ARTICLE OPEN

Check for updates

An integrated network representation of multiple cancer-specific data for graph-based machine learning

Limeng Pu[1,6], Manali Singha[2,6], Hsiao-Chun Wu[3], Costas Busch[4,5], J. Ramanujam[1,3] and Michal Brylinski[1,2]✉

Genomic profiles of cancer cells provide valuable information on genetic alterations in cancer. Several recent studies employed these data to predict the response of cancer cell lines to drug treatment. Nonetheless, due to the multifactorial phenotypes and intricate mechanisms of cancer, the accurate prediction of the effect of pharmacotherapy on a specific cell line based on the genetic information alone is problematic. Emphasizing on the system-level complexity of cancer, we devised a procedure to integrate multiple heterogeneous data, including biological networks, genomics, inhibitor profiling, and gene-disease associations, into a unified graph structure. In order to construct compact, yet information-rich cancer-specific networks, we developed a novel graph reduction algorithm. Driven by not only the topological information, but also the biological knowledge, the graph reduction increases the feature-only entropy while preserving the valuable graph-feature information. Subsequent comparative benchmarking simulations employing a tissue level cross-validation protocol demonstrate that the accuracy of a graph-based predictor of the drug efficacy is 0.68, which is notably higher than those measured for more traditional, matrix-based techniques on the same data. Overall, the non-Euclidean representation of the cancer-specific data improves the performance of machine learning to predict the response of cancer to pharmacotherapy. The generated data are freely available to the academic community at https://osf.io/dzx7b/.

npj Systems Biology and Applications (2022)8:14 ; https://doi.org/10.1038/s41540-022-00226-9

INTRODUCTION

Carcinogenesis is a systems-level phenomenon with a complex phenotype concomitant with the malfunction of signal transduction in a cell[1–3]. Biological networks, including protein-protein interaction (PPI) networks, are often employed to study the alteration of information flow in cancer cells caused by oncogenic changes in protein activity and expression[4–6]. Many early network-based methods build on the observation that gene products associated with similar diseases exhibit similar topological characteristics of PPI networks. For instance, Vavien utilizes the information flow in networks to extract functional disease-gene relationships in order to identify target genes for pharmacotherapy[7]. Systematic benchmarks of Vavien against the Online Mendelian Inheritance in Man database demonstrated that disease-gene candidates can effectively be prioritized based on their topological similarity to known disease-associated genes. A more recent study comprehensively analyzed 150 anticancer drugs approved by the US Food and Drug Administration divided into two groups according to their mechanism of action, cytotoxic and target-based agents[8]. It was found that although the proportion of target-based drugs increased in recent years, cytotoxic agents are used to treat more cancer types. Further, the investigation of the cancer-drug-target network comprising multiple cancer types, drugs, and targets revealed novel drug-cancer associations, most of which are supported by at least one clinical trial study.

These studies exemplify the utility of biological networks to address challenging tasks in cancer research, such as the efficacy prediction for anticancer drugs. Not surprisingly, there is a significant interest in the development of machine-learning systems applicable to graph-structured data[9] in order to further improve the information extraction and induction from biological networks. Graph-based machine-learning approaches can broadly be categorized into two major classes, graph kernels and spectral methods. An exemplar of the former technique is the Weisfeiler–Lehman (WL) algorithm[10], which iteratively assigns a label to each node based on the multi-set hashing of the neighbor labels. Subsequently, graph-level features are computed from either a histogram or another form of summarizing statistics for individual nodes. On the other hand, spectral methods utilize the graph Laplacian, a matrix calculated from the adjacency matrix of a graph[11]. A variety of matrix-related operations, such as eigenvalue decomposition, can be applied to the Laplacian to cluster nodes into different groups and perform graph classification. Alternative techniques to classify nodes and compute various graph statistics include PageRank, a random walk-based algorithm proposed by the founders of Google[12]. This technique determines the importance of a page, represented as a node in the graph, by counting the number and quality of links under the assumption that more important websites receive more links from other websites.

