive by the model; FN denotes the number of genes incorrectly classified as negative by the model; and FP represents the number of genes incorrectly classified as positive by the model. We obtained the approximate area (AP) under the P-R curve by integration:

$$AP = \int_0^1 p(r)dr \tag{16}$$

Because of the discreteness of the data, we actually smoothed the P-R curve rather than compute it directly. For each point on the P-R curve, the precision value took the value of the maximum precision to the right of that point. Next, on the smoothed P-R curve, the precision value of the 10 equipoints of the recall axis (including 11 points and breakpoints) was taken, and its average value was calculated as the final AP value. We calculated AP as follows:

$$P_{smooth}(r) = max_{r' >= r} P(r) \tag{17}$$

$$AP = \frac{1}{11} \sum_{0,0.1 \ldots 1.0} P_{smooth}(i) \tag{18}$$

where $P_{smooth}(r)$ represents the precision value of the model when the recall value is r after the smoothing process. The AP is between zero and one, and the larger the AP, the better the performance (Figure 5).

3.2. Performance of GGraphSAGE by Comparing with SOTA Methods

To evaluate the performance of GGraphSAGE, we first compared it with three SOTA cancer driver gene identification methods, CanDrA [5], 20/20+ [27], and EMOGI [8]. CanDrA takes chromosome number, genomic coordinate, reference allele, mutated allele, and the strand of the mutation as the features of each gene and obtains a score to determine whether this gene is a driver or not. For 20/20+ [27], we used the mutation data from TCGA and 24 attributes acquired from BioGrid [28], MutsigCV [9], and SNVBox [29] as the features and applied a random forest [30] to predict the probability of the genes being drivers. For EMOGI [8], we used DNA methylation, RNA-seq, and CNV data collected from TCGA for different tumor types. We obtained the PPI network from the STRING database and only retained the interaction scores greater than 0.9. We randomized the labeled dataset to the training set (75%) and the test set (25%) for training and used a 10-fold cross-validation for the parameter optimization and improvement of the model stability.

The GGraphSAGE model performed better than all three SOTA methods. Overall, 20/20+ had stable performance across different cancer types but generally had a poorer performance compared with GGraphSAGE. CanDrA achieved good performance on STAD, LUAD, and LIHC, but it was overly dependent on the dataset and had unstable performance on the other five tumor types. EMOGI was a similar approach to GGraphSAGE in model structure and also used multiomics data and graph convolution networks. Although EMOGI was proposed for the identification of pan-cancer genes, it could also work in

predicting cancer driver genes of specific tumor type. The performance of EMOGI was not good on LIHC and LUAD, while it performed well on other tumor types (Figure 5).

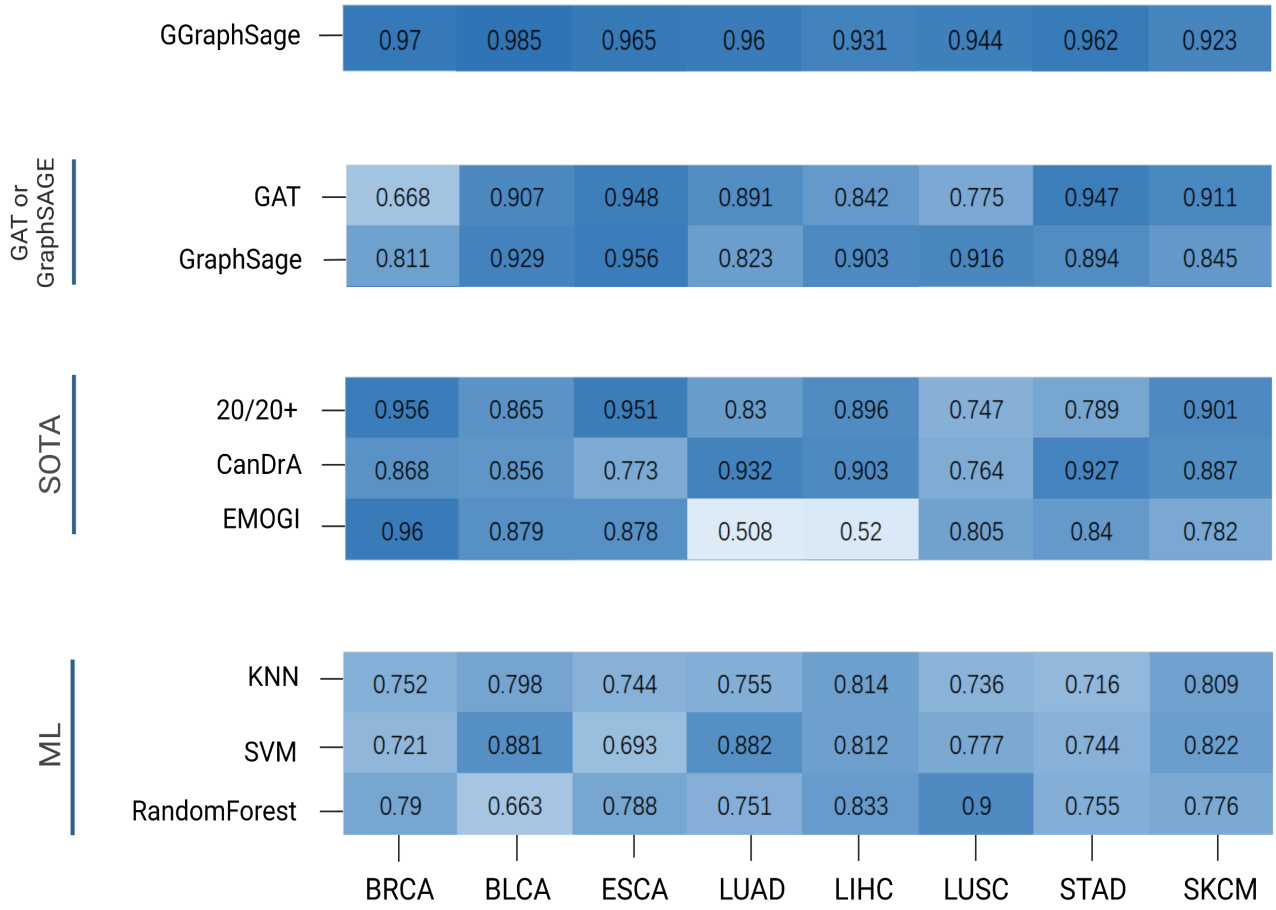

Figure 5. Performance comparison of GGraphSAGE with other methods. The heatmap reveals the performance (AP) comparison of different methods for each tumor type, with darker colors indicating higher AP values. These methods were divided into 4 categories: GGraphSAGE: the combination of GAT and GraphSAGE; GAT or GraphSAGE: GAT or GraphSAGE model only; SOTA methods: 20/20+, CanDrA, and EMOGI; ML (machine learning): KNN, SVM, and random forest. As can be seen from the figure, GGraphSAGE has a high AP value on each tumor type, and its performance is better than other methods.

