www.oncotarget.com Oncotarget, 2022, Vol. 13, pp: 695-706

Research Paper

CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling

Limeng Pu[1,*], Manali Singha[2,*], Jagannathan Ramanujam[1,3] and Michal Brylinski[1,2]

[1]Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA

[2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

[3]Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803, USA

[*]These authors contributed equally to this work

Correspondence to: *Michal Brylinski*, ***email***: *michal@brylinski.org*

Keywords: cancer growth rate; kinase inhibitors; differential gene expression; gene-disease association; cancer-specific networks

Received: *March 22, 2022* ***Accepted***: *May 03, 2022* ***Published***: *May 19, 2022*

ABSTRACT

Development of novel anti-cancer treatments requires not only a comprehensive knowledge of cancer processes and drug mechanisms of action, but also the ability to accurately predict the response of various cancer cell lines to therapeutics. Numerous computational methods have been developed to address this issue, including algorithms employing supervised machine learning. Nonetheless, high prediction accuracies reported for many of these techniques may result from a significant overlap among training, validation, and testing sets, making existing predictors inapplicable to new data. To address these issues, we developed CancerOmicsNet, a graph neural network with sophisticated attention propagation mechanisms to predict the therapeutic effects of kinase inhibitors across various tumors. Emphasizing on the system-level complexity of cancer, CancerOmicsNet integrates multiple heterogeneous data, such as biological networks, genomics, inhibitor profiling, and gene-disease associations, into a unified graph structure. The performance of CancerOmicsNet, properly cross-validated at the tissue level, is 0.83 in terms of the area under the receiver operating characteristics, which is notably higher than those measured for other approaches. CancerOmicsNet generalizes well to unseen data, i.e., it can predict therapeutic effects across a variety of cancer cell lines and inhibitors. CancerOmicsNet is freely available to the academic community at https://github.com/pulimeng/CancerOmicsNet.

INTRODUCTION

Cancer is perhaps best understood as a complex system of interacting molecular-level networks, such as nuclear and cell networks, influenced by local and distant factors [1]. The nuclear network is composed of nucleic acid and protein molecules linked by a variety of biochemical and structural pathways allowing for the production of proteins based on the information encoded in the DNA [2]. Numerous curative cancer treatments have been developed either by targeting a single component within this network or by combining multiple agents to target different levels of the nuclear network in order to interrupt nucleic acid and protein machineries in the nucleus [3]. The cell network consists of various molecules interacting through the linkage of signal transduction pathways and the cytoskeleton [4, 5]. Particularly, the modulation of the activity of receptor tyrosine kinases, important components of the cell network, is an effective strategy against a wide variety of cancers [6]. This therapeutic effect can be achieved by either blocking upstream receptors with antibodies and small molecules or directly suppressing kinase catalytic activity with inhibitors [7]. Another group of therapies targeting the cell network disrupt metabolism by affecting the function of proteasome and chaperone molecules [8]. Since many cancer-specific data, such as molecular interactions, belong to the non-Euclidean space, a network-based

representation of cancer is generally well suited not only to predict the response of tumor cells to pharmacotherapy, but also to help understand drug-cell line interactions. However, utilizing these information-rich data requires advanced graph information processing algorithms and machine learning systems designed specifically to operate on the graph-structured data.

One of the earliest graph information processing techniques is a graph neural network (GNN) that employs a graph structure to learn the representation of the input data [9]. The major limitation of this method is that it restricts the information propagation to the first-order neighbors of every node limiting the information flow in the model. Recently, a graph convolutional network (GCN) was proposed to provide a more flexible model propagating information through many orders of neighbors [10, 11]. More advanced models were developed following the fundamental work on GCN, including a graph-based neural network employing the long-short term memory (LSTM) to carry out the information propagation that was demonstrated to have a significantly improved performance [12]. Another information propagation scheme aggregates the average embeddings of the neighboring nodes yielding a high performance especially for node classification in large graphs [13]. Numerous other techniques implementing minor improvements are currently available to operate on the graph-structured data [14–16].

Compared to other types of biological networks, gene co-expression networks have certain advantages, such as a high coverage of human genes, the additional knowledge obtained from the biomedical literature, and the possibility to study different cancer subtypes. [17, 18]. One of the most important applications of gene co-expression networks is to study the sensitivity of cancer cells to pharmacotherapy. Indeed, networks constructed by connecting those genes having correlated drug-induced expression values, contain a sufficient amount of information to predict drug sensitivity. In a recent study, two feature selection methods, network- and correlation-based, were developed to extract representative features for drug response prediction from gene co-expression networks [19]. The network-based feature selection utilizes assignment vectors describing the importance of individual vertices to predict drug sensitivity, whereas the correlation-based selection employs the Pearson correlation coefficient (PCC) between gene expression and the sensitivity of cell lines to drugs. Benchmarking calculations against non-small cell lung cancer with several canonical prediction algorithms, Elastic Net, Partial Least Squares Regression, Random Forest, Support Vector Regression, and Deep Neural Networks, demonstrated that features extracted with the network-based approach yield the highest performance when predicting the dose-response curve and the median effective dose.

Another group of methods utilize dual-layer cell line-drug networks, constructed by integrating drug similarity and cell line similarity networks in a weighted fashion, to predict the drug sensitivity of cancer cells. These techniques build on the observation that chemically similar drugs exhibit similar inhibitory effects on different cell lines and vice versa, similar cell lines tend to respond comparably to a treatment with the same drug. Dual-layer models typically require the optimization of various parameters, such as weights for individual drugs and cell lines, in order to determine the relative contribution of each network to the final prediction. As an example, a dual-layer network was developed to evaluate separately the response of a known cell line to a new drug and the effect of a known drug against a new cell line using a linear weighted model, followed by combining these two quantities into a sensitivity score for the treatment of a particular cell line with a drug [20]. Encouragingly, comprehensive benchmarks against the Cancer Cell Line Encyclopedia (CCLE) [21] and the Cancer Genome Project (CGP) [22] datasets showed that the predicted and observed therapeutic responses are correlated for most tested drugs with a PCC of 0.6, significantly outperforming an Elastic Net model. Additionally, this dual-layer integrated cell line-drug network model correctly predicted that certain mutant cell lines are more sensitive to inhibitors than the corresponding wild-type cell lines even though no mutation-specific information was provided.

More advanced methods combine genomics with drug chemical and activity information to predict the response to drugs in cancer treatment. For instance, the Cancer Drug Response Profile scan, or CDRscan, predicts anticancer drug responsiveness based on the drug screening assay data, the genomic profiles of human cancer cell lines, and the molecular fingerprints of drugs [23]. The analysis of observed and predicted drug responses showed an exceptionally high accuracy of CDRscan with a mean coefficient of determination of 0.84 and the area under the receiver operating characteristics (ROC) of 0.98. Another technique, DeepDR, predicts drug response purely based on the mutation and expression profiles of cancer cells. The reported overall prediction performance of DeepDR is also exceptionally high with a mean squared error of only 1.96 in the log-scale IC_{50} values. Further, a similarity-regularized matrix factorization method, or SRMF, predicts anticancer drug responses of cell lines solely from the chemical structures of drugs and the baseline gene expression levels in cell lines [24]. Those two features are used as regularization terms, which are incorporated into the drug response matrix factorization model. SRMF yields a drug-averaged mean squared error of 1.73 between predicted and observed responses of sensitive and resistant cell lines.