Recently developed graph-based machine-learning approaches for applications in biology and biomedicine include a new prediction method combining multiple kernels into a tripartite, heterogeneous drug-target-disease interaction space in order to integrate multiple sources of biological information[13]. This novel network-based algorithm adds a disease layer to the traditional drug-target interaction bipartite graph. From the constructed heterogeneous network, new drug-target interactions can be inferred with Gaussian kernel functions and the regularized least

[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. [4] Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA 70803, USA. [5] School of Computer and Cyber Sciences, Augusta University, Augusta, GA 30912, USA. [6] These authors contributed equally: Limeng Pu, Manali Singha. ✉email: michal@brylinski.org

square method of the Kronecker product. Encouragingly, comprehensive benchmarking simulations demonstrated that the network topology can be used to accurately detect drug-target interactions. Another example is a machine-learning framework to identify robust drug biomarkers with network-based analyses of the pharmacogenomic data derived from three-dimensional organoid culture models[14]. These biomarkers were shown to reliably predict responses of colorectal cancer patients to 5-fluorouracil and bladder cancer patients to cisplatin. Finally, biomarkers were confirmed against the external transcriptomic datasets of drug-sensitive and drug-resistant isogenic cancer cell lines, demonstrating that combining the application of gene modules and network-based techniques can be used to predict anticancer-drug efficacy in patients.

A major challenge for the application of machine-learning systems to biological networks is to ensure that these data contain a sufficiently high signal-to-noise ratio for the learning framework to efficiently perform information extraction and high-level induction. In addition, the size of many biological networks, such as PPI networks comprising thousands of nodes, is prohibitively large for many algorithms and, therefore, should preferably be reduced to facilitate a fast-learning process. To address these challenges, we describe a novel procedure to construct compact yet information-rich, cancer-specific graphs by integrating multiple heterogeneous data on differential gene expression, drug profiling, protein-protein interactions, and disease association scores. The resulting networks are characterized by various graph-based statistics and their information content is evaluated with topological and feature entropy measures. Finally, we conduct a comparative analysis of the performance of matrix- and graph-based predictors of the drug efficacy employing a tissue level cross-validation protocol. Overall, the integrated network representation of heterogeneous biological data improves the performance of machine learning to predict the effect of a drug treatment on the cancer cell growth over more traditional matrix-based approaches.

RESULTS

Construction of cancer-specific networks

Integration of the heterogeneous, cancer-specific data was performed by mapping the differential gene expression, disease-gene association scores, and kinase inhibitor profiling onto the human PPI network. The PPI network contains the majority of proteins and connections of the human phosphorylation network, yet it is larger, more diverse and strongly connected[15]. The integration procedure is schematically shown in Fig. 1. An initial network has two types of nodes representing kinases (circles) and non-kinase proteins (rounded squares). For a given cell line, up- (green) and down-regulated (red) genes are marked according to the differential gene expression for that cell line and some nodes are also assigned disease-gene association scores (numbers in bold). If this cell line is treated with a kinase inhibitor, pIC_{50} values against its targets are then added to the graph (numbers in italics). In the resulting graph, kinase nodes have gene expression values, and some kinase nodes also have pIC_{50} values and disease-gene association scores. Non-kinase nodes have gene expression values, and some non-kinases nodes also have disease-gene association scores. Note that all cell line-drug combinations have the same underlying PPI network, however, different cell lines usually have different gene expression values and disease association scores that depend on the tumor type. Similarly, various drugs usually inhibit different sets of kinases, therefore, node features are generally unique for cell line-drug combinations.