3.3. Performance of GGraphSAGE by Comparing with Classical Machine Learning Models

We also compared GGraphSAGE with three classical machine learning models, including K-nearest neighbors (KNN) [31], support vector machines (SVM) [32], and random forests.

KNN is a classical algorithm for supervised learning classification based on the distance between the node and the nearest k nodes and performs well in binary classification tasks. An SVM is a binary classification model. It is the nonlinear classifier defined on the feature space with maximum interval. A random forest is a classifier composed of many decision trees, in which each tree selects the optimal feature recursively and divides the training data according to the feature.

The results showed that the three models generally performed poorly overall on the tumor types, except for the SVM on LUAD and BLCA, and the random forest on LUSC. The AP of GGraphSAGE was generally 20% higher than the three machine learning models on all tumor types, which indicated that the addition of biological network structure information was helpful for cancer driver gene identification (Figure 5).

3.4. Ablation Experiments

To prove the superiority of the combination of a GAT and GraphSAGE, we calculated the AP of the GAT and GraphSAGE, respectively. We set the same inputs as for GGraphSAGE and only changed the model parameters. It can be seen that the performance of the GAT and GraphSAGE was lower than that of GGraphSAGE.

In the GAT, nodes in the graph can be assigned different weights based on the characteristics of their neighbors. The GAT is suitable for calculating the specificity interactions in PPI networks. However, there is noise (false-positive interacting protein) in the topology of the PPI network [33] which affects the performance of the GAT. GraphSAGE introduces neighbor sampling. Through the "concat" method of aggregating neighborhood nodes, GraphSAGE can distinguish itself from neighborhood information to reduce the influence of noise data and thus improve the robustness of the model [17,18]. Although GraphSAGE samples neighborhood nodes to improve the efficiency of training, some neighborhood information is lost. The method of node aggregation in GGraphSAGE improves the robustness of the model, allowing sampling nodes to be aggregated with nonequal weights, while preserving the integrity of the first-order neighborhood structure information.

As a result, GraphSAGE performed well on LIHC, ESCA, BLCA, and LUSC but moderately on the other four tumor types. The GAT performed well on STAD, ESCA, SKCM, and BLCA but poorly on the other tumor types, which indicated the performance of the GAT model was highly dependent on the data set and had a low stability. In contrast, GGraphSAGE performed well and was stable across all tumor types, which suggested that GGraphSAGE was superior to GAT or GraphSAGE alone.

Moreover, to evaluate the contribution of each type of data, we conducted ablation experiments on each type of data. Specifically, we removed one type of data in each experiment and used the remaining data to identify the cancer driver genes by GGraphSAGE. Then, we calculated the AP to assess the performance. The results showed that across the eight cancer types, the complete multiomics data (containing genomic, epigenetic, transcriptome, and network-derived features) performed best, which illustrated the necessity of leveraging the complete multiomics and network data (Figure 6). After removing any one type of features, the performance of the model decreased, which proved that each type of data contributed to the model. For eight cancer types, removing genomic data had the biggest effect on model performance (a mean 39.3% decline in AP), while removing epigenetic data had the least effect (a mean 10.7% decline in AP). It indicated that genomic features contributed the most to the classification and prediction of cancer driver genes, while epigenetic features contributed the least. Compared to other cancer types, the performance was poor after removing transcriptome features in BLCA, LUAD, and STAD, which indicated that transcriptome features contributed significantly to these three cancer types, while network-derived features contributed more to the other types.

3.5. Association between Newly Predicted Genes and Cell Proliferation

Uncontrolled cell proliferation is a determinant of carcinogenesis and driver mutation. To assess the effect of mutations on cell proliferation, we analyzed the differences in abnormal proliferation levels between samples with and without mutations (Figure 7). Specifically, we first retrieved 160 CPRGs from GSEA that were included in DNA replication and cell cycle pathways, generated the transcriptomics data matrix of N samples on the CPRG, and calculated the Spearman correlation coefficient between these genes by setting 0.4 as the threshold. For the active cancer-related biological process, most genes worked together and more than half of genes were significantly coexpressed, so we set the low significance cutoff of correlations as 10^{-3} and selected a set of top 20 genes which were highly correlated with each other as core genes to represent the whole cell proliferation process. Subsequently, we used the linear regression coefficient between core genes' expression value of each sample and core genes' average expression value vector on N

samples, as the cell proliferation activity of each sample. The objective function of the linear regression was:

$$\hat{S}_\theta = \theta \cdot S \tag{19}$$

$$RST = \sum_{i=1}^{n} (S_{mean}^{(i)} - \hat{S}_\theta^{(i)})^2 \tag{20}$$

where θ represents the weights of samples, that is, the parameter that minimizes the squared term of the residual; S is the sample column ($S \in R^{1\times 20}$) in the core gene matrix; $\hat{S}_\theta$ represents the predicted value of a linear regression function that is used to fit the mean vector of the sample ($\hat{S}_\theta \in R^{1\times 20}$); and S_{mean} denotes the mean expression level of core genes across all samples. We took θ as the samples' cell proliferation level.

To evaluate the effect of a mutation on cell proliferation, we divided the cancer samples into two groups: one with such mutation and the other without (M and U in Figure 7). Then, we conducted the Wilcoxon test analysis to assess the significance of differences in cell proliferation between these two groups. We expected the driver mutations to cause abnormal proliferation, but the passenger mutations were not associated with abnormal proliferation. We identified the top 20 candidate genes for each tumor type according to the scores (the score was the probability of being predicted as cancer driver genes in the model) obtained from GGraphSAGE and removed genes that overlapped with those of the intOGene database. Candidate genes with a p-value less than 0.05 indicated a significant alteration of the corresponding gene was closely related to cell proliferation and thus were ultimately selected as cancer driver genes to further demonstrate their association with cancer (Figure 8). The information of the candidate NPCDs is listed in Table 4.

