Computers in Biology and Medicine 176 (2024) 108568

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

CancerGATE: Prediction of cancer-driver genes using graph attention autoencoders

Seunghwan Jung, Seunghyun Wang, Doheon Lee *

Department of Bio and Brain Engineering, KAIST, Daejeon 34141, Republic of Korea

ARTICLE INFO

Dataset link: https://github.com/sktoyo/CancerGATE

Keywords:
Cancer-driver gene
Graph convolutional network
Attention mechanism
Multi-omics data
Self-supervised learning
Interpretability

ABSTRACT

Discovery of the cancer type specific-driver genes is important for understanding the molecular mechanisms of each cancer type and for providing proper treatment. Recently, graph deep learning methods became widely used in finding cancer-driver genes. However, previous methods had limited performance in individual cancer types due to a small number of cancer-driver genes used in training and biases toward the cancer-driver genes used in training the models. Here, we introduce a novel pipeline, CancerGATE that predicts the cancer-driver genes using graph attention autoencoder (GATE) to learn in a self-supervised manner and can be applied to each of the cancer types. CancerGATE utilizes biological network topology and multi-omics data from 15 types of cancer of 20,079 samples from the cancer genome atlas (TCGA). Attention coefficients calculated in the model are used to prioritize cancer-driver genes by comparing coefficients of cancer and normal contexts. CancerGATE shows a higher AUPRC with a difference ranging from 1.5 % to 36.5 % compared to the previous graph deep learning models in each cancer type. We also show that CancerGATE is free from the bias toward cancer-driver genes used in training, revealing mechanisms of the cancer-driver genes in specific cancer types. Finally, we propose novel cancer-driver gene candidates that could be therapeutic targets for specific cancer types.

1. Introduction

Cancer is a complex disease that arises from dysregulation of cellular processes, resulting from uncontrolled alterations to multiple genes. Among such altered genes, cancer-driver genes, are defined as those directly contributing to cell growth and proliferation [1,2]. Identification of the cancer-driver genes is important, as it allows for the understanding of the underlying mechanisms of cancer and development into appropriate treatment [3,4]. Studies have shown that knowledge of cancer-driver genes enables precision oncology in clinical practice leading to better treatment responses [5–7]. One example would be the usage of trastuzumab in treating HER2-subtype breast cancer, where *HER2* is a cancer-driver gene [8–10].

Recent advances in high-throughput technology and international consortia-driven projects have led to an abundance of data, which was used to discover more about cancer-driver genes. The cancer genome atlas (TCGA) which has collected multi-omics data of various cancer types is the most representative example of an international consortia-driven project [11]. Also, there are remarkable efforts to construct comprehensive catalogs of cancer-driver genes such as the network of cancer genes (NCG) [12], and COSMIC cancer gene census [5]. Stemming from the community effort, researchers worldwide have developed different methods to elucidate the cancer-driver genes.

Initial methods of finding cancer-driver genes keyed on finding statistically significant features of cancer, be it genetic mutations [2], or differential gene expression [13]. Some of these methods include algorithms such as MuSiC2 [14], MutSigCV [15], DESeq2 [16]-based pipelines [13]. Further studies introduced an array of machine learning techniques [17,18], or used derived features from aforementioned features, like clustering of genetic mutations [19], measuring the protein structure change from mutations [20], or measuring the potential functional impact of genetic mutations [21]. However, different cancer-driver genes exhibit different patterns of alteration such as the missense mutation of *PIK3CA* [22] and the overexpression of *IGF2* due to imprinting loss [23]. Therefore, to systemically and comprehensively identify cancer-driver genes, integrating multi-omics data is essential.

Another approach to identifying cancer-driver genes is the utilization of network science techniques, as cancer-driver genes influence various biological pathways through their interactions with other biological entities [24–26]. The functional interactions of cancer-driver genes, including protein–protein interactions (PPIs) and co-expressions, are often rewired in the cancer context [27]. For example, the somatic mutations in *BRAF* contribute to carcinogenesis by affecting

* Corresponding author.
E-mail addresses: sktoyo@kaist.ac.kr (S. Jung), kingsarrow@kaist.ac.kr (S. Wang), dhlee@kaist.ac.kr (D. Lee).

https://doi.org/10.1016/j.compbiomed.2024.108568
Received 7 November 2023; Received in revised form 13 April 2024; Accepted 5 May 2024
Available online 9 May 2024

the PPIs associated with *BRAF* [28]. Another instance involves the altered co-expression patterns of *TP53* in the cancer context [29]. Modeling these interactions as networks reveals that cancer-driver genes typically exhibit high connectivity [30]. These network characteristics of cancer-drive genes can be utilized to find and interpret them. HotNet2 identified cancer gene modules using network diffusion with mutational information [31]; LOTUS identified cancer-driver genes by utilizing mutational features and topology of biological networks [32]; DGCA identified cancer-related differential co-expression of cancer-driver genes in a cancer context [29]; inference the PPIs in cancer context using explainable AI method [33,34]; some other previous studies focused more on biological pathways rather than individual cancer-driver genes [35,36]; or identified cancer driver modules by correlating different modes of omics data [37]. In recent years, network-based models opted into utilizing deep learning [38]. The graph convolutional network (GCN) [39] is an emerging graph deep learning architecture that naturally utilizes both the network topology and features of entities. One of the prominent examples of utilizing GCN is EMOGI by Schulte et al. [38], which utilized GCN architecture with multi-omics data and layer-wise relevance propagation [34] for explainability.

However, these network-based approaches suffer from several limitations. Firstly, they highly depend on the prior knowledge of biological pathways, preventing their generalization to less-studied diseases [35,36]. Additionally, they either rely on the unidimensional data from single omics sources [31] or forcefully incorporate the genetic feature or network topology into a single feature type, potentially causing a loss of information [32,35–37].

Given the heterogeneity across cancer types [2], the model for specific cancer types is crucial. Yet, the number of known cancer-driver genes for individual cancer types is limited (Supplementary Figure S1) [12,40]. This limitation is critical when considering the supervised learning method has been widely used for identifying cancer-driver genes. Also, models trained in a supervised manner tend to identify genes similar to those used in training. These limitations could restrict the possibility of discovering sufficiently novel cancer-driver genes with distinct characteristics in individual cancer types.

Here, we introduce CancerGATE, a novel algorithm to predict cancer type-specific cancer-driver genes using a self-supervised learning approach (Fig. 1). Given the changes to interactions of cancer-driver genes within the cancer context [28,29], we aim to identify cancer-driver genes based on changes of interactions with neighboring genes. CancerGATE employs a graph attention autoencoder (GATE) [41–43] in a self-supervised learning manner to overcome the bias introduced by the limited number of known cancer-driver genes for each cancer type and embed biological context into interactions. From multi-omics data of 20,079 samples from TCGA, we were able to construct an attention-based interaction network for each of the 15 cancer types and their corresponding normal samples. By measuring the attention coefficient differences with neighboring genes in cancer and normal contexts, we were able to prioritize genes with significant coefficient differences as cancer-driver gene candidates and provide the rationale for such changes. CancerGATE outperformed previous methods in predicting cancer-driver genes for individual cancer types and exhibited no bias toward cancer-driver genes used in training. Finally, we propose novel gene candidates identified by CancerGATE to provide insights into the mechanisms of a few cancer types and suggest potential therapeutic targets.

The article is organized as follows: Section 2 details the dataset, algorithms of CancerGATE, the structure and training of the GATE models within CancerGATE, and the analysis methods employed; Section 3 presents an evaluation of the performance of CancerGATE, highlights the advantages of CancerGATE, and provides representative cases illustrating the characteristics of CancerGATE; Section 4 summarizes the key features and benefits of CancerGATE and discusses directions for future work.

2. Materials and methods

2.1. Collection of data

Among the cancer types in TCGA, 15 cancer types have available expression, mutation, and methylation data along with known cancer-driver genes. Consequently, we collected expression, mutation, and methylation data of 20,079 samples from TCGA (Supplementary Table S1). Quantile-normalized and batch-corrected RNA-seq expression datasets from Wang et al. were used for expression [44]. Annotated mutation files of samples were downloaded from TCGA. 450k Illumina bead array files of both cancer and normal samples were used for DNA methylation data.