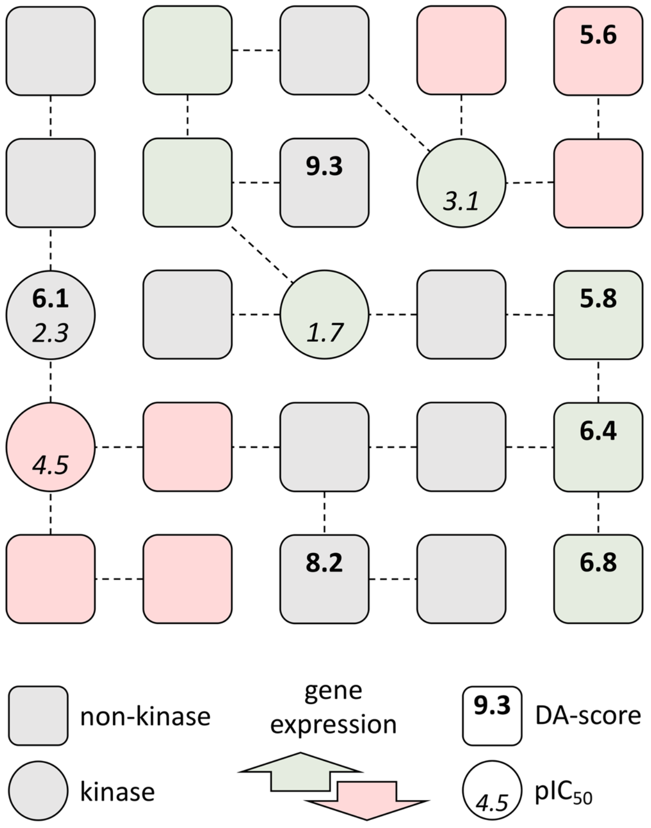

Fig. 1 Schematic of the graph representation of multiple heterogeneous data. Circles are kinases, whereas rounded squares represent non-kinase proteins. Nodes are connected through confident interactions forming a network. Each node is colored according to the differential gene expression: green—up-regulated, red —down-regulated, and gray—normally regulated. Both types of nodes can have gene-disease association scores (numbers in bold), whereas kinases can also have pIC_{50} values according to the kinase profiling data (numbers in italics).

Analysis of full-size networks

In the context of machine learning, full-size graphs corresponding to the original PPI networks are not necessarily the best representation of the cancer-specific data. First, all instances share the same graph topology and differences are only in node features, i.e., gene expression, disease association, and pIC_{50} values, making it difficult for a machine-learning model to gather the information required for effective learning. Second, full-size graphs are very sparse with a density (see Eq. 1) as low as 0.004 wasting the computer memory. A quick fix to this problem could be to employ a sparse representation, such as the coordination (COO) format[16], however, this approach is unworkable because available machine-learning libraries do not support sparse data formats. Third, the majority (98%) of nodes in a graph are non-kinase proteins with no inhibition data and most proteins are normally regulated according to the differential gene expression leading to a significant sparsity of important features. As a result, most entries in the feature matrix carry no effective information, which may result in a poor learning performance. An illustrative analogy is to try training a model to classify images having less than 1% different pixels. Given such tiny differences, any model is going to struggle learning the underlying patterns.

Knowledge-based reduction of cancer-specific networks

In order to address these issues, we devised a knowledge-based graph reduction procedure by edge contraction, which is a fundamental operation in graph theory[17]. Here, the idea is to remove a particular edge and then merge the incident nodes of that edge to form a new node. Edge contraction is widely used in recursive formulas to calculate the number of spanning trees of an arbitrary connected graph[18], and in recurrence formulas to calculate the chromatic polynomial of a simple graph[19]. Nonetheless, a simple edge contraction based solely on the connectivity is not going to produce the desired outcome in our case because we also need to account for the features of nodes. Therefore, we developed a knowledge-based edge contraction algorithm employing both connectivity and biological feature information to satisfy the following conditions: both incident nodes need to be non-kinase proteins, share the same differential gene expression, and belong to the same biological process cluster. The last condition is very important to ensure that the reduction merges only those nodes belonging to the same pathway, thus supporting the biological knowledge. Biological processes in cancer-specific networks are determined by clustering nodes according to the similarity of their Gene Ontology (GO) terms.