Notwithstanding these encouraging reports, there are two drawbacks of currently available techniques to

predict the response of cancer to drug treatment. First, most of these methods employ hand-crafted features simply exploiting similarities between instances, i.e., they essentially look for similar combinations of cell lines and drugs with known therapeutic outcomes. In reality, similar cell line-drug combinations may not necessarily produce the anticipated effects. Explicit similarity-based approaches are also unlikely to reveal the underlying mechanisms of the response of cancer to drug treatment. Second, the performance of many existing algorithms is likely grossly overestimated due to randomly splitting the redundant data into training, validation, and testing subsets resulting in a significant overlap among these sets. To address both issues, we developed CancerOmicsNet, a GNN-based algorithm employing multiple graph convolutional blocks with the attention-based propagation and a sophisticated graph readout mechanism to predict the effect of a drug treatment on the cancer cell growth. This novel method utilizes compact, cancer-specific networks constructed from protein-protein interactions, differential gene expression, disease-gene association, and drug inhibition data. The generalizability of CancerOmicsNet is carefully evaluated in a series of cross-validation benchmarks against different tumor tissues.

RESULTS

Cancer-specific data represented as networks

Input for CancerOmicsNet are cancer-specific networks assembled from multiple heterogeneous data including protein-protein interactions (PPIs), differential gene expression (DGE), disease-gene association (DGA) scores, kinase inhibitor profiling (KIP), and growth rate inhibition (GR). The procedure of data integration is schematically presented in Figure 1 for a combination of breast adenocarcinoma cell line MDA-MB-468 originated from a 51 years old female sample [25], and dasatinib, a dual kinase inhibitor against BCR/ABL and SRC families of tyrosine kinases [26] primarily used to treat chronic myelogenous leukemia and acute lymphoblastic leukemia [27]. In this example subnetwork, nodes (circles) are proteins and dashed lines represent highly confident PPIs. Bold purple circles are kinase nodes and thin blue circles are non-kinase proteins.

After the initial network is constructed (Figure 1A), proteins are annotated with DGE, DGA, and KIP scores (Figure 1B). EGFR is a transmembrane receptor tyrosine kinase having a critical impact on the regulation of apoptosis, cell migration, and cell proliferation. Since it is hyper-expressed in MDA-MB-468 cell line [28], node 1 in Figure 1B is colored green. On the other hand, node i is colored red because ubiquitin ligase CBL is deregulated in breast cancer [29]. In normal cells, CBL mediated ubiquitination negatively regulates EGFR by lysosomal degradation [30], however, CBL mutants escape the degradation of overexpressed EGFR inducing oncogenesis [29]. Next, DGA data for MDA-MB-468 cell line are mapped to proteins in the network; kinase nodes 1 and 4 are assigned DGA scores of 5.2 and 3.4, whereas non-kinase proteins e, f, and g have DGA scores of 2.6, 1.9, and 2.3, respectively. EGFR has the highest DGA score for breast adenocarcinoma likely because it is hyper-expressed in approximately half of the cases of inflammatory breast cancer and triple-negative breast cancer [31].

Subsequently, the inhibition data against dasatinib are added to the network. Dasatinib inhibits SRC with an IC_{50} value of 0.8 nM in a cell-free assay [32] and different variants of EGFR with IC_{50} ranging from 21.7 to 138 nM [33]. Two kinase nodes (1 and 3) are annotated with pIC_{50} values for dasatinib (6.8 and 8.8). Finally, the entire graph is assigned a label describing the effectiveness of the drug therapy against a given cell line. Since the growth of MDA-MB-468 cell line is inhibited by 30% 48 hrs after the treatment with dasatinib at 3 μm [34] and the experimental GR_{max} value [35] is –0.96, the label of the MDA-MB-468-dasatinib combination is a positive pharmacotherapeutic effect.

Network reduction driven by biological knowledge

Cancer-specific networks are subsequently subjected to a reduction procedure devised to produce graphs that are more compact yet richer in the biological information. This algorithm is presented in Figure 1C for the MDA-MB-468-dasatinib subnetwork. Briefly, a group of connected non-kinase proteins having similar DGE values and being part of the same biological processes according to Gene Ontology [36] are merged into a single node. Three such groups are present in the example subnetwork, a-b, c-e, and d-f-g (yellow shapes in Figure 1C). The first group comprises transcription factor P300, a product of EP300 gene, regulating the expression of NANOG that is responsible for pluripotency and self-renewal of stem cells [37]. The second group consists of a transcription activator STAT3 regulating the expression of IL10 [38]. The last cluster contains HSP90AA1 and HIF1A that together regulate the oxygen homeostasis [39] and PXN, a multidomain and multifunctional focal adhesion adaptor protein playing an essential role in the oxidative stress in cells [40]. The resulting virtual nodes in the reduced graph (dashed rounded squares in Figure 1D) representing multiple proteins involved in the same biological processes have a similar expression in cancer cells and are annotated with a median value of the DGE scores of incident nodes.

Information propagation in CancerOmicsNet

CancerOmicsNet implements a GNN model to predict the response of cancer cell lines to a treatment with

kinase inhibitors. The GNN employs graph convolutions, which are functionally equivalent to matrix convolutions in the convolutional neural network (CNN) working with images. Similar to the CNN propagating the information of a pixel to its neighbor pixels, the GNN propagates the information of a node in the graph to its neighbor nodes. The architecture of CancerOmicsNet is presented in Figure 2. An instance consisting of the combination of a cell line and a drug is used to create a cancer-specific network, which is subsequently subjected to the reduction procedure (Figure 2A). The reduced graph is then processed through a cascade of graph convolution blocks (Figure 2B). Each block contains three components, the attention-based propagation, the embedding update, and the generation of new embeddings. Although only the information from 1st order neighbors is passed between nodes in a single block, using multiple sequential blocks propagates the information from higher order neighbors.

This procedure is illustrated in Figure 3 for a simple 4-node graph. Initially, each node has its own information (color coded in Figure 3A), which is used to generate node embeddings. In our model, nodes are proteins connected through PPIs and the information comprises DGE, DGA, and KIP. During the first propagation step, a node of interest, such as node 1 in Figure 3, receives information from its 1st order neighbor, node 2 (Figure 3B). At the same time, node 2 receives information from its 1st order neighbors, nodes 3 and 4. Nodes 1 and 2 now contain more information to generate new embeddings. In the second propagation step, the information from nodes 3 and 4 already present in node 2 is also passed to node 1 (Figure 3C). At this point, new embeddings for node 1 are generated using not only its own information, but also the information propagated from its 1st and 2nd order neighbors. Three graph convolution blocks are employed in our model because we found empirically that adding the fourth block does not improve the performance anymore. Further, there is no point of using more than four blocks because the diameter of the cancer-specific graph is 5, so no new information is propagated beyond 4th order neighbors.