2.2. Collection of network data

Network used in the study was collected from HumanNet v2 [45], which is used in several network-based studies [46–48]. FN-level data of HumanNet v2 including functional relationships such as PPI, co-expression, co-essentiality, associations by pathway database, associations between protein domain profiles, associations by gene neighborhood in a chromosome, and associations between phylogenetic profiles were used to construct the edges between genes in the network.

2.3. Preprocessing of data

For gene expression, we applied the log transformation to the average Fragments Per Kilobase of transcript per Million ($\log_2(FPKM+1)$) values. For mutation, we filtered ultramutated samples from synapse 1729383(syn1729383) following the preprocessing steps of Schulte et al. [38]. For mutation, variant allele frequencies for each of the 12 types of single nucleotide variants are used [49]. Since TCGA derived the mutation information by comparing matching cancer samples and normal samples, we assigned default values to mutation features for normal samples (Supplementary Table S2). For methylation, we again followed the preprocessing steps of Schulte et al. [38]. We averaged methylation signal value beta (β) within ± 1000 bp range of transcription start sites from GENCODE(v28), for each gene for all samples in the cohort. Methylation values were batch-corrected using ComBat [50] for each cancer type.

The result is a bidirectional network with 11,938 genes and 433,750 edges including self-loop. Min–max normalization was performed for input feature data for expression, mutation, and methylation. Each gene is represented as 14-dimensional data, which is a concatenated vector of gene expression, frequencies of 12 types of mutation, and methylation. Any gene with missing values or not present in the network is discarded.

2.4. Preparing dataset

For each cancer type, the train/test set was constructed using 639 known cancer-driver genes collected from NCG [51] as positives and 1270 non-driver genes collected from Schulte et al. [38] as negatives. An Independent set for additional testing was made with 55 cancer-driver genes from OncoKB [40] as positives, with duplicates from NCG removed for each cancer type, and all remaining genes present in the network not listed in OncoKB or the train/test set as negatives. Cancer types of cancer-driver genes for both NCG and OncoKB were categorized based on the primary tissue site. The statistics of known cancer-driver genes of the 15 cancer types are in Supplementary Table S3.

2.5. Self-supervised learning in CancerGATE

CancerGATE employs the self-supervised learning algorithm, GATE [41–43] for its learning processes (Fig. 1b). GATE is trained to use

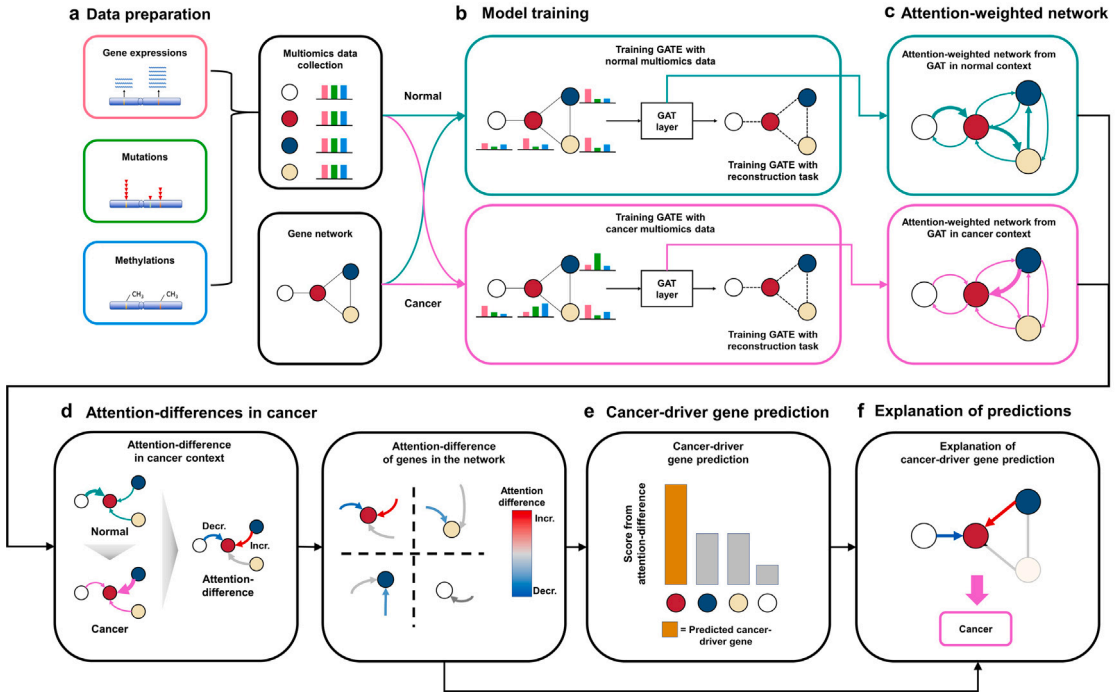

Fig. 1. Overview of CancerGATE. **a.** Data preparation. CancerGATE uses a gene network and 3 types of omics data, expression, mutation, and methylation of cancer and normal samples, respectively. **b.** Attention learning within GATE, which is trained using each normal and cancer multi-omics data. GATE utilizes graph attention network (GAT) layers with multi-omics data to reconstruct gene network topology. **c.** The attention coefficients within GATs are used to construct an attention-weighted network, where the coefficients are represented by the width of the incoming edges for each node. **d.** Attention coefficient difference in the cancer context, highlighting the difference of the coefficients from normal to cancer context. **e.** CancerGATE prioritizes the cancer-driver genes with lower cosine similarity in attention coefficients between cancer and normal contexts. **f.** The difference in attention coefficients provides the basis for understanding the predictions.

the adjacency matrix of the network and feature vectors of nodes to reconstruct the adjacency matrix. The graph attention network (GAT) layers within GATE utilize the feature information of neighboring nodes similarly to that of a convolutional neural network. GAT layers employ an attention mechanism to assign importance to neighboring nodes. GATE learns the attention coefficients for interactions that embed multi-omics features of nodes. Node feature vectors are passed through a dense layer first and then passed to GAT layers. The formula of graph attention as below [41]:

$$\overrightarrow{h_a^{i+1}} = \sigma\left(\sum_{b\in N_a} \alpha_{ab} W \overrightarrow{h_b^i}\right) \tag{1}$$

Here, $\overrightarrow{h_a^{i+1}}$ denotes the hidden vector of node a after passing through GAT layers. σ denotes the activation function, ReLU. The α_{ab} denotes the attention coefficient between node a and node b. W denotes the weight matrix. $\overrightarrow{h_b^i}$ denotes the hidden vector of node b which is the input of GAT layers. The node b is the group of neighbors, N_a, of node a in the network, including itself.

The formula for attention coefficients is below:

$$\alpha_{ab} = \text{softmax}(e_{ab}), \quad b \in N_a \tag{2}$$

$$e_{ab} = \text{LeakyReLU}\left(\vec{a}^T \left[W\overrightarrow{h_a^i} \parallel W\overrightarrow{h_b^i}\right]\right) \tag{3}$$

The attention coefficient α_{ab} is a softmax value of e_{ab}, the importance value of the hidden vector of node b to node a. To calculate the softmax, it sums all importance of neighbors of node a including itself. The sum of the importance of the neighboring nodes is the denominator in the softmax function. e_{ab} is calculated as a product between the attention weight matrix $\vec{a}$ and concatenation of hidden vectors of $\overrightarrow{h_a^i}$ and $\overrightarrow{h_b^i}$. LeakyReLU is the activation function to calculate e_{ab}.