GOGO is a method to calculate semantic similarities between GO terms using Directed Acyclic Graphs (DAGs)[20]. GO consists of three DAGs created based on molecular function (MF), cellular component (CC), and biological process (BP) ontologies[21]. In order to verify that the network locality is preserved when using GOGO similarities derived from the BP ontology, we first calculated similarity values between 1st, 2nd, 3rd, and 4th order neighbors in the full-size PPI network. Figure 2 shows that GOGO similarities are the highest for the 1st order neighbors and decrease with the increasing order. These results corroborate previous studies demonstrating that the closer the two proteins are in the network the more similar are their biological functions[22]. Next, using GOGO similarities and the hierarchical clustering analysis (HCA), all proteins in the graph were partitioned into 30 (HCA-30), 100 (HCA-100), and 300 (HCA-300) clusters. Only those nodes belonging to the same cluster are allowed to be merged during graph reduction.

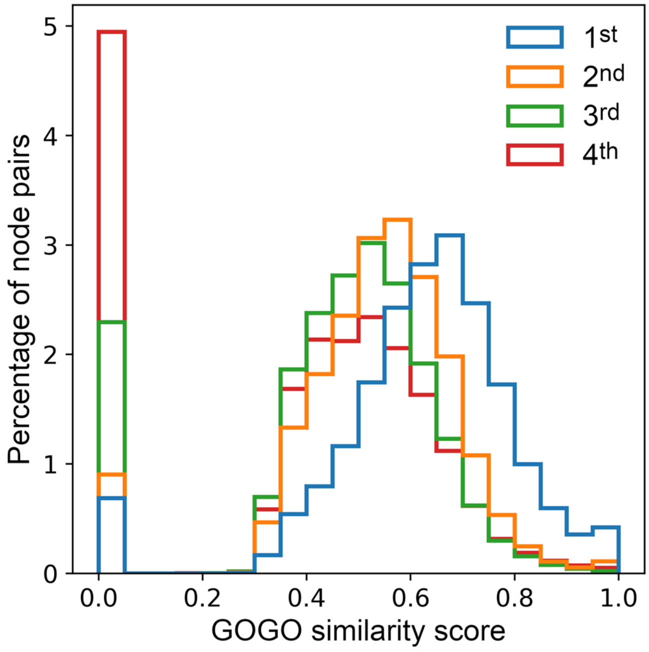

Fig. 2 Histogram of the pairwise GOGO similarity scores across the protein-protein interaction network. GOGO similarities are calculated using the biological process ontology for 1st, 2nd, 3rd, and 4th order neighbors in the network.

The reduction procedure for a graph is schematically shown in Fig. 3. An edge may be contracted only if both incident nodes are non-kinases, share the same differential gene expression, and belong to the same GOGO cluster. Yellow squares in Fig. 3A delineate groups of nodes that can be merged by contracting edges connecting them. The resulting reduced graph shown in Fig. 3B has the same number of kinases (circles), but fewer non-kinase proteins (rounded squares and diamonds representing merged nodes). The rationale behind this procedure is not only to reduce the size of a graph, but also to create more diversity among cell lines, which is highly beneficial for further machine-learning applications. Note that the reduction is performed on different cell lines without any drug information because we consider only the differential gene expression, the type of node (kinase or non-kinase protein), and the biological process assignment of nodes, with the latter two properties being independent on the cell line type.

Analysis of full-size and reduced networks

Compared to the original PPI network (Fig. 3C), the reduced graph (Fig. 3D) has a much higher ratio of kinase (red dots) to non-kinase (green dots) proteins. Table 1 shows that after the reduction, the number of nodes decreased from 19,144 to 1349 and the graph density increased from 0.004 to 0.014, saving computational resources in future graph processing. Further, a higher clustering coefficient for reduced graphs implies that they have better structures in terms of the information exchange than full-size networks. A high maximum betweenness centrality is to be expected since reduced graphs contain a super-hub node connected to many other nodes in the graph. A significant increase of the average betweenness centrality indicates that the information is going to flow more efficiently throughout reduced graphs. Overall, statistics reported in Table 1 demonstrate that in contrast to full-size networks, reduced graphs are compact, yet information-rich offering an effective representation of the biological data for machine-learning applications.