Graph information extraction

Once all embeddings are generated, the information on the entire graph can be extracted with a readout mechanism to predict the final drug response (Figure 2C). Standard readout techniques, such as global pooling, are unsuitable for our model comprising multiple graph convolutional blocks and learning from highly heterogeneous input graphs. In CancerOmicsNet, node embeddings generated by consecutive graph convolutional blocks contain distinct information. Therefore, a jumping knowledge network (JK-Net) is employed to exploit all information collected from different blocks. JK-Net was specifically developed to efficiently integrate the output from different layers into a single representation [41]. It is based on the concept of an influence radius corresponding to the radius of neighbors whose output is to be aggregated. The selection of an optimal radius is

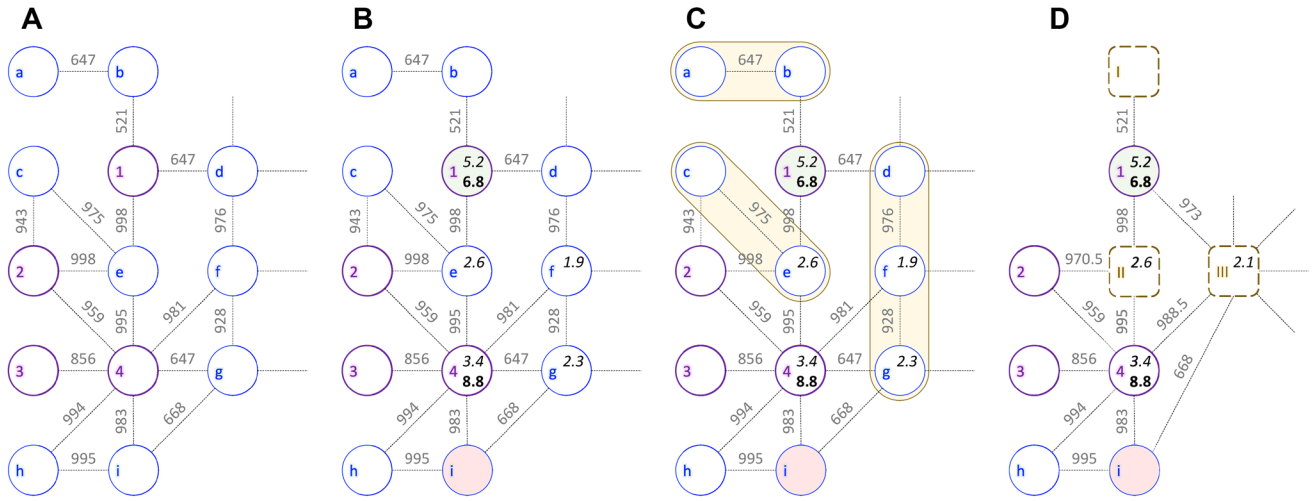

Figure 1: Example of a cancer-specific subnetwork. The graph shows a portion of protein-protein interaction network for breast adenocarcinoma cell line MDA-MB-468 and kinase inhibitor dasatinib. Bold purple circles represent kinase nodes (1 – EGFR, 2 – JAK2, 3 – JAK1, and 4 – SRC), whereas non-kinase nodes are shown as thin blue circles (a – NANOG, b – EP300, c – IL10RA, d – HIF1A, e – STAT3, f – HSP90AA1, g – PXN, h – CRK, and i – CBL). Edge weights are confidence scores for protein-protein interactions with a threshold value of ≥500. (**A**) Initial subnetwork constructed from interactions obtained from the STRING database. (**B**) Subnetwork integrating kinase inhibitor profiling (pIC_{50}, in bold), disease-gene association scores (in italics), and the differential gene expression: up- (green), down- (red), and normally (gray) regulated. (**C**) Graph reduction procedure with orange shapes outlining groups of non-kinase nodes that have similar differential gene expression and belong to the same GOGO cluster. (**D**) Reduced cancer-specific subnetwork with merged nodes shown as dashed brown rounded boxes (I – constructed from incident nodes a-b, II – c-e, and III – d-f-g). Node features and edge weights for merged nodes are calculated as median values of incident nodes.

crucial because large radii may cause excessive averaging and small radii may result in an insufficient information aggregation. JK-Net learns the effective neighborhood size for each layer in order to generate the best representation of the entire graph.

Global pooling of the embeddings of all nodes is appropriate only for homogeneous networks. In contrast, cancer-specific networks are highly heterogeneous comprising nodes of varying importance to one another and to the overall graph. Therefore, we added a mechanism to emphasize on important nodes rather than treating all nodes equally. Although such techniques have successfully been used in the CNN and the RNN [42], unlike images or text, graphs are orderless, i.e. an image does not remain the same if pixels are rearranged, while a graph remains the same if nodes are reordered. To account for the lack of order in graphs, we added a Set2Set layer converting a set to another set [43]. This model employs a set of LSTMs recursively combining the state of the previous processing step with the current embeddings to generate attention values. These attention values and embeddings form new states for the next processing step. By using Set2Set, we ensure that any permutation performed on the original vector does not affect the final read vector. The information summarized by JK-Net and Set2Set for the entire graph is then passed to a set of fully connected layers to make the final prediction, which is the effect of pharmacotherapy on the cancer cell growth.

Performance of CancerOmicsNet compared to other methods

In order to properly evaluate the generalizability of CancerOmicsNet, we performed a cross-validation at the tissue level. The entire dataset was first divided into nine groups of different tissues, digestive system, respiratory system, haematopoietic and lymphoid tissue, breast tissue, female reproductive system, skin, nervous system, excretory system, and others. Next, we conducted a 9-fold cross-validation, each time using cancer