The objective function of GATE is the reconstruction of the adjacency matrix of the network. We followed the reconstruction method and loss function from Kipf et al. [42]

$$\hat{A} = \sigma(ZZ^T)$$
$$Z = \text{GAT}(X, A) \tag{4}$$

Here, A denotes the adjacency matrix of the network. $\hat{A}$ denotes the reconstructed adjacency matrix. X denotes the input feature matrix. Z is the resultant embedding matrix from the embedding layer and GAT layers. The formula of the loss function is below:

$$\text{Loss} = \frac{1}{2}(\text{Loss}_{pos} + \text{Loss}_{neg}) \tag{5}$$

$$\text{Loss}_{pos} = -\frac{1}{E}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i} e_{ij}\log\hat{e}_{ij} \tag{6}$$

$$\text{Loss}_{neg} = -\frac{1}{N^2-E}\sum_{i=1}^{N}\sum_{j\notin\mathcal{N}_i}(1-e_{ij})\log(1-\hat{e}_{ij}) \tag{7}$$

Loss is calculated as an average of the cross-entropy of edges and non-edges. Edges represent elements with value 1 in the adjacency matrix, and non-edges 0. Loss_{pos} represents a loss of cross-entropy of edges in original and reconstructed adjacency matrix (6) and Loss_{neg}, non-edges (7). E denotes the number of edges in the network and N denotes the number of nodes in the network. e_{ij} denotes the value of the edge between node i and node j of the adjacency matrix, and $\hat{e}_{ij}$ denotes the value of the reconstructed edge between node i and node j of the reconstructed adjacency matrix.

2.6. CancerGATE scoring

The trained GATE provides attention coefficients for edges (Fig. 1c). The attention coefficients are calculated from each attention head in GAT layers. Therefore, the total number of attention coefficients in GATE is obtained by multiplying the number of GAT layers by the number of attention heads. We used the sum of attention coefficients

from all attention heads in the GAT layers. The sum of the attention coefficient becomes the edge weight of the directed network.

GATE is trained separately for cancer and normal samples, resulting in two networks with attention coefficient weights. The score using the difference in attention coefficient between cancer and normal contexts is calculated by the following formula (Fig. 1d, e).

$$\text{Score}_a = 1 - \frac{\overrightarrow{E_{a,C}} \cdot \overrightarrow{E_{a,N}}}{\|\overrightarrow{E_{a,C}}\| \|\overrightarrow{E_{a,N}}\|} \quad (8)$$

Here, Score_a denotes the CancerGATE score of gene a which denotes reversed cosine similarity. $\overrightarrow{E_{a,N}}$ and $\overrightarrow{E_{a,C}}$ denote vectors that represent the attention coefficient for the neighbor genes of gene a in the normal and cancer network, respectively. Cosine similarity measures the similarity between attention coefficients for a gene between normal and cancer contexts. All elements of $\overrightarrow{E_{a,N}}$ and $\overrightarrow{E_{a,C}}$ are non-negative, and thus cosine similarity value ranges from 0 to 1. CancerGATE score is designed to reflect the dissimilarity of $\overrightarrow{E_{a,N}}$ and $\overrightarrow{E_{a,C}}$. Genes with significantly high scores (one-tailed t-test, p-value < 0.05) are considered potential cancer-driver gene candidates.

2.7. Performance measure

To measure the model performance, we calculated the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area under the Precision-Recall Curve (AUPRC). AUROC represents the area under the plot of the true positive rate against the false positive rate. AUPRC represents the area under the plot of precision versus recall. These metrics evaluate a model's ability to accurately classify positives and negatives.

2.8. Setting hyperparameters

The dimension size of the embedding layers was chosen as 128. The dimensions of the two GAT layers were each chosen as 300 and 100. The number of attention heads of the GAT layers was chosen through grid search (Supplementary Figure S2). The dropout rate of attention weight and feature of GATE layers is 0.2. Dropout layer with a dropout rate of 0.5 was used. ADAM optimizer [52] was used with a weight decay of 0.005 and a learning rate of 0.001. Dimension size of GAT layers, optimizer choice, weight decay, and learning rate were chosen following the hyperparameters of Schulte et al. [38]. The hyperparameters of CancerGATE are detailed in Supplementary Table S4. The learning results of the GATE models, using the selected hyperparameters, are detailed in Supplementary Table S5.

2.9. Essentiality analysis

To see if cancer-driver gene candidates predicted by CancerGATE are functionally impactful, we labeled some genes 'essential' by measuring the essentiality of candidate genes for cell growth. Project Achilles [53] calculated the essentiality of genes across 764 cell lines utilizing CERES scores [54]. If a gene has a negative CERES score, the gene results in a reduction in cell growth. We selected the essential genes with CERES score < -0.5 and genes that affect more than the average number of cell lines affected by a single gene. Of the 764 cell lines in Project Achilles, on average 116 cell lines were affected by a single gene. We labeled 2844 genes as essential for total cell lines (for the statistics of the individual cancer types, see Supplementary Table S6).

3. Results

3.1. Performance comparison on cancer-driver gene prediction

We hypothesized that CancerGATE could outperform supervised learning methods in predicting cancer-driver genes of each cancer type, which typically have a limited number of known cancer-driver genes. We implemented GCN models, GAT [41] and EMOGI [38] to compare with (Detailed parameters of implemented models in Supplementary Table 4). We adopted AUPRC as the main measure, as AUPRC is a more reliable measure than AUROC in a dataset with much fewer positives than negatives. We validated each model ten times each with 5-fold cross-validation.

Table 1
Performance comparison on the test set with AUPRC.

Cancer types[a]	EMOGI	GAT	CancerGATE
BLCA	0.1890 ± 0.0327	0.3190 ± 0.0262	**0.3493 ± 0.0839**
BRCA	0.4836 ± 0.0702	0.3515 ± 0.0209	**0.5197 ± 0.0693**
CESC	0.0063 ± 0.0000	0.1327 ± 0.0256	**0.5296 ± 0.2793**
COAD	0.1983 ± 0.0232	**0.3278 ± 0.0224**	0.2484 ± 0.0596
ESCA	0.1088 ± 0.0041	0.2806 ± 0.0471	**0.3073 ± 0.0838**
HNSC	0.1945 ± 0.0421	0.2546 ± 0.0127	**0.2698 ± 0.0666**
KIRC	0.0176 ± 0.0023	0.2005 ± 0.0140	**0.2394 ± 0.1298**
KIRP	0.0182 ± 0.0025	0.0621 ± 0.0069	**0.2501 ± 0.1441**
LIHC	0.0645 ± 0.0025	0.1422 ± 0.0151	**0.3236 ± 0.1205**
LUAD	0.2838 ± 0.0980	0.3004 ± 0.0239	**0.3834 ± 0.0751**
LUSC	0.0106 ± 0.0011	0.0988 ± 0.0353	**0.2031 ± 0.1219**
PRAD	0.0496 ± 0.0063	0.1645 ± 0.0486	**0.5298 ± 0.1173**
STAD	0.0669 ± 0.0149	0.2098 ± 0.0469	**0.3353 ± 0.0904**
THCA	0.0742 ± 0.0408	0.1656 ± 0.0362	**0.2764 ± 0.1024**
UCEC	0.1018 ± 0.0056	0.1944 ± 0.0168	**0.4040 ± 0.0789**

Bold values highlight the metric of best performing models.
[a] BLCA, Bladder Urothelial Carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma; HNSC, Head and Neck squamous cell carcinoma; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; PRAD, Prostate adenocarcinoma; STAD, Stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma.

We compared the performances of CancerGATE and other models in 15 types of cancer. We computed AUPRC on the test set (Table 1 and Supplementary Figure S3), the independent set (Table 2), and AUROC on the test set (Supplementary Figure S4 and Supplementary Table S7), independent set (Supplementary Table S8). In the test set, CancerGATE achieved the best AUPRC metric of the test set on every individual cancer type, except for colorectal adenocarcinoma (COAD). In the independent set, CancerGATE showed the best performances in most cancer types, except for head-neck squamous cell carcinoma (HNSC), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) (Table 2). Due to the differences in the number of genes labeled non-driver in the independent set, AUPRC values are much lower. Although CancerGATE did not rank first in most cancer types based on AUROC, it performed comparably with the GAT model (Supplementary Figure S4 and Supplementary Table S7, 8).