The graph reduction procedure greatly increases the diversity of graph topologies and features across all 359 cancer cell lines, while maintaining the important information and biological knowledge in each graph. In order to demonstrate the effectiveness of our reduction scheme, we calculated the information gain/loss after the reduction using the Shannon entropy of features and the graph-feature entropy (see Eq. 3). Ideally, the graph reduction should increase the information content for features without any decrease in the graph-feature information. The information gain/loss is shown in Fig. 4 for several reduction schemes. Although the simplest reduction requiring incident nodes to share at least one GO-BP term increases the feature-only entropy by 2.3 ± 0.6 (green bar), it causes a detrimental decrease of the graph-feature entropy by -0.4 ± 0.04 (red bar). In contrast, HCA increases the feature-only entropy while preserving the valuable graph-feature information. In particular, partitioning nodes into 30 clusters not only yields the highest information gain for features of 3.5 ± 0.9, but also slightly increases the graph-feature entropy. Based on these results and due to the fact that there are actually 30 level-1 biological processes in GO[23], we incorporated HCA-30 into the graph reduction procedure.

Comparison of matrix- and graph-based machine-learning approaches

In order to properly evaluate the performance of methods to predict the response of cancer to drugs, we conducted 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

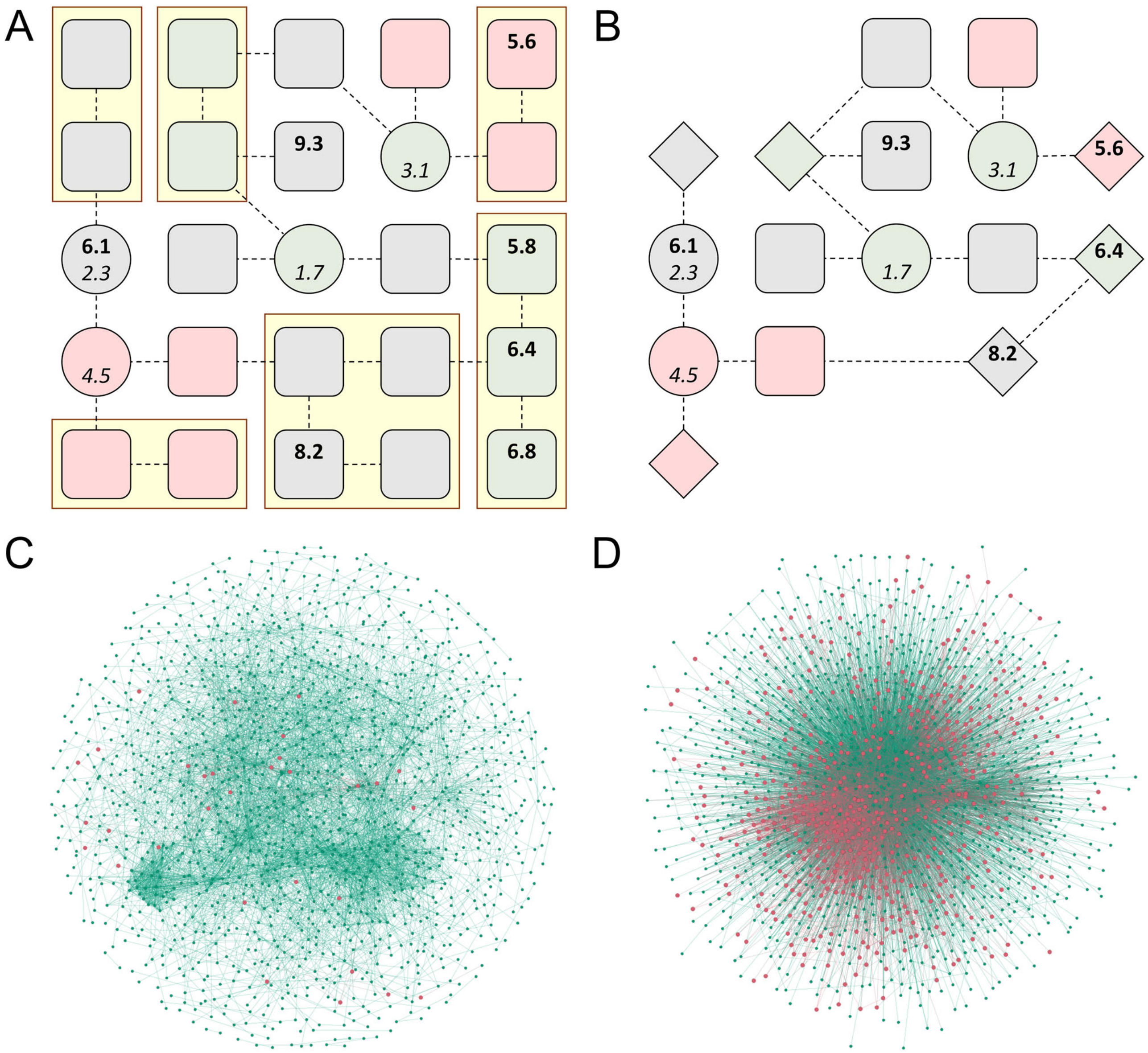

Fig. 3 Graph reduction of cancer-specific networks. **A** A schematic of the initial graph with yellow boxes outlining groups of nodes that can be merged by contracting their edges. **B** A schematic graph of the reduced graph in which merged nodes are represented by diamonds. **C** The initial (sub)network for glioblastoma (cell line A172) with red nodes representing kinases and green nodes representing other proteins. **D** The reduced network for glioblastoma colored the same as in (**C**). The network in (**C**) is a randomly sampled subgraph from the original network with the same number of nodes as (**D**).