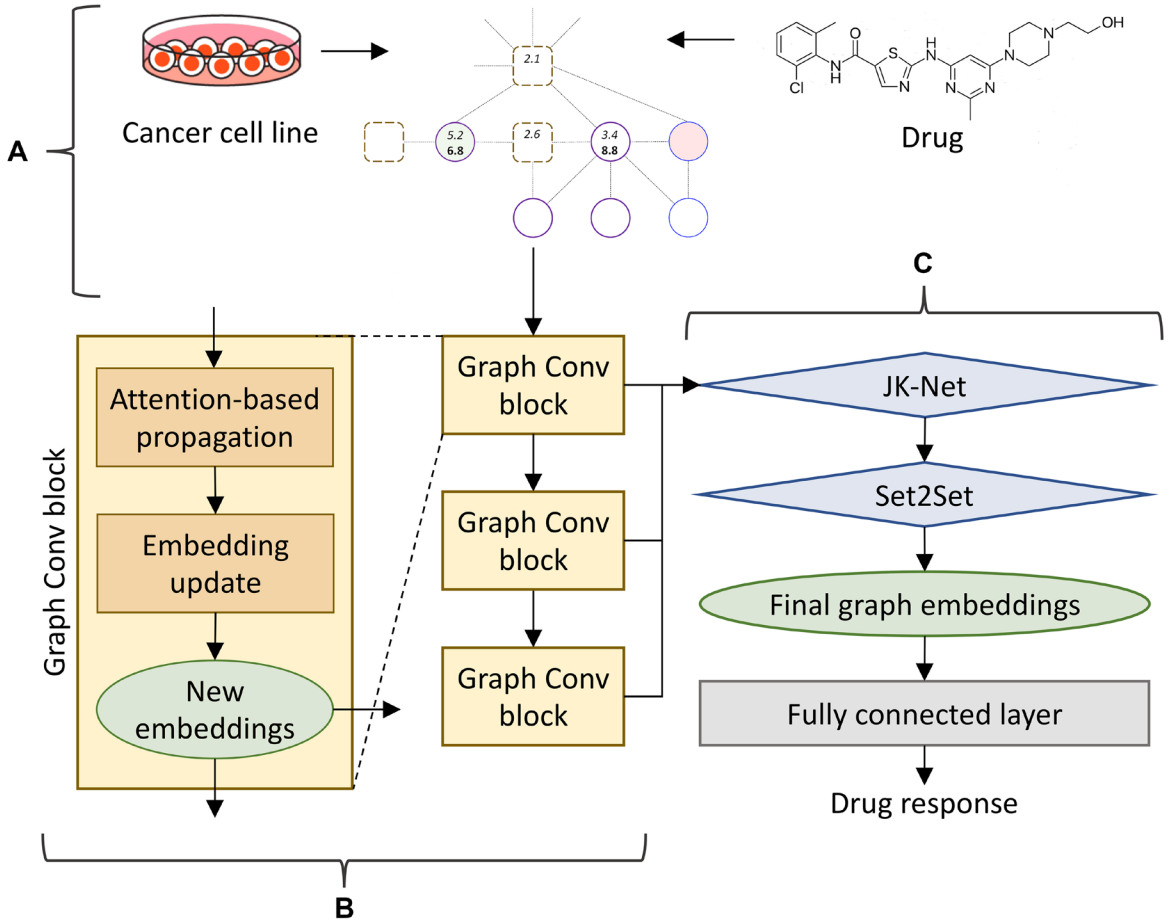

Figure 2: CancerOmicsNet architecture. (**A**) The input is a reduced graph constructed for the combination of a cell line and a small molecule inhibitor. (**B**) The graph is processed through a cascade of three graph convolutional blocks. Within each block, an attention-based propagation is first utilized to pass the information among nodes, and then a graph isomorphism network is employed to update the embeddings for each node. (**C**) Node embeddings generated by all blocks in B are combined using a JK-Net layer and passed to a Set2Set pooling layer serving as the read-out function to acquire the final graph embeddings. At the end, graph embeddings are sent to a fully connected layer to predict the drug response.

cell lines from one tissue as a validation set while the remaining cancer cell lines were used for model training. Since cell lines collected from different tissues have different gene expression patterns, this cross-validation scheme eliminates the overlap between training and validation data because the reduced graphs have different topologies. In addition, there is also a desired variability in feature matrices on account of different gene-disease associations, which depend on the cell line and tissue type. Essentially, each fold has entirely different training and validation data. Figure 4 shows a cross-validated ROC plot for CancerOmicsNet compared to other graph-based methods. Indeed, CancerOmicsNet not only gives the highest area under the curve (AUC) of 0.83 ± 0.02,

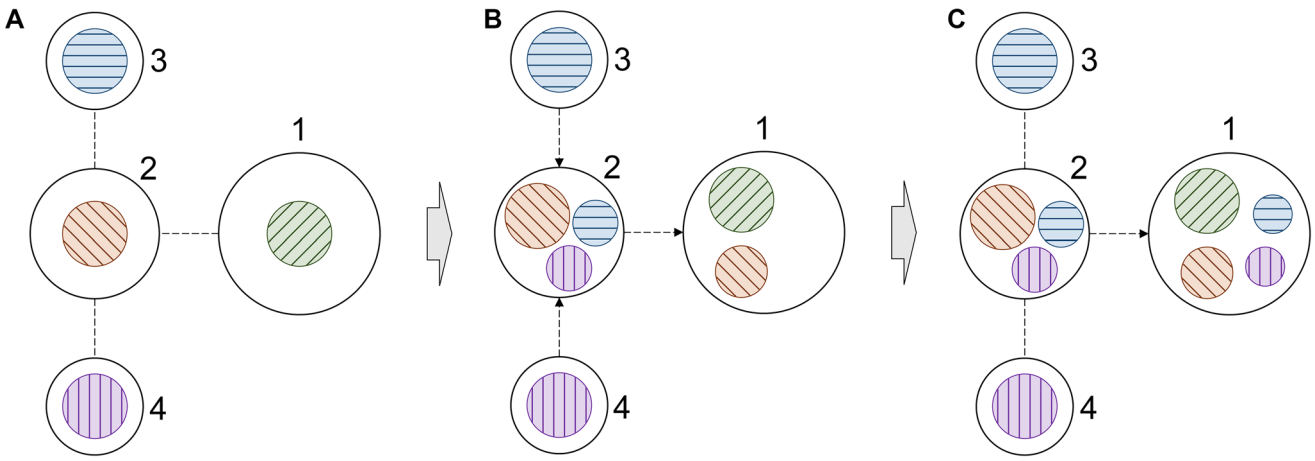

Figure 3: Schematic of information propagation in a graph. (**A**) A simple 4-node graph, in which each node contains its own information. The information is color coded, node 1 – green, node 2 – orange, node 3 – blue, and node 4 – purple. (**B**) The distribution of information within the graph after the first propagation step. (**C**) The distribution of information within the graph after the second propagation step. Only the information propagation to node 1 is illustrated in order to demonstrate how it receives information from higher order neighbors.