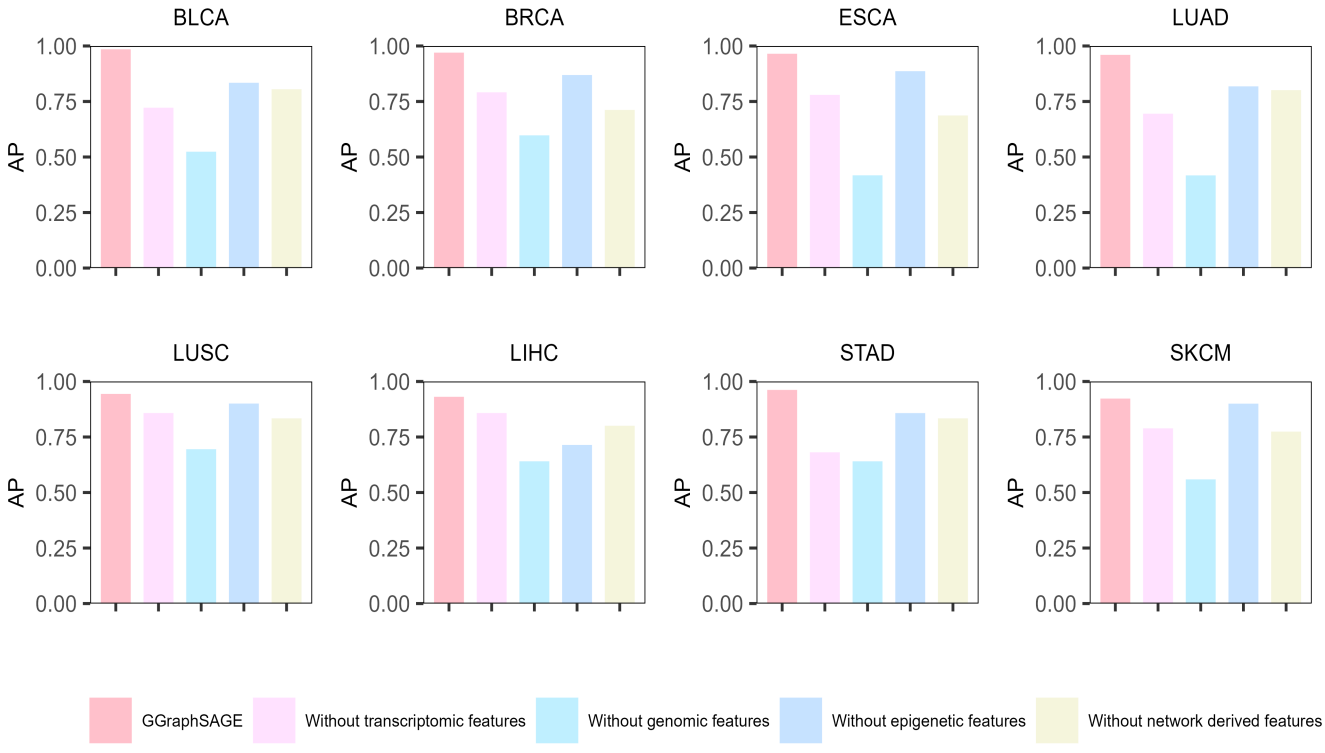

Figure 6. Results of ablation experiment of GGraphSAGE in eight cancer types. The bar chart represents the model performance (AP) of GGraphSAGE across eight cancer types when removing genomic, epigenetic, transcriptome, and network data, respectively.

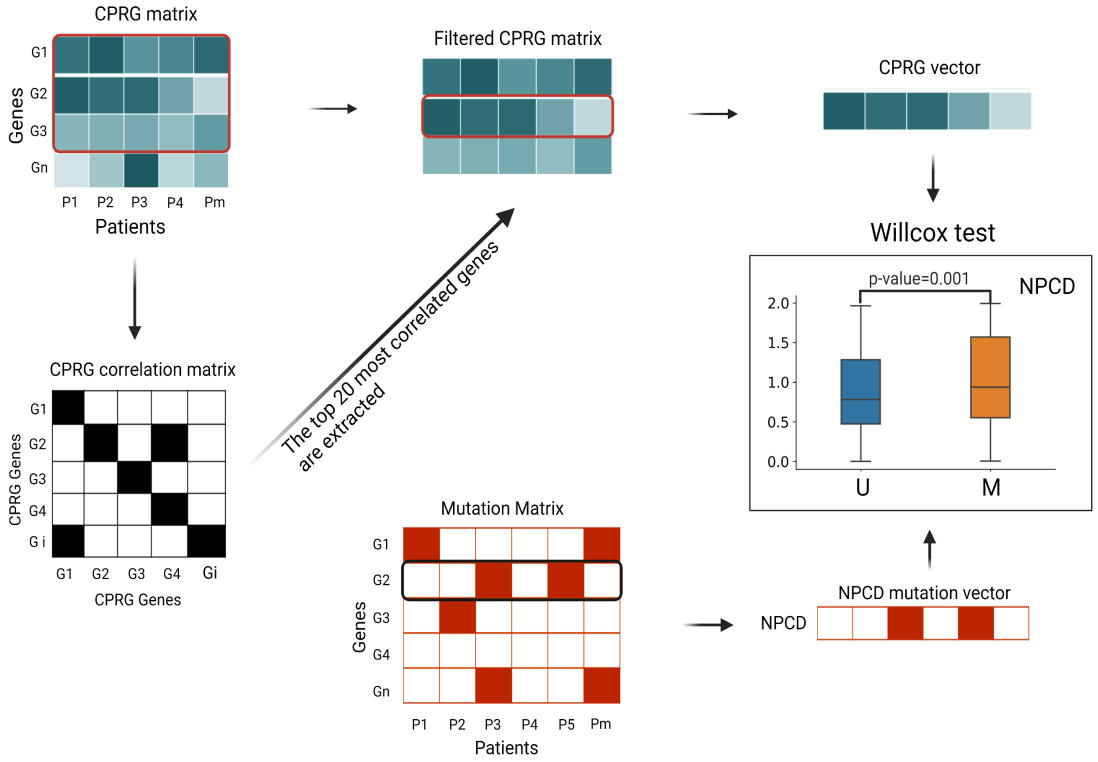

Figure 7. Flow chart of DNA replication function correlation test. (1) We extracted the fraction of all CPRGs from the *gene expression matrix* $\in R^{20500\times n}$ (n is the number of samples) of each tumor type and created the *CPRG matrix* $\in R^{160\times n}$. (2) According to the Spearman correlation coefficient threshold, we generated *CPRG correlation Matrix* $\in R^{160\times 160}$(symmetric matrix). (3) We selected the top 20 genes with the highest correlation in the CPRG matrix and extracted the top 20 genes from the CPRG matrix to produce the *Filtered CPRG matrix* $\in R^{20\times n}$. (4) The mean expression level of each CPRG in the filtered CPRG matrix was used to calculate the linear regression parameters for each sample, as the CPRG vector for the sample. The CPRG vectors were divided into mutation/nonmutation vectors according to the mutation matrix, and the Wilcoxon test was performed on the distributions of these two vectors to compute the *p*-values. The box plots are drawn for each NPCD by the two sample-population distributions (Figure 8).