using cancer 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. In addition, there is also a desired feature variability 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.

Table 2 reports the cross-validated performance of matrix- and graph-based algorithms. The first matrix-based approach employs a multilayer perceptron (MLP) to the differential gene expression data along with ligand embeddings. This method represents an appropriate baseline because many available tools reported in the literature utilize similar data frameworks[24]. The cross-validated accuracy of the baseline algorithm is only 0.55 indicating that the model is unable to discover the underlying patterns connecting the genomic data and small molecule descriptors. We also tested another matrix-based approach utilizing the same set of features as that used in the graph-based model. The accuracy of this classifier indeed improved to 0.60 demonstrating that kinase inhibitor profiling and disease-gene associations are more effective compared to ligand embeddings. One possible reason for this improved accuracy is that chemically dissimilar drugs can exhibit similar pharmacological effects in cell lines sensitive to these compounds[25], which remains undetected by using ligand embeddings.

In addition to MLP, we tested two other machine-learning approaches, Support Vector Machines (SVM) and Random Forest (RF), both employing the Principal Component Analysis (PCA) to reduce the dimensionality of input feature vectors. Although using SVM-PCA yields a higher accuracy than MLP, the Matthews correlation coefficient[26] and F-score values are notably lower. These results demonstrate that MLP achieves a robust

Table 1. Properties of full-size and reduced graphs.

Property	Full- size graph	Reduced graphs
Number of nodes	19,144	1349 ± 80
Number of edges	685,198	12,613 ± 608
Average degree	71	19 ± 0.3
Density	0.004	0.014 ± 0.0009
Diameter	8	4.073 ± 0.26
Clustering coefficient	0.287	0.659 ± 0.006
Maximum betweenness centrality	0.021	0.596 ± 0.011
Average betweenness centrality	1.11×10^{-4}	$7.88 \times 10^{-4} \pm 4.49 \times 10^{-6}$

Statistics are calculated from the graph topology without considering node features. Values for reduced graphs are reported as the average ±standard deviation across the dataset.