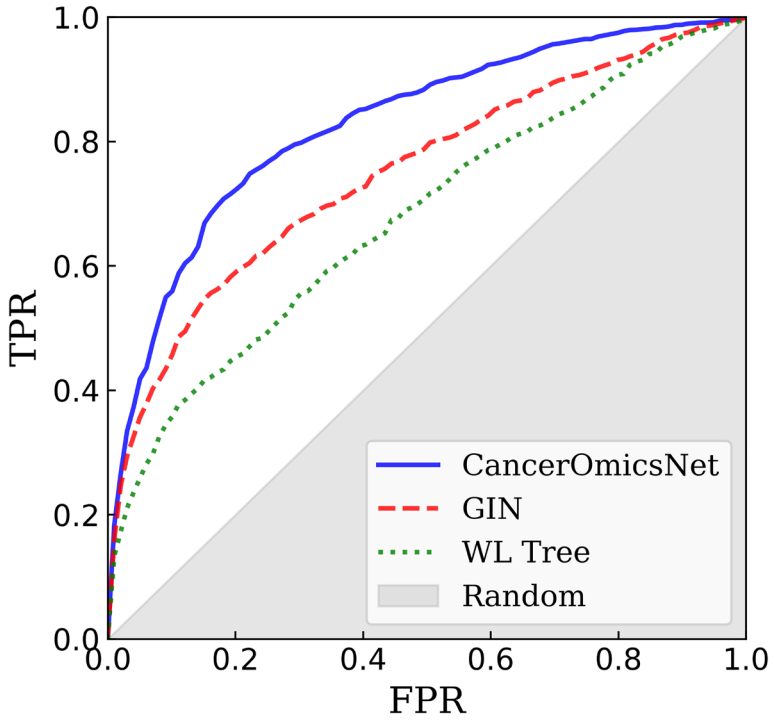

Figure 4: Performance of graph-based algorithms to predict the response of cancer cell lines to drugs. The performance of each method is cross-validated at the tissue level. CancerOmicsNet (solid blue line) is compared to the graph isomorphism network (GIN, dashed red line) utilizing equal propagation, and WL Tree (dotted green line) employing the Weisfeiler-Lehman graph kernel. TPR is the true positive rate, FPR is the false positive rate, and the gray area corresponds to the performance of a random predictor.

but the AUC values do not vary significantly for different tissues, digestive system (0.85), respiratory system (0.80), haematopoietic and lymphoid tissue (0.81), breast tissue (0.82), female reproductive system (0.86), skin (0.85), nervous system (0.83), excretory system (0.83), and others (0.81).

Removing the attention mechanism, which detects important nodes and puts more weight on them (labeled as GIN in Figure 4), decreases the AUC to 0.75 ±0.04 demonstrating that the propagation attention is an important component of CancerOmicsNet. Further, the performance of CancerOmicsNet is compared to that of the Weisfeiler-Lehman (WL) Tree [44]. Not only the AUC for WL Tree of 0.68 ±0.03 is lower than that for CancerOmicsNet, but since WL Tree processes one graph at a time, its runtimes are much longer than those for CancerOmicsNet featuring batch processing. Finally, Table 1 reports several performance metrics for two deep learning-based methods, CancerOmicsNet and CDRScan [23]. The precision quantifies the number of positive class predictions actually belonging to the positive class, whereas the recall quantifies the number of positive class predictions made out of all positive examples in the dataset. The balanced accuracy is computed as the average recall over all classes to addresses the imbalanced dataset problem [45]. The F-measure provides a single score balancing the concerns of both precision and recall in one number [46]. Encouragingly, using CancerOmicsNet yields up to 14% performance improvement over CDRScan. Overall, these results demonstrate that CancerOmicsNet outperforms other graph kernel and deep learning approaches.

DISCUSSION

In this study, we developed CancerOmicsNet, a graph neural network model to predict the growth rate of a cancer cell line after drug treatment. CancerOmicsNet is more advanced than many deep learning techniques operating in the Euclidean space [47], because it extracts knowledge directly from biological networks providing a more adequate representation of complex diseases such as cancer. Further, we implemented a sophisticated attention mechanism to propagate information more efficiently from the most important nodes in the graph when generating node embeddings. Attention mechanisms assigning trainable weights to nodes during information propagation are used to improve not only the classification performance [48, 49], but also the capability to generalize to larger, more complex, and noisy graphs [50, 51]. In our case, this technique allows the GNN model to direct more attention to kinase nodes since many of them contain valuable information on differential gene expression and the level of inhibition by small molecules across different cancer cell lines. As a result, the GNN achieves a better performance, especially against highly heterogeneous networks, such as cancer-specific networks employed in this study.

In order to evaluate the performance of CancerOmicsNet, we conducted a cross-validation at the tissue level by removing from model training all cell lines originating from a particular tissue and then analyzing the accuracy for these cell lines. We put a special attention to design a proper benchmarking protocol since in the context of predictive models, misunderstanding cross-validation very often yields an impressive, yet grossly overestimated predictor performance [52]. Numerous examples of exaggerated results in biomedical studies due to a problematic cross-validation include cancer prediction [53], the prediction of cancer cell line sensitivity and compound potency [54], the identification of drug-target interactions [55], the prediction of optimal drug therapies [56], the estimation of drug-target binding affinities [57], and virtual screening [58]. Since multiple instances in our dataset share cell lines originating from the same tissue, employing cross-validation at the tissue level is critical because splitting the dataset randomly into folds would cause training and validation instances to have a significant overlap with respect to graph topology as well as certain features such as gene expression and gene-disease associations. Encouragingly, the cross-validated accuracy of CancerOmicsNet at the tissue level is significantly higher than those measured for other approaches on the same data. Nonetheless, we note that the applicability of CancerOmicsNet is at present limited to kinase inhibitors, while alternative methods are applicable to other classes of therapeutics as well. Overall, CancerOmicsNet offers a high performance and the desired generalizability in the prediction of the effect of kinase-targeted therapies on the cancer cell growth.

MATERIALS AND METHODS

Cancer-specific molecular networks

Input graphs are constructed by mapping multiple heterogeneous data, DGE, KIP, DGA, and GR, on the human PPI network. STRING v11 database [59] has been used to construct the PPI network with an edge confidence threshold of ≥500. The resulting network comprises 19,144 proteins and 685,198 interactions. The DGE data were obtained from the curated Cancer Cell Line Encyclopedia (CCLE) containing the information on normally, up- and down-regulated genes for 749,551 associations between 18,022 genes and 1,035 cancer cell lines [21]. The KIP data on the half maximal inhibitory concentration (IC_{50}) for 49,348 small molecules and 411 kinases were collected from Team-SKI [60] and filtered at a minimum threshold of pIC_{50} (the negative logarithm of IC_{50}) of 6.3, which is equivalent to 500 nM in terms of IC_{50}. The DGA data were obtained from the DISEASE database [61] of 8,330 diseases and 20,715 genes, and the DisGeNET database

Table 1: Performance of CancerOmicsNet and CDRScan in predicting the response of cancer cell lines to drugs

Method	Balanced accuracy	Precision	Recall	F-measure
CancerOmicsNet	0.781	0.764	0.770	0.766
CDRScan	0.637	0.711	0.637	0.632

Accuracy, precision, recall, and F-score are calculated based on the cross-validation at the tissue level.

[62] of 24,166 diseases and 17,545 genes. The association scores range from 1 to 10 in DISEASE and from 0.01 to 1 in DisGeNET databases.

Growth rate inhibition data

Recent drug response metrics, GR_{50} and GR_{max}, quantify the proliferation with the value of growth rate inhibition (GR) based on time course and endpoint assays [35]. GR_{50} is the concentration of a drug at which GR is 0.5, whereas GR_{max} is the maximum measured GR value. Negative GR_{max} values correspond to the cytotoxic response and positive values correspond to the cytostatic response. In this study, we employ six LINCS-Dose-Response datasets, Broad-HMS LINCS Joint Project, LINCS MCF10A Common Project, HMS LINCS Seeding Density Project, MEP-HMS LINCS Joint Project, Genentech Cell Line Screening Initiative, and Cancer Therapeutics Response Portal [35]. The original dataset contains 632 cell lines from different cancer tissues and 795 small molecules tested against those cancer cell lines, totaling 83,162 combinations. After mapping the GR data to the constructed cancer-specific molecular networks and removing those cases having either GR_{50} values set to infinity or multiple GR_{50} values for a particular cell line-drug combination, the final dataset comprises 359 cell lines, 29 drugs, and 3,549 cell line-drug combinations. The number of positive instances (the cytotoxic effects of drugs on cell lines) is 2,124, whereas the number of negative instances (cytostatic responses) is 1,425.