CancerGATE achieved the best AUPRC performances in individual cancer-type models which had significant disparity between the number of positive and negative labels. This suggests that CancerGATE could reliably predict the cancer-driver genes in individual cancer types, despite the limited number of known cancer-driver genes.

3.2. Ablation study

To better assess the performance, we investigated the contribution of different components of CancerGATE. We evaluated two aspects of performance; the network reconstruction ability of the GATE models, and the effectiveness in predicting cancer-driver genes. Performances of several permutations to components of the embedding layer and each multi-omics data were measured. A detailed list of permutations is in Supplementary Table S9.

The performance comparison of CancerGATE and permutations to its components is in Supplementary Table S10. We observed that the embedding layer has a significant impact on performance in both

Table 2
Performance comparison on the independent set with AUPRC.

Cancer types	EMOGI	GAT	CancerGATE
BLCA	0.0017 ± 0.0004	0.0045 ± 0.0002	**0.0080 ± 0.0013**
BRCA	0.0012 ± 0.0002	0.0020 ± 0.0004	**0.0023 ± 0.0010**
CESC	0.0008 ± 0.0000	0.0048 ± 0.0008	**0.0175 ± 0.0072**
COAD	0.0018 ± 0.0001	0.0037 ± 0.0005	**0.0082 ± 0.0022**
ESCA	0.0014 ± 0.0002	0.0045 ± 0.0003	**0.0078 ± 0.0017**
HNSC	0.0019 ± 0.0004	**0.0038 ± 0.0003**	0.0035 ± 0.0010
KIRC	0.0008 ± 0.0001	0.0023 ± 0.0003	**0.0029 ± 0.0007**
KIRP	0.0011 ± 0.0001	0.0034 ± 0.0007	**0.0265 ± 0.0123**
LIHC	0.0018 ± 0.0000	0.0033 ± 0.0003	**0.0044 ± 0.0014**
LUAD	0.0016 ± 0.0005	0.0025 ± 0.0002	**0.0039 ± 0.0017**
LUSC	0.0013 ± 0.0002	0.0058 ± 0.0011	**0.0143 ± 0.0083**
PRAD	0.0033 ± 0.0003	0.0066 ± 0.0005	**0.0087 ± 0.0016**
STAD	0.0019 ± 0.0003	**0.0036 ± 0.0005**	0.0028 ± 0.0009
THCA	0.0008 ± 0.0001	0.0035 ± 0.0003	**0.0082 ± 0.0020**
UCEC	0.0018 ± 0.0001	**0.0069 ± 0.0005**	0.0052 ± 0.0015

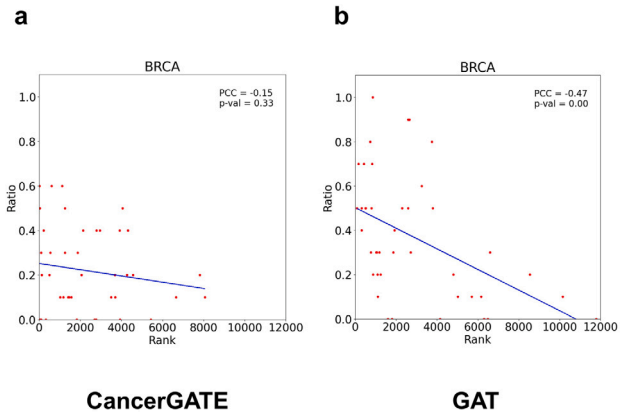

Fig. 2. Correlation between the rank of scores and the ratio of known cancer-driver genes. The horizontal axis is the predicted rank of the known cancer-driver genes, and the vertical axis is the ratio of the known cancer-driver genes in genes that show high similarity in features with cancer-driver genes. Details in the text. **a.** The correlation of genes ranked with CancerGATE in BRCA, **b.** The correlation of genes ranked with the GAT model in BRCA. Pearson correlation (PCC) was applied as the correlation measure.

network reconstruction and cancer-driver gene prediction. Also, we confirmed that using all omics data improves the reconstruction performance in all cancer types. Both the multi-omics data and the components of CancerGATE play crucial roles in the accurate prediction of cancer-driver genes.

3.3. Independence to the prior knowledge

Next, we hypothesized that CancerGATE would show no bias towards known cancer-driver genes. To show this, we ranked all genes based on CancerGATE scores and calculated the correlation between the calculated rank of known cancer-driver genes in the test set and the ratio of known cancer-driver genes in the training/test set that have similar feature values of genes. If there was bias, then feature similarity would be represented and show some levels of correlation.

As hypothesized, no significant correlations were found in CancerGATE results for every cancer type (Fig. 2a, Supplementary Figure S5). CancerGATE was able to discover cancer-driver genes that have distinct genetic features. However, GAT models that were trained under the supervised learning schema showed correlations (Fig. 2b, Supplementary Figure S6). Fig. 2 presents the correlation results for BRCA, which possesses strong statistical power due to having the most KCDGs among cancer types. These results suggest that if there exists a cancer-driver gene dissimilar to cancer-driver genes used in training, supervised learning models would struggle to find it.

3.4. Selection of cancer-driver gene candidates

Driver gene candidates were needed for further analysis of CancerGATE. Ten separate CancerGATE models each with different initial weights were trained while keeping all other settings unchanged, and from the ten models, the model with the best AUROC and AUPRC performance in the test set was chosen. cancer-driver gene candidates with significantly high scores (p-value < 0.05) were chosen (Supplementary Figure S7). The candidate genes for each cancer type are listed in Supplementary Table S11. Further analyses use cancer-driver genes predicted from this model, and newly predicted cancer-driver genes from the list are referred to as novel cancer-driver gene candidates.

3.5. Explainability of CancerGATE

The difference in attention coefficients between cancer and normal contexts represents the loss or gain of functional interactions between genes and their neighbors. Because we trained CancerGATE to embed similar hidden vectors between gene pairs as having functional interactions, pairs of genes with low feature similarity would have high attention coefficients to embed similar hidden vectors. Finally, we hypothesized that an increase in attention coefficient in cancer would mean the loss of functional interaction and vice versa.

Because manual inspection for all known cancer-driver genes and their neighbors in biological networks in every cancer type is unfeasible (Supplementary Table S12), we further inspected neighboring genes of *DAXX*, a known cancer-driver gene, which had the highest CancerGATE score in breast invasive carcinoma (BRCA) and thyroid carcinoma (THCA) in cancer context. Other genes labeled as pan-cancer in NCG show similar mechanisms in multiple cancer types, so those genes were excluded from the inspection. *DAXX*, death domain associated protein, is a favorable prognostic marker in BRCA [55–57] and a known to be mutated in THCA [58,59].

TP53 and *CASP10* were neighboring genes of *DAXX* with significant differences in attention coefficients in BRCA and THCA (Fig. 3a, b). *TP53* and *CASP10* were enriched in cellular mechanism of apoptosis (Fig. 3c, d). *DAXX* is known to regulate *TP53* through physical interactions [60,61] and mutation of *TP53* in BRCA and THCA is known to disrupt interactions between *DAXX* and p53 [56,62]. A gene that showed a decrease in attention was *FASLG*. *FASLG* is, along with *DAXX*, a gene associated with apoptosis. There are reports of association of *FASLG* with BRCA and THCA [63,64] and *FASLG* is known to be upregulated with *DAXX* in BRCA [63]. Both examples suggest that an increase in attention means a loss of interaction and a decrease, a gain of interaction.

We also identified cancer type-specific differences in attention coefficient, including *PAX3* as a neighbor with decreased attention and *DNMT3A* as a neighbor with increased attention in THCA, both supported by previous reports [65,66]. *DAXX* negatively regulates *PAX3* [67] and interacts with *DNMT3A* [68]. Interestingly, the loss of negative regulation seems to result in a decrease in attention. Another example is *HAT1* in BRCA (Fig. 3a) which, while not statistically significant, showed up as a neighboring gene with one of the top ten attention increases. *HAT1* is known to be highly expressed in BRCA [69] and is involved in histone assembly with *DAXX* [70]. While there were no known studies on the interaction of *DAXX* and *HAT1* in BRCA, these attention differences may suggest perturbation of interactions specific to BRCA.