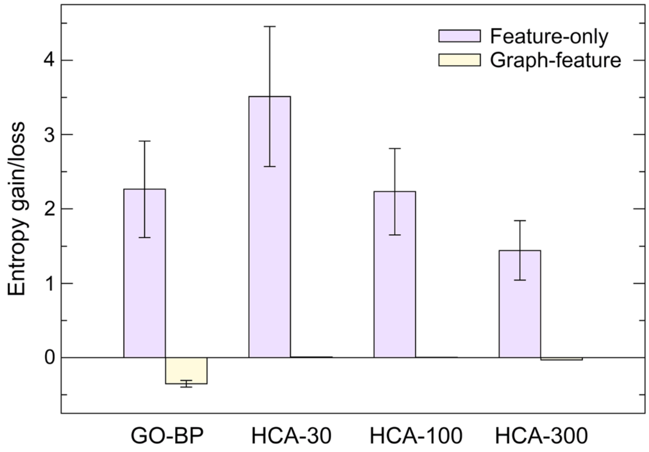

Fig. 4 Entropy gain/loss for different reduction schemes. Purple bars represent the Shannon entropy calculated using the feature matrix only, while yellow bars correspond to the graph-feature entropy computed using both feature and topological information of a graph. GO-BP requires that two incident nodes have a common biological process term to be assigned to the same cluster. HCA bars correspond to the clustering using GOGO similarities into 30, 100, and 300 clusters.

performance across four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally to the size of positive and negative instances[27]. It is noteworthy that similar to MLP, SVM-PCA and RF-PCA perform better when kinase inhibitor profiling and disease-gene associations are used instead of ligand embeddings. Encouragingly, a graph-based approach employing the WL Tree, a widely adopted graph kernel method for graph machine learning[10], yields the highest classification accuracy of 0.68, the MCC of 0.32, and the F-score of 0.65, outperforming all matrix-based algorithms. These benchmarking calculations demonstrate that the non-Euclidean representation offers a powerful approach to fully utilize biological data in machine learning.

Table 2. Performance of algorithms to predict the response of cancer cell lines to drugs.

Data type	Model	Features	ACC	PPV	TPR	MCC	F-score
Matrix	MLP	DGE, LE	0.55	0.63	0.64	0.27	0.55
Matrix	MLP	DGE, KIP, DGA	0.60	0.60	0.60	0.20	0.60
Matrix	SVM-PCA	DGE, LE	0.61	0.70	0.51	0.10	0.40
Matrix	SVM-PCA	DGE, KIP, DGA	0.62	0.72	0.53	0.16	0.45
Matrix	RF-PCA	DGE, LE	0.42	0.56	0.52	0.06	0.33
Matrix	RF-PCA	DGE, KIP, DGA	0.44	0.56	0.53	0.09	0.39
Graph	WL Tree	DGE, KIP, DGA	0.68	0.67	0.65	0.32	0.65

A graph-based approach is compared to two matrix-based methods. The performance of each algorithm is cross-validated at the tissue level.
MLP multilayer perceptron, *SVM-PCA* Support Vector Machines with Principal Component Analysis, *RF-PCA* Random Forest with Principal Component Analysis, *WL Tree* Weisfeiler–Lehman graph kernel, *DGE* differential gene expression, *LE* ligand embeddings, *KIP* kinase inhibitor profiling, *DGA* disease-gene associations, *ACC* accuracy, *PPV* precision, *TPR* recall, *MCC* Matthews correlation coefficient.