3.6. Newly Predicted Cancer Driver Genes and Their Verified Functions

On the basis of the cell proliferation analysis, we further verified the functions of the candidate driver genes. Most of the genes have been extensively studied and are directly and indirectly associated with tumor development in the published literature. Some genes are identified as molecular markers of carcinogenesis. For example, CUL1 is defined as a biomarker for breast cancer because it significantly promotes breast cancer cell migration, invasion, and test-tube formation, as well as angiogenesis and metastasis in vivo [34]. Alternatively, the expression levels of some genes affect the functions involved in the transformation of normal cells into cancer cells. For example, ACTN2 overexpression in human hepatoma cells shows an enhanced cell viability and invasion, suggesting that it may play a role in late metastasis, such as exosmosis and lung colonization [35]. Besides extensive studies on gene expression level and function, some genes influence the occurrence and development of cancer at the genomic and epigenetic levels. Sucularli et al. reported that the deleterious mutations of DYNC1H1 led to the formation of associated cancers [36]. Lin et al. found the methylation of RILP in lung cancer promoted tumor cell proliferation and invasion [37]. Hence, we conclude that the NPCDs are expected to provide guidance for researchers on further studies.

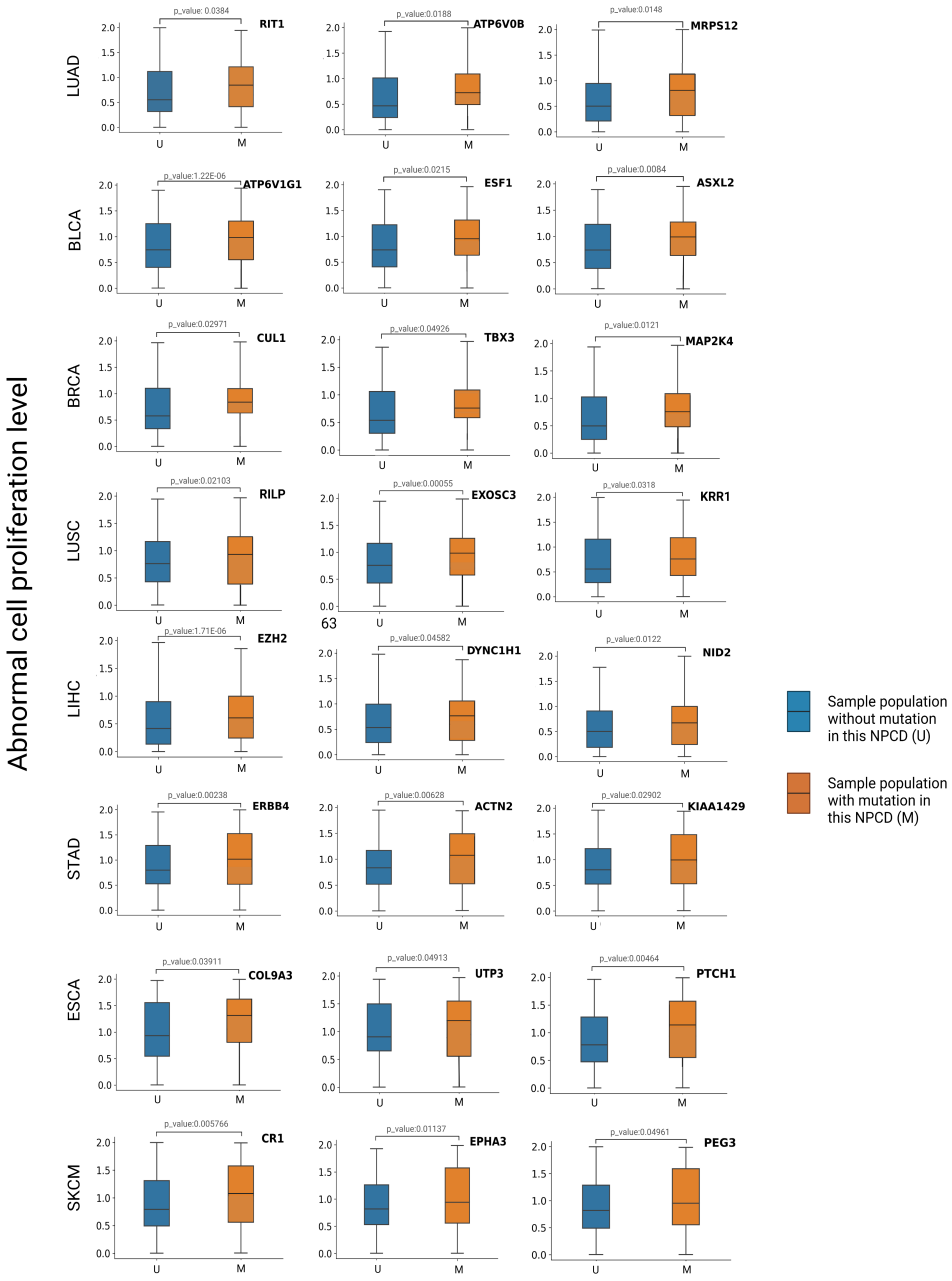

Figure 8. Box plots of abnormal cell proliferation level of CPRG. The box plots show the differences in the distribution of abnormal proliferation levels between samples with and without mutations in different tumor types. The blue and orange boxes represent the distribution of the U and M sample populations, respectively. The vertical axis is the abnormal cell proliferation level of each sample. We conclude that mutations in candidate genes with a p-value < 0.05 in the CPRG abnormal cell proliferation analysis result in elevated levels of abnormal cell proliferation.

Moreover, eight genes predicted by GGraphSAGE were also coauthenticated by the most popular 30 SOTA methods (e.g., CHASM [29], e-Driver3D [38], OncoIMPACT [39], PolyPhen2 [40], CTAT-score [41]). These genes include TBX3, CUL1, and MAP2K4 for BRCA, PTCH1 for ESCA, ASXL2 for BLCA, EZH2 for LIHC, RIT1 for LUAD, and ERBB4 for STAD.

In summary, we demonstrated that GGraphSAGE had a higher performance than other state-of-the-art methods based only on multiomics, biological networks, or using simple graph neural networks in the classification of cancer driver genes by tumor type. We also

provided strong evidence for the credibility of new driver genes predicted by GGraphSAGE at both theoretical and function levels through a literature review and functional correlation analysis (Table A1).

Table 4. Number of samples of NPCD mutant/nonmutant group and median of abnormal cell proliferation level of sample groups.