3.6. Driver gene candidates as therapeutical targets

We also hypothesized that cancer-driver gene candidates are significantly enriched with cancer-specific driver genes and cancer-specific therapeutic targets. We measured the sensitivity of each with previously known cancer genes from NCG and 109 therapeutic targets were from the TARGET [71], a database of therapeutic target genes

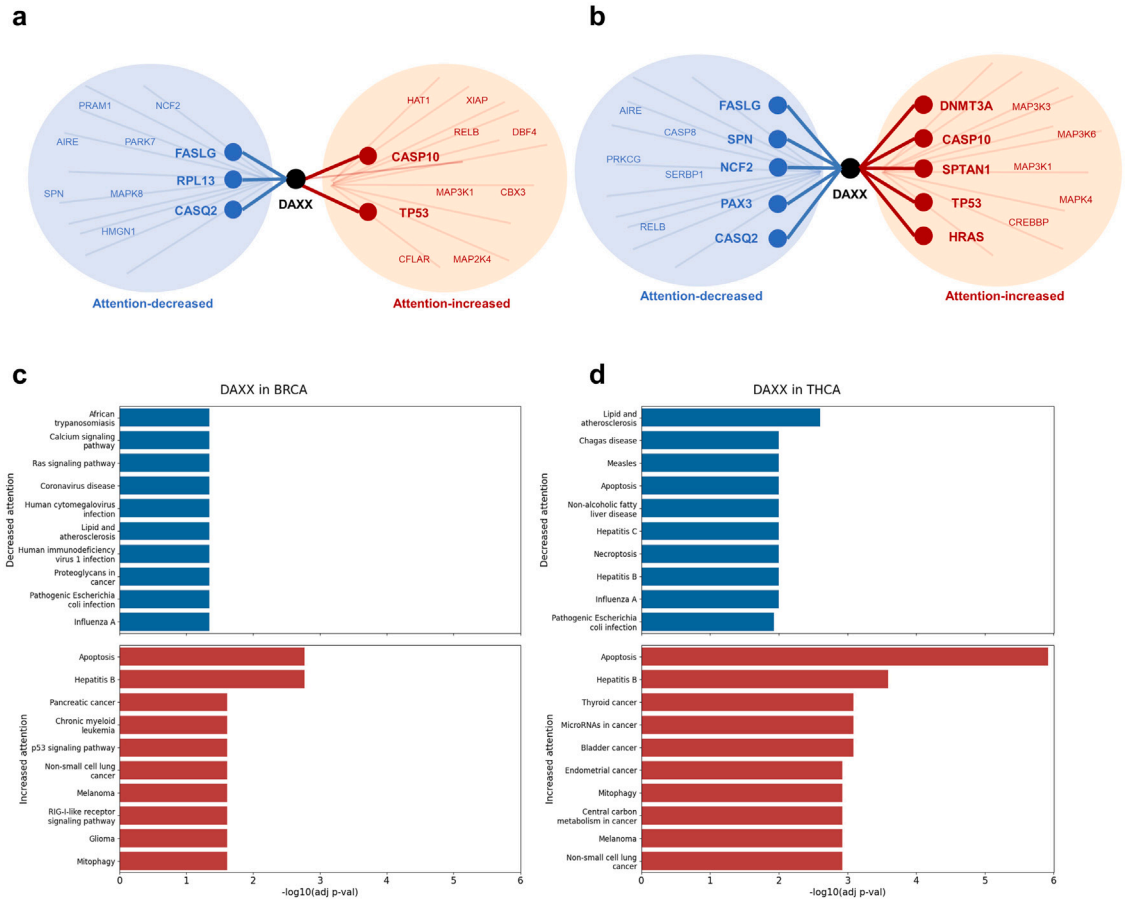

Fig. 3. The interpretability of CancerGATE to oncogenic mechanisms of known cancer-driver genes using attention differences. **a.** Neighbors of *DAXX* in BRCA with attention differences. **b.** Neighbors of *DAXX* in THCA with attention differences, **c.** KEGG enrichment test of significantly different attention neighbors of *DAXX* in BRCA, **d.** KEGG enrichment test of significantly different attention neighbors of *DAXX* in THCA. Every adjusted *p*-value is less than 0.05. Blue/red color indicates decreased/increased attention, respectively. Emphasized nodes/edges indicate the neighbors with significant differences in attention coefficients.

in cancer (Supplementary Figure S8 a). Among predicted genes, the recall of known cancer-driver genes from NCG was from 18.7% in Lung adenocarcinoma (LUAD) to 69.2% in Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and the average was 34.4% for all cancer types. The recall of therapeutic targets in individual cancer types was from 35.8% in LUAD to 58.72% in UCEC and THCA, and the average was 50.6%. We also confirmed significant enrichment of known cancer-driver genes and therapeutic targets among cancer-driver gene candidates for every cancer type (p-value < 0.05, hypergeometric test, Supplementary Figure S8 b, c). Consequently, CancerGATE demonstrated its ability to identify not only previously known cancer-driver genes but also potential therapeutic targets.

3.7. Functional analysis of novel cancer-driver gene candidates

We further investigated the functional impact of novel cancer-driver gene candidates, excluding known cancer-driver genes, by analyzing their characteristics in the biological network and assessing the enrichment of 'essential' genes among these candidates. We also observed significant correlations between the ranks of CancerGATE score-ranked genes and the ranks of genes ranked by number of interactions with known cancer-driver genes (Fig. 4, Supplementary Figure S9), implying that novel cancer-driver candidates tended to have more interactions with known cancer-driver genes. This result hints at a potential co-functionality between novel cancer-driver gene candidates and known cancer-driver genes, suggesting that targeting these genes could impact cancer cell lines through these interactions. Additionally, cancer-type-specific 'essential' genes [53], which significantly affect the cancer cell survival in the different cancer cell lines through loss-of-function experiments, were significantly more enriched in novel cancer-driver gene candidates than in non-candidates across all cancer types (Supplementary Table S13). These results indicate the possibility of the candidates being potential therapeutic targets.

3.8. Prediction of novel cancer-driver candidates

Finally, we manually inspected predicted genes to assess the potential of these genes as novel cancer-driver genes and therapeutic targets. As mentioned, it is infeasible to inspect all candidates, so focused on the top 30 cancer-driver gene candidates of BRCA, where CancerGATE exhibited the best performance among the cancer types (Table 3). The complete candidate gene list for each cancer type is in Supplementary Table S14. The top 10 enrichment pathways of these 30 genes were pathways in cancer, Hepatitis B, prostate cancer, chronic myeloid leukemia, Kaposi sarcoma-associated herpesvirus infection, human T-cell leukemia virus 1 infection, FoxO signaling pathway, Epstein-Barr virus infection, viral carcinogenesis, and pancreatic cancer. We found that 3 genes, *NFKB1*, *CHUK*, and *RELA* were enriched in 9 pathways of the top 10 enriched pathways. These genes are subunits (*NFKB1*, *RELA*) or regulators (*CHUK*) of the transcription factor, NF-κB. All three genes are related to the progress of breast cancer [72–74]. *RELA* is previously reported to be associated with inhibition of triple-negative breast cancer growth [75] and *CHUK*, drug resistance in BRCA [76]. These genes may potentially be therapeutic targets in BRCA.