Graph reduction

A procedure devised to reduce the size of drug-cell line networks employs the topological information and the biological knowledge. Two neighboring nodes are merged when the following conditions are met, both nodes are kinase proteins, share the same gene expression, and belong to the same GOGO [63] cluster representing proteins involved in similar biological processes. Applying the graph reduction procedure produces smaller graphs with the average number of nodes of 1,349 and the average number of edges of 12,613. Even though the graph sizes are significantly reduced by more than 90%, the percentage of kinase nodes carrying most of the meaningful information increases from 2% to 30%. Another advantage of reduced graphs over full-size networks is their topological diversity created by differences in the gene expression profiles of various cancer cell lines.

Information propagation

The most widely adopted propagation protocol transmit the information equally without considering the importance of a node to its neighbors and to the graph. This protocol can be expressed as

$$X^{(t)} = D^{-1/2} A D^{-1/2} X^{(t-1)} \qquad \text{Equation 1}$$

where t is the propagation step, D is the degree matrix of the adjacency matrix A, and $X^{(t)}$ represents embeddings at the propagation step t. Note that the original node features can be denoted as the 0-th propagation step, $X^{(0)}$. It is obvious that not all nodes have the same importance to their neighbors. For instance, many non-kinases in our dataset contain no useful information because these proteins are normally expressed, have no association with a disease, and are not targets for inhibitors. The information propagated from such proteins should be less important compared to the information transmitted from kinases and other proteins differentially expressed and having high disease associations. On that account, we added a propagation attention mechanism to increase the importance of these nodes. Specifically, we implemented a mechanism to learn a dynamic and adaptive summary of the local neighborhood, which operates only in the feature space [64]. The attention from node i to node j, $\gamma_{i,j}$, is defined as

$$\gamma_{i,j} = \frac{e^{\beta \cos(x_i, x_j)}}{\sum_{k \in N(i) \cup \{i\}} e^{\beta \cos(x_i, x_k)}} \qquad \text{Equation 2}$$

where $N(i)$ denotes the neighbors of node i and β is a trainable parameter. Essentially, the attention is the softmax of feature cosine similarities between center nodes and their neighbors. By utilizing the attention mechanism, the original propagation matrix calculated from the degree and adjacency matrices shown in Equation 1 can be replaced with a new propagation matrix Γ, which adaptively adjusts propagation weights based on neighbor features. This new propagation scheme addressing the problem of equal weights can be expressed as

$$X^{(t)} = \Gamma X^{(t-1)} \qquad \text{Equation 3}$$

where each entry of the propagation matrix Γ is calculated using Equation 2.

Node embeddings

After the information is propagated, the embeddings of each node need to be updated. Many techniques are available to generate node embeddings, and each has its advantages and disadvantages. Based on a series of preliminary experiments, we decided to implement a model inspired by the graph isomorphism network (GIN) [65]. The GIN offers an exceptional performance and has a relatively simple structure, which is important for our model because even after reduction, the cancer input data are much larger than typical datasets used in other fields. Briefly, the GIN transforms the graph isomorphism to the context of deep learning. Nonetheless, it employs a rather basic propagation scheme summing up features from all neighbor nodes. In order to further increase the performance, we replaced this simple propagation step with the attention-based propagation scheme shown in Equation 3. Combining the GIN update protocol with the propagation attention results in a very efficient graph convolution block expressed as

$$X' = \Theta\left(\left(\Gamma + (1+\varepsilon)\cdot I\right)\cdot X\right) \qquad \text{Equation 4}$$

where Θ denotes a neural network, Γ is the propagation matrix calculated using Equation 2 and ε is a trainable parameter.

Graph readout mechanism

CancerOmicsNet employs JK-Net followed by a Set2Set model to generate a global representation of the input graph from the node-wise information. JK-Net exploits varying influence radii of different layers to learn the best representation of the entire graph. This model can integrate outputs from individual graph convolutional blocks with three strategies, the concatenation, the max-pooling, and the LSTM attention. Considering the size of our data, we decided to employ the max-pooling strategy since it does not introduce any additional hyperparameters. This particular strategy performs a feature-wise max-pooling with lower layers favoring the local information and higher layers mostly containing the global graph information. With the max-pooling scheme, JK-Net automatically selects the most informative neighborhood size for each feature coordinate. Once the information from different layers is aggregated, we adopted the Set2Set model [43] as a final attention-based readout mechanism. A conventional method to simply flatten all embeddings is unsuitable for orderless graphs, which require a premutation-invariant readout mechanism instead. Set2Set comprises three blocks, a reading block, a process block, and a write block. In CancerOmicsNet, the reading block generating embeddings for each item in the set is replaced by JK-Net aggregating information from multiple graph convolutional blocks. The process block is an LSTM that reads the embeddings and state generated from the previous processing step, and outputs a new hidden state. Finally, the write block is also an LSTM, which takes the hidden state as a context to generate the attention for each item in the set. Subsequently, the attention vector is combined with the embedding matrix using a weighted summation to generate new, permutation-invariant embeddings.

Other methods to predict cancer drug response

CancerOmicsNet is compared to several other methods to predict the growth rate of cancer cell lines after drug treatment against the same dataset and employing the same cross-validation protocol. The graph isomorphism network (GIN) incorporates the graph isomorphism test to generate node embeddings preserving the original graph structure at each propagation step [65]. As a result, the propagation process contains not only the propagated information, but also the node information in the original graph as an extra term. The Weisfeiler-Lehman (WL) Tree is a widely adopted graph kernel method for graph machine learning [44]. This algorithm utilizes kernel functions and the WL graph isomorphism test to iteratively generate new labels for nodes and new representations for graphs. By iteratively propagating the information, the final information for each node and the entire graph can be extracted.

Cancer Drug Response Profile scan (CDRscan) is a deep learning model predicting drug response from cancer genomic signature [23]. CDRscan employs two input data, the genetic mutation information and the molecular profiles of drugs represented by PaDEL-descriptors [66]. In order to apply CDRscan to our dataset, the mutation information was substituted with the gene expression of cancer cell lines. Following the original implementation, the input data are passed through CNNs to extract features, which are then concatenated to make the final prediction. In the original paper, five slightly different models were employed in order to create an ensemble model. However, since there neither fundamental differences among these models nor a significant performance improvement of the ensemble model, we implemented the best performing single model according to the original benchmarks.

Data availability

CancerOmicsNet is open sourced and freely available to the academic community at https://github.com/pulimeng/CancerOmicsNet.