Tumor Type	Gene Name	M/U Group Size	Median of Abnormal Cell Proliferation Level (M/U)
BRCA	CUL1	50/1045	0.971/0.551
	TBX3	85/1010	0.896/0.548
	MAP2K4	104/991	0.743/0.51
BLCA	ATP6V1G1	22/389	0.998/0.75
	ESF2	53/358	0.999/0.749
	ASXL2	64/347	1.12/0.561
ESCA	UTP3	23/161	1.25/0.987
	COL9A3	19/165	1.3/0.99
	PTCH1	32/152	1.131/0.748
LUAD	RIT1	86/481	0.801/0.562
	ATP6V0B	30/537	0.731/0.433
	MRPS12	30/537	0.799/0.561
LIHC	EZH2	79/292	0.688/0.49
	DYNC1H1	63/308	0.789/0.55
	NID2	48/323	0.727/0.501
LUSC	RILP	53/448	0.975/0.691
	EXOSC3	39/462	1.081/0.75
	KRR1	45/456	1.191/0.744
STAD	ERBB4	35/402	1.011/0.771
	ACTN2	48/389	1.232/0.822
	KIAA1429	28/156	1.088/0.752
SKCM	CR1	68/399	1.11/0.786
	PEG3	85/382	0.977/0.812
	EPHA3	48/413	1.134/0.78

4. Discussion

Carcinogenesis is usually thought to result from the accumulation of key alterations, both at the genetic and nongenetic level. Hence, the identification of cancer driver genes is important to discover drug targets for specific cancer types. In this study, we proposed a novel semisupervised graph-neural-network-based framework, GGraphSAGE, which integrated multiomics and network-derived features to identify cancer driver genes for each cancer type. We evaluated the performance of GGraphSAGE against three state-of-the-art methods and three classic machine learning methods under the average precision (AP) index across eight tumor types. Experiments showed that GGraphSAGE performed better than all these methods in both precision and stability. Considering cancer is a highly heterogeneous and complex disease, the disease initiation mechanism, microenvironment composition, and key alterations of various cancer types should be intuitively distinct [42]. Hence, different from most existing methods focusing on pan-cancer driver genes, we proposed to identify cancer driver genes based on the characteristics specific to each cancer from various aspects. Besides widely studied genomic features, we also focused on nongenetic alterations including transcriptomic and epigenomic features, as well as biological-network-derived features, and generated a feature vector of length 36 for each gene at the end. It is well known that genes in the biological systems do not work independently, and the disruption of some driver genes may promote the initiation and progression of cancer. Hence, taking advantage of the information contained in biological networks is highly important for predicting cancer driver genes. Indeed, most advanced

methods also take this into account. The combination of valuable information contained in graph and deep learning approaches, graph neural networks have gradually been used in classifying nodes with outstanding performance [8,13]. Herein, considering large numbers of genes cannot be definitely determined as cancer driver genes or not, we proposed a semisupervised classification method to identify cancer driver genes, applying a GAT and GraphSAGE to improve the performance of the model. Ablation experiments demonstrated that the performance of jointly applying the GAT and GraphSAGE models was much better than using either one alone. Moreover, the application of GraphSAGE was able to provide an availability assurance for scalable networks. After years of efforts, some public databases, such as COSMIC [43] and intOGene, have release parts of driver genes. Beyond these, GGraphSAGE was able to identify new driver genes with a high potential. Through a literature review and functional correlation analysis (**see Results**), we demonstrated the credibility of the newly predicted ones. We took abnormal cell proliferation as an indicator of carcinogenesis and determined whether a candidate gene was a driver by evaluating the association between its alterations and this indicator. In addition, other key biological functions, such as EMT [44], which is an important cancer development indicator, will also be considered in our further studies for identifying key drivers of cancer progression.

The application area of the GGraphSAGE model is broad, as it can be applied to any multiomics data and biological network other than this study. Moreover, the predicted diseases are not limited to cancer but can also be applied to other complex diseases. Our GGraphSAGE is designed to classify cancer driver genes and passenger genes in various tumor types. The semisupervised mechanism of the model also identifies NPCDs that play a critical role in tumor development and the impact on cell growth function (Table A1). These NPCDs are highly similar to the cancer driver genes in terms of multiomics embedding and biological network structure. The GGraphSAGE framework provides an important analytical tool for the future of precision medicine and for understanding the process of tumor development and targeted therapy.

5. Conclusions

GGraphSAGE is a graph neural network framework for accurately and efficiently distinguishing cancer driver genes, which is empowered by the generated comprehensive multiomics features and the combination of GraphSAGE and GAT models. Through statistical analyses, we found that the alterations of some newly identified cancer driver genes influenced cell proliferation function and were reportedly associated with cancer initiation and development. The GGraphSAGE framework provides a new insight to identify the driver genes of complex disease and is helpful to understand the process of disease development and design targeted therapy.

Author Contributions: H.S. (Hongzhi Song) and H.S. (Huiyan Sun) conceived the idea of GGraphSAGE. H.S. (Hongzhi Song) and K.F. collected the multiomics data and preformed the experiments. H.S. (Hongzhi Song) conducted the data analyses. H.S. (Huiyan Sun) and K.F. helped with the components of the feature matrix. H.S. (Huiyan Sun) supervised the study and provided the resources. H.S. (Huiyan Sun), Y.C., C.Y., Y.G. and Z.L. helped to review and edit the manuscript. H.S. (Hongzhi Song) wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the funding support from the National Natural Science Foundation of China (no. 61902144) and Special Project for Medical and Sanitary Talent of Jilin Province (JLSWSRCZZX2021-039).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at https://github.com/JP909/GGraphSAGE accessed on 3 July 2023.

Acknowledgments: The authors are grateful to Sun Hui Yan for her help with the preparation of the figures and the language in this paper. This work was supported by the School of Artificial Intelligence at Jilin University. We would like to thank the anonymous reviewers for their helpful remarks.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Newly predicted cancer driver genes by the GGraphSAGE model. Screening of the literature and *p*-value, including gene ID, gene functions, statistical *p*-value of the association between the gene's mutation and cell proliferation, and reference of existing reports.

Tumor	Gene ID	Literature Description	*p*-Value	Citation
STAD	ERBB4	Members of the ErbB subfamily of receptor tyrosine kinases are important regulators of normal breast physiology, and abnormalities in their signaling have been associated with breast tumorigenesis.	0.00238	[41,45]
	ACTN2	ACTN2-overexpression in human hepatocellular carcinoma cells showed enhanced cell motility and invasive ability, suggesting possible functions in late metastatic stages, such as extravasation and lung colonization.	0.00628	[35]
	KIAA1429	M6 RNA methylation has a huge impact on RNA production/ metabolism and is involved in the pathogenesis of many diseases, including cancer. M6 is modified by m-mounted 6 methyltransferases (METTL3/14, WTAP, RBM15/15B and KIAA1429, referred to as "writers"), reduced by (FTO and ALKBH5, referred to as "erasers") and recognized by m-mount6 binding proteins (YTHDF1/2/3, IGF2BP1 and HNRNPA2B1, called "readers").	0.02902	[46]
BLCA	ESF1	Transcription of the E1A gene of highly oncogenic adenovirus 12 (Ad12) starts at two initiation sites (TS1 and TS2). We have previously shown that the distal ends of the E2F and ATF motifs of TS1 are synergistically involved in E1A self-stimulation in the TS1 promoter region. Here, we report the identification of a second E2F-like target region (E2DFII) upstream of the E1A stimulatory factor 1 binding site (ESF-1), which is important for 13S-mediated self-activation of TS2.	0.0215	[47]
	ATP6V1G1	Low mRNA and protein expression of ATP6V1s members were found to be significantly associated with clinical cancer stage, lymph node metastasis status, and patient gender in KIRC patients. In addition, ATP6V1A, ATP6V1B2, ATP6V1C1, ATP6V1C2, ATP6V1D, ATP6V1E1, ATP6V1E1, ATP6V1E1, ATP6V1E1, ATP6V1E1, ATP6V1G1, and ATP6V1H had lower mRNA expression but shorter OS.	1.22×10^{-6}	[48]
	ASXL2	Upregulation of ASXL2 was associated with advanced clinical staging. Patients exhibiting high levels of ASXL2 expression had poorer overall survival, while those with low ASXL2 expression levels survived longer. A multifactorial Cox regression analysis showed that ASXL2 expression could be considered as an independent prognostic factor for CRC. Inhibition or overexpression of ASXL2 significantly affected the proliferation of CRC cells.	0.0084	[41,49]