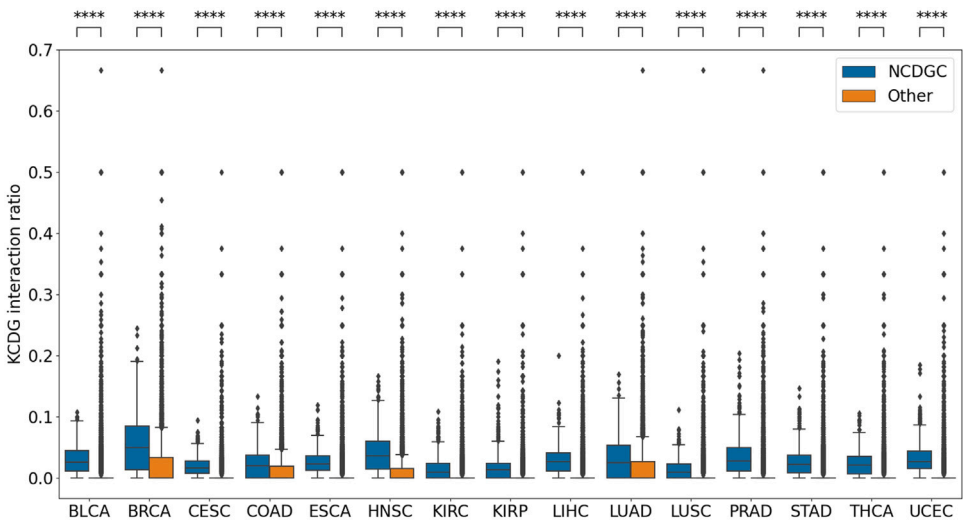

Fig. 4. Novel cancer-driver gene candidates had significantly more interactions with known cancer-driver genes than others. ****: p <= 1.00e−04. NCDGC, novel cancer-driver gene candidate. KCDG, known cancer-driver gene.

Table 3
Analysis of the top 30 cancer-driver gene candidates of BRCA by CancerGATE. NCG, OncoKB, and TARGET columns denote whether the gene has been found in BRCA for each database. Pancancer_NCG and pancancer_OncoKB columns denote that the gene is labeled as pan-cancer in each database. NCDGC denotes whether the gene can be found in NCG, OncoKB, including pan-cancer data, and TARGET. #CancerGATE, #EMOGI, and #GAT columns denote gene's ranking in each model.

Gene	#CancerGATE	#EMOGI	#GAT	NCG	OncoKB	TARGET	Essential	pancancer_NCG	pancancer_OncoKB	NCDGC
NFKBIA	1	133	240	0	0	1	0	0	0	0
EGFR	2	5423	527	1	0	1	0	1	1	0
MAPK8	3	61	282	0	0	0	0	1	0	0
JUN	4	143	411	0	0	0	0	1	0	0
PPP2CA	5	449	568	0	0	0	1	0	0	1
PRKCA	6	190	430	0	0	0	0	0	0	1
PIK3CG	7	44	215	0	0	0	0	0	0	1
RAC1	8	341	680	0	0	0	1	1	0	0
NFKB1	9	260	377	0	0	0	0	0	0	1
HRAS	10	159	153	1	0	1	0	1	1	0
RELA	11	289	380	0	0	0	0	0	0	1
DAXX	12	151	391	1	0	0	0	1	0	0
PIK3CA	13	34	69	1	1	1	1	1	1	0
SUMO1	14	11 616	4864	0	0	0	0	0	0	1
MAX	15	435	619	0	0	0	1	1	0	0
CDK1	16	79	1496	0	0	0	1	0	0	1
HDAC1	17	800	1422	0	0	0	0	0	0	1
CHUK	18	49	239	0	0	0	0	0	0	1
PLK1	19	1028	4094	0	0	0	1	0	0	1
CDKN1A	20	383	502	0	0	1	0	1	0	0
PIK3R1	21	170	108	1	0	1	0	1	0	0
MYC	22	327	824	1	0	1	1	1	0	0
AKT1	23	269	248	1	1	1	0	1	1	0
MDM2	24	735	722	1	0	1	0	1	1	0
CDK2	25	1537	694	0	0	0	1	0	0	1
CREBBP	26	320	477	1	0	1	0	1	0	0
IKBKB	27	117	210	0	0	0	0	1	0	0
EP300	28	393	486	1	0	0	1	1	0	0
CCND1	29	159	319	1	0	1	1	1	0	0
TGFBR1	30	806	365	0	0	0	0	0	0	1

4. Discussion

Finding cancer-driver genes is critical for furthering the treatment and diagnosis of cancer; many methods have been developed ranging technique-wise from statistical tests to deep learning, or data-wise from single-omics to multi-omics. Statistical methods faced the problem of limited interpretation, and deep learning, especially supervised learning methods, showed limitations due to the lack of available information if looked at from the resolution of individual cancer types.

In this research, we developed CancerGATE, a cancer-driver gene prediction model. CancerGATE adopts an attention mechanism and self-supervised learning to overcome the limitations of previous works. We found cancer-driver genes using the differences in attention coefficient between cancer and normal contexts. As far as we know, CancerGATE is the first model that directly utilizes attention coefficients for prediction and analysis. We confirmed that CancerGATE had outperformed the previous methods in individual cancer types with a limited number of cancer-driver gene data. CancerGATE also showed that it is free from biases arising from the nature of supervised learning. Using CancerGATE, we identified potential therapeutic targets for specific cancer types. We also elucidated the rationale behind the predictions by analyzing changed relationships in cancer contexts. Finally, through CancerGATE, we identified novel cancer-driver gene candidates that can affect cancer cell growth or support existing therapies.

CancerGATE is the first initiative to utilize attention coefficients for prediction, outperforming the previous models and elucidating the

mechanism of the cancer-driver gene through specific cancer interactions, such as *DAXX* and *DNMT3A* in THCA. However, its performance was not superior across all cancers, notably in COAD. Currently, the application of attention coefficients for prediction and interpretation is novel, suggesting potential areas for enhancement in their biomedical application. For further research, we expect to apply CancerGATE as a general method for identifying disease-related genes, particularly Parkinson's disease where known gene information is limited. This approach is promising given CancerGATE's ability to operate independently of prior disease knowledge.

Funding

This work was supported by the Ministry of Science and ICT through the National Research Foundation (RS-2023-00262747). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CRediT authorship contribution statement

Seunghwan Jung: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Seunghyun Wang:** Writing – review & editing, Visualization, Investigation. **Doheon Lee:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

The source code and dataset used in this research are available at https://github.com/sktoyo/CancerGATE.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108568.