Author contributions

Conceptualization: LP, MS, JR, MB. Data curation: MS, LP, MB. Methodology: LP, MS, MB. Funding

acquisition: JR, MB. Software: LP. Supervision: MB. Manuscript draft: LP, MS. Final manuscript: MB.

ACKNOWLEDGMENTS

Portions of this research were conducted with computing resources provided by Louisiana State University. The authors thank Dr. Hsiao-Chun Wu for his valuable comments on the model development.

CONFLICTS OF INTEREST

Authors have no conflicts of interest to declare.

FUNDING

This work has been supported in part by the National Institute of General Medical Sciences of the National Institutes of Health award R35GM119524, the US National Science Foundation award CCF-1619303, the Louisiana Board of Regents contract LEQSF(2016-19)-RD-B-03, and the Center for Computation and Technology at Louisiana State University.

REFERENCES

1. Camacho DF, Pienta KJ. Disrupting the networks of cancer. Clin Cancer Res. 2012; 18:2801–8. https://doi.org/10.1158/1078-0432.CCR-12-0366. [PubMed]
2. Cozzolino F, Iacobucci I, Monaco V, Monti M. Protein-DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era. J Proteome Res. 2021; 20:3018–30. https://doi.org/10.1021/acs.jproteome.1c00074. [PubMed]
3. Kwak EL, Clark JW, Chabner B. Targeted agents: the rules of combination. Clin Cancer Res. 2007; 13:5232–37. https://doi.org/10.1158/1078-0432.CCR-07-1385. [PubMed]
4. Janmey PA. The cytoskeleton and cell signaling: component localization and mechanical coupling. Physiol Rev. 1998; 78:763–81. https://doi.org/10.1152/physrev.1998.78.3.763. [PubMed]
5. Pawson T, Nash P. Protein-protein interactions define specificity in signal transduction. Genes Dev. 2000; 14:1027–47. [PubMed]
6. Takeuchi K, Ito F. Receptor tyrosine kinases and targeted cancer therapeutics. Biol Pharm Bull. 2011; 34:1774–80. https://doi.org/10.1248/bpb.34.1774. [PubMed]
7. Montor WR, Salas AROSE, Melo FHM. Receptor tyrosine kinases and downstream pathways as druggable targets for cancer treatment: the current arsenal of inhibitors. Mol Cancer. 2018; 17:55. https://doi.org/10.1186/s12943-018-0792-2. [PubMed]
8. Cvek B, Dvorak Z. The ubiquitin-proteasome system (UPS) and the mechanism of action of bortezomib. Curr Pharm Des. 2011; 17:1483–99. https://doi.org/10.2174/138161211796197124. [PubMed]
9. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw. 2009; 20:61–80. https://doi.org/10.1109/TNN.2008.2005605. [PubMed]
10. Kipf TN, Welling M. Semi-supervised classification with Graph Convolutional Networks. arXiv. 2016. https://doi.org/10.48550/arXiv.1609.02907.
11. Kipf TN, Welling M. Variational graph auto-encoders. arXiv. 2016. https://doi.org/10.48550/arXiv.1611.07308.
12. Li Y, Tarlow D, Brockschmidt M, Zemel R. Gated graph sequence neural networks. arXiv. 2015. https://doi.org/10.48550/arXiv.1511.05493.
13. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. arXiv. 2018. https://doi.org/10.48550/arXiv.1706.02216.
14. Bacciu D, Errica F, Micheli A. Contextual Graph Markov Model: A deep and generative approach to graph processing. arXiv. 2018. https://doi.org/10.48550/arXiv.1805.10636.
15. Chen J, Ma T, Xiao C. FastGCN: Fast learning with graph convolutional networks via importance sampling. arXiv. 2018. https://doi.org/10.48550/arXiv.1801.10247.
16. Liang X, Shen X, Feng J, Lin L, Yan S. Semantic object parsing with graph LSTM. arXiv. 2016. https://doi.org/10.48550/arXiv.1603.07063.
17. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. J Biopharm Stat. 2010; 20:281–300. https://doi.org/10.1080/10543400903572753. [PubMed]
18. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun. 2014; 5:3231. https://doi.org/10.1038/ncomms4231. [PubMed]
19. Ahmed KT, Park S, Jiang Q, Yeu Y, Hwang T, Zhang W. Network-based drug sensitivity prediction. BMC Med Genomics. 2020; 13:193. https://doi.org/10.1186/s12920-020-00829-3. [PubMed]
20. Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. PLoS Comput Biol. 2015; 11:e1004498. https://doi.org/10.1371/journal.pcbi.1004498. [PubMed]
21. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–7. https://doi.org/10.1038/nature11003. [PubMed]
22. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483:570–75. https://doi.org/10.1038/nature11005. [PubMed]

23. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, Jung J, Shin JM. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci Rep. 2018; 8:8857. https://doi.org/10.1038/s41598-018-27214-6. [PubMed]
24. Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer. 2017; 17:513. https://doi.org/10.1186/s12885-017-3500-5. [PubMed]
25. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. J Biomol Tech. 2018; 29:25–38. https://doi.org/10.7171/jbt.18-2902-002. [PubMed]
26. Kamath AV, Wang J, Lee FY, Marathe PH. Preclinical pharmacokinetics and in vitro metabolism of dasatinib (BMS-354825): a potent oral multi-targeted kinase inhibitor against SRC and BCR-ABL. Cancer Chemother Pharmacol. 2008; 61:365–76. https://doi.org/10.1007/s00280-007-0478-8. [PubMed]
27. Keating GM. Dasatinib: A Review in Chronic Myeloid Leukaemia and Ph+ Acute Lymphoblastic Leukaemia. Drugs. 2017; 77:85–96. https://doi.org/10.1007/s40265-016-0677-x. [PubMed]
28. Jackson NM, Ceresa BP. Protein Kinase G facilitates EGFR-mediated cell death in MDA-MB-468 cells. Exp Cell Res. 2016; 346:224–32. https://doi.org/10.1016/j.yexcr.2016.07.001. [PubMed]
29. Ahmed SF, Buetow L, Gabrielsen M, Lilla S, Sibbet GJ, Sumpton D, Zanivan S, Hedley A, Clark W, Huang DT. E3 ligase-inactivation rewires CBL interactome to elicit oncogenesis by hijacking RTK-CBL-CIN85 axis. Oncogene. 2021; 40:2149–64. https://doi.org/10.1038/s41388-021-01684-x. [PubMed]
30. Hong SY, Kao YR, Lee TC, Wu CW. Upregulation of E3 Ubiquitin Ligase CBLC Enhances EGFR Dysregulation and Signaling in Lung Adenocarcinoma. Cancer Res. 2018; 78:4984–96. https://doi.org/10.1158/0008-5472.CAN-17-3858. [PubMed]
31. Masuda H, Zhang D, Bartholomeusz C, Doihara H, Hortobagyi GN, Ueno NT. Role of epidermal growth factor receptor in breast cancer. Breast Cancer Res Treat. 2012; 136:331–45. https://doi.org/10.1007/s10549-012-2289-9. [PubMed]
32. O'Hare T, Walters DK, Stoffregen EP, Jia T, Manley PW, Mestan J, Cowan-Jacob SW, Lee FY, Heinrich MC, Deininger MW, Druker BJ. In vitro activity of Bcr-Abl inhibitors AMN107 and BMS-354825 against clinically relevant imatinib-resistant Abl kinase domain mutants. Cancer Res. 2005; 65:4500–5. https://doi.org/10.1158/0008-5472.CAN-05-0259. [PubMed]
33. Formisano L, D'Amato V, Servetto A, Brillante S, Raimondo L, Di Mauro C, Marciano R, Orsini RC, Cosconati S, Randazzo A, Parsons SJ, Montuori N, Veneziani BM, et al. Src inhibitors act through different mechanisms in Non-Small Cell Lung Cancer models depending on EGFR and RAS mutational status. Oncotarget. 2015; 6:26090–103. https://doi.org/10.18632/oncotarget.4636. [PubMed]
34. Nautiyal J, Majumder P, Patel BB, Lee FY, Majumdar AP. Src inhibitor dasatinib inhibits growth of breast cancer cells by modulating EGFR signaling. Cancer Lett. 2009; 283:143–51. https://doi.org/10.1016/j.canlet.2009.03.035. [PubMed]
35. Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. Nat Methods. 2016; 13:521–27. https://doi.org/10.1038/nmeth.3853. [PubMed]
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. https://doi.org/10.1038/75556. [PubMed]
37. Wang T, Liu H, Ning Y, Xu Q. The histone acetyltransferase p300 regulates the expression of pluripotency factors and odontogenic differentiation of human dental pulp cells. PLoS One. 2014; 9:e102117. https://doi.org/10.1371/journal.pone.0102117. [PubMed]
38. Hedrich CM, Rauen T, Apostolidis SA, Grammatikos AP, Rodriguez Rodriguez N, Ioannidis C, Kyttaris VC, Crispin JC, Tsokos GC. Stat3 promotes IL-10 expression in lupus T cells through trans-activation and chromatin remodeling. Proc Natl Acad Sci U S A. 2014; 111:13457–62. https://doi.org/10.1073/pnas.1408023111. [PubMed]
39. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009; 37:D674–79. https://doi.org/10.1093/nar/gkn653. [PubMed]
40. López-Colomé AM, Lee-Rivera I, Benavides-Hidalgo R, López E. Paxillin: a crossroad in pathological cell migration. J Hematol Oncol. 2017; 10:50. https://doi.org/10.1186/s13045-017-0418-y. [PubMed]
41. Xu K, Li C, Tian Y, Sonobe T, Kawarabayashi K, Jegelka S. Representation learning on graphs with jumping knowledge networks. arXiv. 2018. https://doi.org/10.48550/arXiv.1806.03536.
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. arXiv. 2017. https://doi.org/10.48550/arXiv.1706.03762.
43. Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets. arXiv. 2015. https://doi.org/10.48550/arXiv.1511.06391.
44. Shervashidze N, Schweitzer P, van Leeuwen EJ, Mehlhorn K, Borgwardt KM. Weisfeiler-Lehman graph kernels. J Mach Learn Res. 2011; 12:2539–61.
45. Kelleher JD, Mac Namee B, D'arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press; 2020.

46. Zheng Y, Wu Z. A Machine Learning-Based Biological Drug-Target Interaction Prediction Method for a Tripartite Heterogeneous Network. ACS Omega. 2021; 6:3037–45. https://doi.org/10.1021/acsomega.0c05377. [PubMed]

47. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: Going beyond Euclidean data. IEEE Signal Process Mag. 2017; 34:18–42. https://doi.org/10.1109/MSP.2017.2693418.

48. Zhang S, Xie L. Improving Attention Mechanism in Graph Neural Networks via Cardinality Preservation. IJCAI (U S). 2020; 2020:1395–402. https://doi.org/10.24963/ijcai.2020/194. [PubMed]

49. Zhang Y, Wang X, Jiang X, Shi C, Ye Y. Hyperbolic graph attention network. arXiv. 2019. https://doi.org/10.48550/arXiv.1912.03046.

50. Knyazev B, Taylor GW, Amer MR. Understanding attention and generalization in graph neural networks. arXiv. 2019. https://doi.org/10.48550/arXiv.1905.02850.

51. Shi M, Tang Y, Zhu X, Zhuang Y, Lin M, Liu J. Feature-Attention Graph Convolutional Networks for Noise Resilient Learning. IEEE Trans Cybern. 2022. [Epub ahead of print]. https://doi.org/10.1109/TCYB.2022.3143798. [PubMed]

52. Neunhoeffer M, Sternberg S. How cross-validation can go wrong and what to do about it. Political Analysis. 2019; 27:101–6. https://doi.org/10.1017/pan.2018.39.

53. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed. 2018; 153:1–9. https://doi.org/10.1016/j.cmpb.2017.09.005. [PubMed]

54. Cortés-Ciriano I, Bender A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. J Cheminform. 2019; 11:41. https://doi.org/10.1186/s13321-019-0364-5. [PubMed]

55. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug-target interactions based on graph convolutional network and deep neural network. Brief Bioinform. 2021; 22:2141–50. https://doi.org/10.1093/bib/bbaa044. [PubMed]

56. Huang C, Mezencev R, McDonald JF, Vannberg F. Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. PLoS One. 2017; 12:e0186906. https://doi.org/10.1371/journal.pone.0186906. [PubMed]

57. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018; 34:i821–29. https://doi.org/10.1093/bioinformatics/bty593. [PubMed]

58. Liu Z, Du J, Fang J, Yin Y, Xu G, Xie L. DeepScreening: a deep learning-based screening web server for accelerating drug discovery. Database (Oxford). 2019; 2019:baz104. https://doi.org/10.1093/database/baz104. [PubMed]

59. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47:D607–13. https://doi.org/10.1093/nar/gky1131. [PubMed]

60. Sorgenfrei FA, Fulle S, Merget B. Kinome-Wide Profiling Prediction of Small Molecules. ChemMedChem. 2018; 13:495–99. https://doi.org/10.1002/cmdc.201700180. [PubMed]

61. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015; 74:83–89. https://doi.org/10.1016/j.ymeth.2014.11.020. [PubMed]

62. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017; 45:D833–39. https://doi.org/10.1093/nar/gkw943. [PubMed]

63. Zhao C, Wang Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. Sci Rep. 2018; 8:15107. https://doi.org/10.1038/s41598-018-33219-y. [PubMed]

64. Thekumparampil KK, Wang C, Oh S, Li LJ. Attention-based Graph Neural Network for Semi-supervised Learning. arXiv. 2018. https://doi.org/10.48550/arXiv.1803.03735.

65. Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? arXiv. 2018. https://doi.org/10.48550/arXiv.1810.00826.

66. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011; 32:1466–74. https://doi.org/10.1002/jcc.21707. [PubMed]