Table A1. *Cont.*

Tumor	Gene ID	Literature Description	*p*-Value	Citation
BRCA	CUL1	CUL1 significantly promoted breast cancer cell migration, invasion, and in vitro tube formation, as well as angiogenesis and metastasis in vivo. Mechanistically, global transcriptional changes in MDA-MB-231 cells were determined using human gene expression profiling, and autocrine expression of cytokines CXCL8 and IL11 were identified as target genes of CUL1 in breast cancer cell migration, invasion, metastasis, and angiogenesis.	0.02971	[34,41]
BRCA	TBX3	TBX3 has no known function in adult tissues but is frequently overexpressed in a wide range of epithelial- and mesenchymal-derived cancers. This overexpression greatly affects several features of cancer, including senescent bypass, apoptosis and deactivation, promotion of proliferation, tumor formation, angiogenesis, invasive and metastatic capacity, and cancer stem cell expansion.	0.04926	[41,50]
	MAP2K4	Genetic variants (copy number variants and single nucleotide polymorphisms) and acquired somatic copy number aberrations (CNAs) were associated with approximately 40% of gene expression, with the landscape dominated by cis and trans CNAs. By depicting genes with expression aberrations driven by CNA in cis, we identified putative cancer genes, including deletions in PPP2R2A, MTAP, and MAP2K4.	0.0121	[41,51]
ESCA	COL9A3	USP3 promotes GC progression and metastasis by deubiquitinating C-OL9A3 and COL6A5. These findings identify a mechanism for the USP3-mediated deubiquitination of enzymatic activity in GC metastasis and suggest a potential therapeutic target for GC management.	0.03911	[52]
	PTCH1	Dysregulation of the Hh signaling pathway is associated with developmental abnormalities and cancers (including Gorlin syndrome) and sporadic cancers (e.g., basal cell carcinoma, medulloblastoma, pancreatic, breast, colon, ovarian, and small-cell lung cancers). Abnormal activation of the Hh signaling pathway is caused by mutations in related genes (ligand nondependent signaling) or by overexpression of Hh signaling molecules (ligand-dependent signaling—autocrine or paracrine).	0.00464	[41,53]
	UTP3	UTP3, the small subunit process group homolog (UTP3), and prostaglandin E synthase 3 (PTGES3). Thus, pathway functions in dynamic module 3 (ubiquitin-mediated protein hydrolysis and ribosomes) and several seed genes (PPP1R12A, UTP3, and PTGES3) may be associated with OS progression and could serve as potential therapeutic targets in OS.	0.04913	[54]
LIHC	EZH2	Recent findings on the role of EZH2 in cancer genesis, progression, metastasis, metabolism, drug resistance and immune regulation. In addition, we highlight the progress of targeted EZH2 therapies in experimental and clinical studies.	1.71×10^{-6}	[41,55]
	DYNC1H1	The deleterious mutations were found to affect the function of DYNC1H1 leading to the formation of associated cancers.	0.04582	[36]
	NID2	NID2, SPARC, and MFAP2 were upregulated in gastric tumor tissues and significantly associated with poor overall survival. Thus, by using these 3 upregulated DEGs, the predictive value of the risk score model used for gastric cancer prognosis could be improved.	0.0122	[56]

Table A1. *Cont.*

Tumor	Gene ID	Literature Description	*p*-Value	Citation
LUSC	RILP	Methylation of RILP in lung cancer promotes tumor cell proliferation and invasion.	0.02103	[37]
LUSC	EXOSC3	In inflamed mucosa, EXOSC3 and CNOT4-mediated RNA stabilization, including that of MYD88, may trigger cancer development and could serve as potential predictive markers and innovative therapies to control cancer progression.	0.00055	[57]
	KRR1	Tumor-associated antigens KRR1 and ZRF1 and their cognate autoantibodies may be considered as potential molecular markers for breast cancer.	0.0318	[58]
LUAD	ATP6V0B	miRNA-15a, which could regulate ATPase, H(+) transporting, lysosomal21 kDa, V0 subunit b(ATP6V0B), and miRNA-155, were found to be the most significant.	0.0188	[59]
	RIT1	The results provide a genome-wide map of oncogenic RIT1 functional interactions and identify components of the RAS pathway, spindle assembly checkpoint, and Hippo/YAP1 network as candidate therapeutic targets for RIT1 mutant lung cancer.	0.0384	[41,60]
	MRPS12	MRPS12 functions as a potential oncogene in ovarian cancer and is a promising prognostic candidate.	0.0148	[61]
SKCM	CR1	The HER2 CR1 domain can be antagonized by the clinically used HER2 antibody pertuzumab. Pertuzumab significantly reduced tumorigenicity in TNBC cells expressing circ-HER2/HER2-103.	0.005766	[62]
	EPHA3	EphA3, originally isolated from leukemia and melanoma cells, is currently one of the most promising therapeutic targets, with multiple tumor-promoting effects in multiple cancer types.	0.01137	[63]
	PEG3	Paternal expression gene 3 (PEG3) was the only imprinted gene associated with prognosis of colon cancer patients at the mRNA level, and a high expression of PEG3 mRNA was associated with a poor prognosis.	0.04961	[64]