References

[1] Michael R. Stratton, Peter J. Campbell, P. Andrew Futreal, The cancer genome, Nature 458 (7239) (2009) 719–724.
[2] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, Kenneth W. Kinzler, Cancer genome landscapes, Science 339 (6127) (2013) 1546–1558.
[3] Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, et al., Comprehensive characterization of cancer driver genes and mutations, Cell 173 (2) (2018) 371–385.
[4] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, et al., A compendium of mutational cancer driver genes, Nat. Rev. Cancer (2020) 1–18.
[5] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, Simon A. Forbes, The COSMIC cancer gene census: describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (11) (2018) 696–705.
[6] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L. Baehner, Michael G. Walker, Drew Watson, Taesung Park, et al., A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (27) (2004) 2817–2826.
[7] Richard G. Gray, Philip Quirke, Kelly Handley, Margarita Lopatin, Laura Magill, Frederick L. Baehner, Claire Beaumont, Kim M. Clark-Langone, Carl N. Yoshizawa, Mark Lee, et al., Validation study of a quantitative multigene reverse transcriptase–polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer, J. Clinical Oncol. 29 (35) (2011) 4611–4619.
[8] Dihua Yu, Mien-Chie Hung, Overexpression of ErbB2 in cancer and ErbB2-targeting strategies, Oncogene 19 (53) (2000) 6115–6121.
[9] José Baselga, Debasish Tripathy, John Mendelsohn, Sharon Baughman, Christopher C. Benz, Lucy Dantis, Nancy T. Sklarin, Andrew D. Seidman, Clifford A. Hudis, Jackie Moore, et al., Phase II study of weekly intravenous recombinant humanized anti-p185her2 monoclonal antibody in patients with HER2/neu-overexpressing metastatic breast cancer., J. Clinical Oncol. 14 (3) (1996) 737–744.
[10] Kate McKeage, Caroline M. Perry, Trastuzumab: a review of its use in the treatment of metastatic breast cancer overexpressing HER2, Drugs 62 (1) (2002) 209–243.
[11] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, Cancer Genome Atlas Research Network, et al., The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113.
[12] Lisa Dressler, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, Lucia Montorsi, Neshika Wijewardhane, Dimitra Repana, Joel Nulsen, et al., Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (NCG) resource, Genome Biol. 23 (1) (2022) 1–22.
[13] Cheng Wang, Yayun Gu, Erbao Zhang, Kai Zhang, Na Qin, Juncheng Dai, Meng Zhu, Jia Liu, Kaipeng Xie, Yue Jiang, et al., A cancer-testis non-coding RNA LIN28B-AS1 activates driver gene LIN28B by interacting with IGF2BP1 in lung adenocarcinoma, Oncogene 38 (10) (2019) 1611–1624.
[14] Nathan D. Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C. Wendl, William Schierding, Daniel C. Koboldt, Thomas B. Mooney, Matthew B. Callaway, David Dooling, Elaine R. Mardis, et al., MuSiC: identifying mutational significance in cancer genomes, Genome Res. 22 (8) (2012) 1589–1598.
[15] Michael S. Lawrence, Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, Gad Getz, Discovery and saturation analysis of cancer genes across 21 tumour types, Nature 505 (7484) (2014) 495–501.
[16] Michael I. Love, Wolfgang Huber, Simon Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 1–21.
[17] Collin J. Tokheim, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, Rachel Karchin, Evaluating the evaluation of cancer driver genes, Proc. Natl. Acad. Sci. 113 (50) (2016) 14330–14335.
[18] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, Wayne Xu, Applications of support vector machine (SVM) learning in cancer genomics, Cancer Genom.-Proteom. 15 (1) (2018) 41–51.
[19] David Tamborero, Abel Gonzalez-Perez, Nuria Lopez-Bigas, OncodriveclUSt: exploiting the positional clustering of somatic mutations to identify cancer genes, Bioinformatics 29 (18) (2013) 2238–2244.
[20] Beifang Niu, Adam D. Scott, Sohini Sengupta, Matthew H. Bailey, Prag Batra, Jie Ning, Matthew A. Wyczalkowski, Wen-Wei Liang, Qunyuan Zhang, Michael D. McLellan, et al., Protein-structure-guided discovery of functional mutations across 19 cancer types, Nat. Genet. 48 (8) (2016) 827–837.
[21] Abel Gonzalez-Perez, Nuria Lopez-Bigas, Functional impact bias reveals cancer drivers, Nucleic Acids Res. 40 (21) (2012) e169.
[22] B. Karakas, K.E. Bachman, B.H. Park, Mutation of the PIK3ca oncogene in human cancers, Br. J. Cancer 94 (4) (2006) 455–459.
[23] Aldona Kasprzak, Agnieszka Adamek, Insulin-like growth factor 2 (IGF2) signaling in colorectal cancer—from basic research to potential clinical applications, Int. J. Molecular Sci. 20 (19) (2019) 4915.
[24] Alan Ashworth, Christopher J. Lord, Jorge S. Reis-Filho, Genetic interactions in cancer progression and treatment, Cell 145 (1) (2011) 30–38.
[25] Mark M. Moasser, The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis, Oncogene 26 (45) (2007) 6469–6487.
[26] Keith T. Flaherty, Igor Puzanov, Kevin B. Kim, Antoni Ribas, Grant A. McArthur, Jeffrey A. Sosman, Peter J. O'Dwyer, Richard J. Lee, Joseph F. Grippo, Keith Nolop, et al., Inhibition of mutated, activated BRAF in metastatic melanoma, N. Engl. J. Med. 363 (9) (2010) 809–819.
[27] Eiru Kim, Lance C. Novak, Chenchu Lin, Medina Colic, Lori L. Bertolet, Veronica Gheorghe, Christopher A. Bristow, Traver Hart, Dynamic rewiring of biological activity across genotype and lineage revealed by context-dependent functional interactions, Genome Biol. 23 (1) (2022) 140.
[28] Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R. Martin, Ruisheng Wang, Jin Huang, Tong Hao, et al., Comprehensive characterization of protein–protein interactions perturbed by disease mutations, Nat. Genet. 53 (3) (2021) 342–353.
[29] Andrew T. McKenzie, Igor Katsyv, Won-Min Song, Minghui Wang, Bin Zhang, DGCA: a comprehensive r package for differential gene correlation analysis, BMC Syst. Biol. 10 (1) (2016) 1–25.
[30] Takeshi Hase, Samik Ghosh, Sucheendra K. Palaniappan, Hiroaki Kitano, Cancer network medicine, Netw. Med. (2017) 294–323.
[31] Mark D.M. Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al., Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes, Nat. Genet. 47 (2) (2015) 106–114.

[32] Olivier Collier, Véronique Stoven, Jean-Philippe Vert, LOTUS: A single-and multitask machine learning algorithm for the prediction of cancer driver genes, PLoS Comput. Biol. 15 (9) (2019) e1007381.

[33] Philipp Keyl, Michael Bockmayr, Daniel Heim, Gabriel Dernbach, Grégoire Montavon, Klaus-Robert Müller, Frederick Klauschen, Patient-level proteomic network prediction by explainable artificial intelligence, NPJ Precis. Oncol. 6 (1) (2022) 35.

[34] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, Wojciech Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.

[35] Marta Paczkowska, Jonathan Barenboim, Nardnisa Sintupisut, Natalie S. Fox, Helen Zhu, Diala Abd-Rabbo, Miles W. Mee, Paul C. Boutros, Jüri Reimand, Integrative pathway enrichment analysis of multivariate omics data, Nat. Commun. 11 (1) (2020) 1–16.

[36] Lieven P.C. Verbeke, Jimmy Van den Eynden, Ana Carolina Fierro, Piet Demeester, Jan Fostier, Kathleen Marchal, Pathway relevance ranking for tumor samples through network-based data integration, PLoS One 10 (7) (2015) e0133503.

[37] Dana Silverbush, Simona Cristea, Gali Yanovich-Arad, Tamar Geiger, Niko Beerenwinkel, Roded Sharan, Simultaneous integration of multi-omics data improves the identification of cancer driver modules, Cell Syst. 8 (5) (2019) 456–466.

[38] Roman Schulte-Sasse, Stefan Budach, Denes Hnisz, Annalisa Marsico, Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms, Nat. Mach. Intell. 3 (6) (2021) 513–526.

[39] Thomas N. Kipf, Max Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[40] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, et al., Oncokb: a precision oncology knowledge base, JCO Precis. Oncol. 1 (2017) 1–16.

[41] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[42] Thomas N. Kipf, Max Welling, Variational graph auto-encoders, 2016, arXiv preprint arXiv:1611.07308.

[43] Amin Salehi, Hasan Davulcu, Graph attention auto-encoders, 2019, arXiv preprint arXiv:1905.10715.

[44] Qingguo Wang, Joshua Armenia, Chao Zhang, Alexander V. Penson, Ed Reznik, Liguo Zhang, Thais Minet, Angelica Ochoa, Benjamin E. Gross, Christine A. Iacobuzio-Donahue, et al., Unifying cancer and normal RNA sequencing data from different sources, Sci. Data 5 (1) (2018) 1–8.

[45] Sohyun Hwang, Chan Yeong Kim, Sunmo Yang, Eiru Kim, Traver Hart, Edward M. Marcotte, Insuk Lee, HumanNet v2: human gene networks for disease research, Nucleic Acids Res. 47 (D1) (2019) D573–D580.

[46] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, Panos Kalnis, GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 705–713.

[47] Tianyi Zhao, Yang Hu, Jiajie Peng, Liang Cheng, DeepLGP: a novel deep learning method for prioritizing lncRNA target genes, Bioinformatics 36 (16) (2020) 4466–4472.

[48] Xinru Tang, Jiawei Luo, Cong Shen, Zihan Lai, Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction, Brief. Bioinform. 22 (6) (2021) bbab174.

[49] Ping Luo, Yulian Ding, Xiujuan Lei, Fang-Xiang Wu, Deepdriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks, Front. Genet. 10 (2019) 13.