References

1. Liu, F.; Gai, X.; Wu, Y.; Zhang, B.; Wu, X.; Cheng, R.; Tang, B.; Shang, K.; Zhao, N.; Deng, W.; et al. Oncogenic β -catenin stimulation of AKT2–CAD-mediated pyrimidine synthesis is targetable vulnerability in liver cancer. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2202157119. [CrossRef] [PubMed]
2. Stratton, M. Patterns of somatic mutation in human cancer genomes. *EJC Suppl.* **2008**, *9*, 6. [CrossRef]
3. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. [Special Issue Review] Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef] [PubMed]
4. Wensheng, Z.; Flemington, E.K.; Kun, Z. Driver gene mutations based clustering of tumors: Methods and applications. *Bioinformatics* **2018**, *34*, i404–i411.
5. Mao, Y.; Chen, H.; Liang, H.; Meric-Bernstam, F.; Mills, G.B.; Chen, K. CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. *PLoS ONE* **2013**, *8*, e77945. [CrossRef]
6. Bashashati, A.; Haffari, G.; Ding, J.; Ha, G.; Lui, K.; Rosner, J.; Huntsman, D.G.; Caldas, C.; Aparicio, S.A.; Shah, S.P. DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **2012**, *13*, R124. [CrossRef]
7. Carter, H. Computational Assessment of Somatic Missense Mutations Detected in Tumor Sequencing Studies with Cancer-Specific High-Throughput Annotation of Somatic Mutations (CHASM). Ph.D. Thesis, The Johns Hopkins University, Baltimore, MD, USA, 2012.
8. 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. *Nat. Mach. Intell.* **2021**, *3*, 513–526. [CrossRef]
9. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A.; et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214–218. [CrossRef]
10. Bradner, J.E.; Hnisz, D.; Young, R.A. Transcriptional Addiction in Cancer. *Cell* **2017**, *168*, 629–643. [CrossRef]
11. Baylin, S.B.; Jones, P.A. Epigenetic Determinants of Cancer. *Cold Spring Harb. Perspect. Biol.* **2016**, *8*, a019505. [CrossRef]

12. Creixell, P.; Reimand, J.; Haider, S.; Wu, G.; Shibata, T.; Vazquez, M.; Mustonen, V.; Gonzalez-Perez, A.; Pearson, J.; Sander, C.; et al. Pathway and network analysis of cancer genomes. *Nat. Methods* **2015**, *12*, 615–621.
13. Yin, C.; Cao, Y.; Sun, P.; Zhang, H.; Li, Z.; Xu, Y.; Sun, H. Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration. *Front. Genet.* **2022**, *13*, 884028. [CrossRef]
14. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
15. Gong, L.; Cheng, Q. Exploiting Edge Features in Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
16. Xu, D.; Ruan, C.; Korpeoglu, E.; Kumar, S.; Achan, K. Inductive representation learning on temporal graphs. *arXiv* **2020**, arXiv:2002.07962.
17. Zhu, J.; Yan, Y.; Zhao, L.; Heimann, M.; Akoglu, L.; Koutra, D. Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7793–7804.
18. Zhu, J.; Jin, J.; Loveland, D.; Schaub, M.T.; Koutra, D. On the Relationship between Heterophily and Robustness of Graph Neural Networks. *arXiv* **2021**, arXiv:2106.07767.
19. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 20.
20. Liu, J.; Ong, G.P.; Chen, X. GraphSAGE-Based Traffic Speed Forecasting for Segment Network with Sparse Data. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1755–1766. [CrossRef]
21. Weinstein, J.; Collisson, E.; Mills, G.; Shaw, K.; Ozenberger, B.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.; Network, C. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
22. Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Tamborero, D.; Schroeder, M.P.; Jene-Sanz, A.; Santos, A.; Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **2013**, *10*, 1081–1082. [CrossRef]
23. Hamosh, A.; Scott, A.F.; Amberger, J.; Valle, D.; Mckusick, V.A. Online Mendelian Inheritance In Man (OMIM). *Hum. Mutat.* **2000**, *15*, 57–61. [CrossRef]
24. D'Antonio, M.; Pendino, V.; Sinha, S.; Ciccarelli, F.D. Network of Cancer Genes (NCG 3.0): Integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res.* **2012**, *40*, D978–D983. [CrossRef] [PubMed]
25. Pleasance, E.D.; Cheetham, R.K.; Stephens, P.J.; McBride, D.J.; Humphray, S.J.; Greenman, C.D.; Varela, I.; Lin, M.L.; Ordóñez, G.R.; Bignell, G.R.; et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **2010**, *463*, 191–196. [CrossRef] [PubMed]
26. Imambi, S.; Prakash, K.B.; Kanagachidambaresan, G.R. PyTorch. *Program. Tensorflow Solut. Edge Comput. Appl.* **2021**, 87–104. Available online: https://www.semanticscholar.org/paper/PyTorch-Imambi-Prakash/d668f12be54174141e6197fad737006b7b0c0571 (accessed on accessed on 1 May 2021).
27. Tokheim, C.J.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B.; Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14330–14335. [CrossRef] [PubMed]
28. Chatr-Aryamontri, A.; Breitkreutz, B.J.; Oughtred, R.; Boucher, L.; Heinicke, S.; Chen, D.; Stark, C.; Breitkreutz, A.; Kolas, N.; O'Donnell, L.; et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **2015**, *43*, D470–D478. [CrossRef]
29. Wong, W.C.; Kim, D.; Carter, H.; Diekhans, M.; Ryan, M.C.; Karchin, R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **2011**, *27*, 2147–2148. [CrossRef]
30. Pavlov, Y.L. *Random Forests*; Karelian Research Centre Russian Academy of Sciences: Petrozavodsk, Russia, 1997.
31. Abeywickrama, T.; Cheema, M.A.; Taniar, D. k-Nearest Neighbors on Road Networks: A Journey in Experimentation and In-Memory Implementation. *arXiv* **2016**, arXiv:1601.01549.
32. Cortes, C.; Vapnik, V.N. Support-vector networks. *Mach. Learn.* **2004**, *20*, 273–297. [CrossRef]
33. Zhang, Y.; Liu, J.; Liu, X.; Fan, X.; Hong, Y.; Wang, Y.; Huang, Y.L.; Xie, M.Q. Prioritizing disease genes with an improved dual label propagation framework. *BMC Bioinform.* **2018**, *19*, 47. [CrossRef]
34. Huang, Y.F.; Zhang, Z.; Zhang, M.; Chen, Y.S.; Song, J.; Hou, P.F.; Yong, H.M.; Zheng, J.N.; Bai, J. CUL1 promotes breast cancer metastasis through regulating EZH2-induced the autocrine expression of the cytokines CXCL8 and IL11. *Cell Death Dis.* **2018**, *10*, 2. [CrossRef] [PubMed]
35. Lo, L.H.; Lam, C.Y.; To, J.C.; Chiu, C.H.; Keng, V.W. Sleeping Beauty insertional mutagenesis screen identifies the pro-metastatic roles of CNPY2 and ACTN2 in hepatocellular carcinoma tumor progression. *Biochem. Biophys. Res. Commun.* **2021**, *541*, 70–77. [CrossRef] [PubMed]
36. Sucularli, C.; Arslantas, M. Computational prediction and analysis of deleterious cancer associated missense mutations in DYNC1H1. *Mol. Cell. Probes* **2017**, *34*, 21–29. [CrossRef] [PubMed]
37. Lin, J.; Zhuo, Y.; Yin, Y.; Qiu, L.; Li, X.; Lai, F. Methylation of RILP in lung cancer promotes tumor cell proliferation and invasion. *Mol. Cell. Biochem.* **2021**, *476*, 853–861. [CrossRef]
38. Porta-Pardo, E.; Garcia-Alonso, L.; Hrabe, T.; Dopazo, J.; Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* **2015**, *11*, e1004518. [CrossRef]
39. Bertrand, D.; Chng, K.R.; Sherbaf, F.G.; Kiesel, A.; Chia, B.K.; Sia, Y.Y.; Huang, S.K.; Hoon, D.S.; Liu, E.T.; Hillmer, A.; et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* **2015**, *43*, e44. [CrossRef]
40. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*,7–20. [CrossRef]

41. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **2018**, *173*, 371–385. [CrossRef]
42. Kumar, A.; Masand, N.; Patil, V.M. Understanding Molecular Process and Chemotherapeutics for the Management of Breast Cancer. *Curr. Chem. Biol.* **2021**, *15*, 69–84. [CrossRef]
43. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **2018**, *18*, 696–705. [CrossRef]
44. Craene, B.D.; Berx, G. Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer* **2013**, *13*, 97–110. [CrossRef]
45. Kilpinen, S. Role of ErbB4 in breast cancer. *J. Mammary Gland. Biol. Neoplasia* **2008**, *13*, 259–268.
46. Chen, X.Y.; Zhang, J.; Zhu, J.S. The role of m6A RNA methylation in human cancer. *Mol. Cancer* **2019**, *18*, 1285–1292. [CrossRef]
47. Pützer, B.M.; Gnauck, J.; Kirch, H.C.; Brockmann, D.; Esche, H. A cis-acting element 7 bp upstream of the ESF-1-binding motif is involved in E1A 13S autoregulation of the adenovirus 12 TS2 promoter. *J. Gen. Virol.* **1997**, *78*, 879–891. [CrossRef]
48. Wang, P.; Wang, L.; Sha, J.; Lou, G.; Lu, N.; Hang, B.; Mao, J.H.; Zou, X. Expression and transcriptional regulation of human ATP6V1A gene in gastric cancers. *Sci. Rep.* **2017**, *7*, 3015. [CrossRef]
49. Cui, R.; Yang, L.; Wang, Y.; Zhong, M.; Yu, M.; Chen, B. Elevated Expression of ASXL2 is Associated with Poor Prognosis in Colorectal Cancer by Enhancing Tumorigenesis and Inducing Cell Proliferation. *Cancer Manag. Res.* **2020**, *12*, 10221. [CrossRef]
50. Khan, S.F.; Damerell, V.; Omar, R.; Du Toit, M.; Khan, M.; Maranyane, H.M.; Mlaza, M.; Bleloch, J.; Bellis, C.; Sahm, B.D.; et al. The roles and regulation of TBX3 in development and disease. *Gene* **2020**, *726*, 144223. [CrossRef]
51. Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **2012**, *486*, 346–352. [CrossRef]
52. Wu, X.; Wang, H.; Zhu, D.; Chai, Y.; Wang, J.; Dai, W.; Xiao, Y.; Tang, W.; Li, J.; Hong, L.; et al. USP3 promotes gastric cancer progression and metastasis by deubiquitination-dependent COL9A3/COL6A5 stabilisation. *Cell Death Dis.* **2021**, *13*, 10. [CrossRef]
53. Skoda, A.M.; Simovic, D.; Karin, V.; Kardum, V.; Vranic, S.; Serman, L. The role of the Hedgehog signaling pathway in cancer: A comprehensive review. *Bosn. J. Basic Med. Sci.* **2018**, *18*, 8. [CrossRef]
54. Sontag, J.M.; Nunbhakdi-Craig, V.; Mitterhuber, M.; Ogris, E.; Sontag, E. Regulation of protein phosphatase 2A methylation by LCMT1 and PME-1 plays a critical role in differentiation of neuroblastoma cells. *J. Neurochem.* **2010**, *115*, 1455–1465. [CrossRef] [PubMed]
55. Duan, R.; Du, W.; Guo, W. EZH2: A novel target for cancer treatment. *J. Hematol. Oncol.* **2020**, *13*, 104. [CrossRef] [PubMed]
56. Shan, Z.; Wang, W.; Tong, Y.; Zhang, J. Genome-scale analysis identified NID2, SPARC, and MFAP2 as prognosis markers of overall survival in gastric cancer. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **2021**, *27*, e929558-1. [CrossRef] [PubMed]
57. Tsuda, M.; Noguchi, M.; Kurai, T.; Ichihashi, Y.; Ise, K.; Wang, L.; Ishida, Y.; Tanino, M.; Hirano, S.; Asaka, M.; et al. Aberrant expression of MYD88 via RNA-controlling CNOT4 and EXOSC3 in colonic mucosa impacts generation of colonic cancer. *Cancer Sci.* **2021**, *112*, 5100. [CrossRef]
58. Dyachenko, L.; Havrysh, K.; Lytovchenko, A.; Dosenko, I.; Antoniuk, S.; Filonenko, V.; Kiyamova, R. Autoantibody response to ZRF1 and KRR1 SEREX antigens in patients with breast tumors of different histological types and grades. *Dis. Markers* **2016**, *2016*, 5128720. [CrossRef]
59. Chen, Y.; Teng, L.; Liu, W.; Cao, Y.; Ding, D.; Wang, W.; Chen, H.; Li, C.; An, R. Identification of biological targets of therapeutic intervention for clear cell renal cell carcinoma based on bioinformatics approach. *Cancer Cell Int.* **2016**, *16*, 16. [CrossRef]
60. Vichas, A.; Riley, A.K.; Nkinsi, N.T.; Kamlapurkar, S.; Parrish, P.C.; Lo, A.; Duke, F.; Chen, J.; Fung, I.; Watson, J.; et al. Integrative oncogene-dependency mapping identifies RIT1 vulnerabilities and synergies in lung cancer. *Nat. Commun.* **2021**, *12*, 4789. [CrossRef]
61. Qiu, X.; Guo, D.; Du, J.; Bai, Y.; Wang, F. A novel biomarker, MRPS12 functions as a potential oncogene in ovarian cancer and is a promising prognostic candidate. *Medicine* **2021**, *100*, e24898. [CrossRef]
62. Li, J.; Ma, M.; Yang, X.; Zhang, M.; Luo, J.; Zhou, H.; Huang, N.; Xiao, F.; Lai, B.; Lv, W.; et al. Circular HER2 RNA positive triple negative breast cancer is sensitive to Pertuzumab. *Mol. Cancer* **2020**, *19*, 142. . [CrossRef]
63. Janes, P.W.; Slape, C.I.; Farnsworth, R.H.; Atapattu, L.; Scott, A.M.; Vail, M.E. EphA3 biology and cancer. *Growth Factors* **2014**, *32*, 176–189. . [CrossRef]
64. Zhou, T.; Lin, W.; Zhu, Q.; Renaud, H.; Liu, X.; Li, R.; Tang, C.; Ma, C.; Rao, T.; Tan, Z.; et al. The role of PEG3 in the occurrence and prognosis of colon cancer. *OncoTargets Ther.* **2019**, *12*, 6001–6012. [CrossRef]