[50] Abdelkader Behdenna, Maximilien Colange, Julien Haziza, Aryo Gema, Guillaume Appé, Chloé-Agathe Azencott, Akpéli Nordor, PyComBat, a python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods, BMC bioinformatics 24 (1) (2023) 459.

[51] Dimitra Repana, Joel Nulsen, Lisa Dressler, Michele Bortolomeazzi, Santhilata Kuppili Venkata, Aikaterini Tourna, Anna Yakovleva, Tommaso Palmieri, Francesca D. Ciccarelli, The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens, Genome Biol. 20 (2019) 1–12.

[52] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[53] Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, William F. Harrington, Sasha Pantel, John M. Krill-Burger, et al., Defining a cancer dependency map, Cell 170 (3) (2017) 564–576.

[54] Robin M. Meyers, Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, Phillip G. Montgomery, Glenn S. Cowley, Sasha Pantel, et al., Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells, Nat. Genet. 49 (12) (2017) 1779–1784.

[55] Daniel S. Peiffer, Debra Wyatt, Andrei Zlobin, Ali Piracha, Jeffrey Ng, Andrew K. Dingwall, Kathy S. Albain, Clodia Osipo, DAXX suppresses tumor-initiating cells in estrogen receptor–positive breast cancer following endocrine therapy, Cancer Res. 79 (19) (2019) 4965–4977.

[56] Yaqin Shi, Juan Jin, Xin Wang, Wenfei Ji, Xiaoxiang Guan, DAXX, as a tumor suppressor, impacts DNA damage repair and sensitizes BRCA-proficient TNBC cells to PARP inhibitors, Neoplasia 21 (6) (2019) 533–544.

[57] Daniel S. Peiffer, Emily Ma, Debra Wyatt, Kathy S. Albain, Clodia Osipo, DAXX-inducing phytoestrogens inhibit er+ tumor initiating cells and delay tumor development, NPJ Breast Cancer 6 (1) (2020) 37.

[58] Raj K. Gopal, Kirsten Kübler, Sarah E. Calvo, Paz Polak, Dimitri Livitz, Daniel Rosebrock, Peter M. Sadow, Braidie Campbell, Samuel E. Donovan, Salma Amin, et al., Widespread chromosomal losses and mitochondrial DNA alterations as genetic drivers in Hürthle cell carcinoma, Cancer Cell 34 (2) (2018) 242–255.

[59] Natalia Pstrąg, Katarzyna Ziemnicka, Hans Bluyssen, Joanna Wesoły, Thyroid cancers of follicular origin in a genomic light: in-depth overview of common and unique molecular marker candidates, Molecular Cancer 17 (2018) 1–17.

[60] Jan Brazina, Jan Svadlenka, Libor Macurek, Ladislav Andera, Zdenek Hodny, Jiri Bartek, Hana Hanzlikova, DNA damage-induced regulatory interplay between DAXX, p53, ATM kinase and Wip1 phosphatase, Cell Cycle 14 (3) (2015) 375–387.

[61] Lisa Y. Zhao, Jilin Liu, Gurjit S. Sidhu, Yuxin Niu, Yue Liu, Ruipeng Wang, Daiqing Liao, Negative regulation of p53 functions by daxx and the involvement of MDM2, J. Biol. Chem. 279 (48) (2004) 50566–50579.

[62] Iñigo Landa, Tihana Ibrahimpasic, Laura Boucai, Rileen Sinha, Jeffrey A. Knauf, Ronak H. Shah, Snjezana Dogan, Julio C. Ricarte-Filho, Gnana P. Krishnamoorthy, Bin Xu, et al., Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers, J. Clinical Investigat. 126 (3) (2016) 1052–1066.

[63] Xiaolu Yang, Roya Khosravi-Far, Howard Y. Chang, David Baltimore, Daxx, a novel fas-binding protein that activates JNK and apoptosis, Cell 89 (7) (1997) 1067–1076.

[64] Nicholas Mitsiades, Vassiliki Poulaki, George Mastorakos, Sophia Tseleni-Balafouta, Vassiliki Kotoula, Demetrios A. Koutras, Maria Tsokos, Fas ligand expression in thyroid carcinomas: a potential mechanism of immune evasion, J. Clinical Endocrinol. Metabol. 84 (8) (1999) 2924–2932.

[65] Wei Liu, Fang Sui, Jiazhe Liu, Meichen Wang, Sijia Tian, Meiju Ji, Bingyin Shi, Peng Hou, PAX3 is a novel tumor suppressor by regulating the activities of major signaling pathways and transcription factor FOXO3a in thyroid cancer, Oncotarget 7 (34) (2016) 54744.

[66] Ling-Chuan Guo, Wei-Dong Zhu, Xiang-Yuan Ma, Hao Ni, En-Jian Zhong, Yang W. Shao, Jie Yu, Dong-Mei Gu, Shun-Dong Ji, Hao-Dong Xu, et al., Mutations of genes including DNMT3a detected by next-generation sequencing in thyroid cancer, Cancer Biol. Therapy 20 (3) (2019) 240–246.

[67] Jennifer D. Kubic, Kacey P. Young, Rebecca S. Plummer, Anton E. Ludvik, Deborah Lang, Pigmentation PAX-ways: the role of Pax3 in melanogenesis, melanocyte stem cell maintenance, and disease, Pigment Cell Melanoma Res. 21 (6) (2008) 627–645.

[68] Iqbal Mahmud, Daiqing Liao, DAXX in cancer: phenomena, processes, mechanisms and regulation, Nucleic Acids Res. 47 (15) (2019) 7734–7752.

[69] Vincenza Capone, Laura Della Torre, Daniela Carannante, Mehrad Babaei, Lucia Altucci, Rosaria Benedetti, Vincenzo Carafa, HAT1: Landscape of biological function and role in cancer, Cells 12 (7) (2023) 1075.

[70] Hui Zhang, Junhong Han, Bin Kang, Rebecca Burgess, Zhiguo Zhang, Human histone acetyltransferase 1 protein preferentially acetylates H4 histone molecules in H3. 1-h4 over H3. 3-h4, J. Biol. Chem. 287 (9) (2012) 6573–6581.

[71] Eliezer M. Van Allen, Nikhil Wagle, Petar Stojanov, Danielle L. Perrin, Kristian Cibulskis, Sara Marlow, Judit Jane-Valbuena, Dennis C. Friedrich, Gregory Kryukov, Scott L. Carter, et al., Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine, Nat. Med. 20 (6) (2014) 682–688.

[72] Dong W. Kim, Lee Gazourian, Shafat A. Quadri, David H. Sherr, Gail E. Sonenshein, et al., The rela NF-κb subunit and the aryl hydrocarbon receptor (AhR) cooperate to transactivate the c-myc promoter in mammary cells, Oncogene 19 (48) (2000) 5498–5506.

[73] Julia Concetti, Caroline L. Wilson, NFKB1 and cancer: friend or foe? Cells 7 (9) (2018) 133.

[74] Pei-Yen Yeh, Yen-Shen Lu, Da-Liang Ou, Ann-Lii Cheng, Iκb kinases increase myc protein stability and enhance progression of breast cancer cells, Molecul. Cancer 10 (1) (2011) 1–12.

[75] Hirotaka Kanzaki, Avradip Chatterjee, Hanieh Hossein Nejad Ariani, Xinfeng Zhang, Stacey Chung, Nan Deng, V. Krishnan Ramanujan, Xiaojiang Cui, Mark I. Greene, Ramachandran Murali, Disabling the nuclear translocalization of rela/NF-κB by a small molecule inhibits triple-negative breast cancer growth, Breast Cancer: Targets Therapy (2021) 419–430.

[76] Jian Liao, Qing-hong Qin, Fa-you Lv, Zhen Huang, Bin Lian, Chang-yuan Wei, Qin-guo Mo, Qi-xing Tan, IKKα inhibition re-sensitizes acquired adriamycin-resistant triple negative breast cancer cells to chemotherapy-induced apoptosis, Sci. Rep. 13 (1) (2023) 6211.