DISCUSSION

In this communication, we describe a new procedure to construct cancer-specific biological networks by integrating multiple heterogeneous data, including protein-protein interactions, differential gene expression, gene-disease associations, and kinase inhibitor profiling. In order to improve the effectiveness of machine learning, a graph reduction protocol was implemented to convert initially information-sparse biological networks to information-rich graphs preserving the biological knowledge. In contrast to the original PPI networks that share exactly the same topology, the reduced graphs provide a diverse set of graph topologies across different cell lines. In addition, because of the much smaller size compared to full-size PPI networks, the reduced graphs can be used with any machine-learning model, even those that are computationally expensive. Overall, our procedure to construct cancer-specific molecular networks enables a more efficient learning process on account of the more diverse and information-rich data. This approach improves the performance of machine learning to predict the responsiveness of cancer to pharmacotherapy.

MATERIALS AND METHODS

Protein-protein interaction network

Protein-protein interaction data were acquired from the STRING database[28] that comprises 19,354 human proteins forming 11,355,804 interactions including those identified experimentally and predicted computationally. Each interaction is assigned a confidence score ranging from 150 for low-confidence to 999 for high-confidence interactions. A human PPI network was constructed from confident interactions with a score of ≥500. Single proteins disconnected from the main network and those forming small, isolated networks were removed, resulting in the final PPI network containing 19,144 proteins and 685,198 interactions.

Differential gene expression data

Original gene expression data are available from the Cancer Cell Line Encyclopedia (CCLE) project comprising a detailed genetic characterization of a large number of cancer cell lines[29]. We obtained the curated CCLE data from Harmonizome, a comprehensive repository of processed genomics, proteomics, epigenomics, transcriptomics, and metabolomics data[30]. The differential gene expression (DGE) data contain 18,022 genes, 1035 cancer cell lines, and 749,551 gene-cell associations categorized as down-, up-, and normally regulated with respect to the expression level in healthy cells.

Kinase inhibitor profiling

Inhibitor profiling refers to a large-scale experimental measurement of the activity of an inhibitor against a panel of target proteins. The kinase inhibitor profiling (KIP) data used in this study were collected and curated by Team-SKI[31]. The activity is reported as a pIC$_{50}$ value, which is the negative logarithm of the half-maximal inhibitory concentration (IC$_{50}$) measuring the potency of a compound in inhibiting a specific biochemical function. The cutoff for pIC$_{50}$ values was set at 6.3 corresponding to the IC$_{50}$ of 500 nM. The Team-SKI dataset contains 49,348 small molecules tested against 411 protein kinases.

Disease-gene associations

Disease-gene association (DGA) scores were obtained from two sources, DISEASES[32] and DisGeNET[33]. The DISEASES database integrates evidence on associations collected through automatic text mining, manually curated from biomedical literature, cancer mutation data, and genome-wide association studies. It contains 8330 diseases and 20,715 genes with association scores ranging from 1 to 10. The DisGeNET database provides information on associations pulled from various repositories including Mendelian, complex and environmental diseases, and integrated using gene/disease vocabulary mapping and the DisGeNET ontology. It comprises 24,166 diseases and 17,545 genes with association scores ranging from 0.01 to 1.

Data integration

Data integration was performed by mapping DGE, KIP, and DGA values directly onto the PPI network. Gene expression data contains normalized scores for 18,022 genes, whereas the PPI network comprises 19,144 proteins. The missing values for nodes in the graph were replaced by median scores calculated over 1st order neighbors. The kinase component of the PPI network was determined by running Basic Local Alignment Search Tool (BLAST)[34] against known human kinases[35]. Using a similarity threshold of 95% identified 508 kinases. pIC$_{50}$ values are available from the Team-SKI database for 411 kinase proteins in our network and 29 small molecules that are also present in the growth rate inhibition data from LINCS. The Disease Ontology ID (DOID) and the Concept ID for each cancer cell line were identified with the Cellosaurus resource portal[36], and mapped to, respectively, DISEASES and DisGeNET databases in order to assign disease-gene association scores to nodes in the PPI network. After integrating all data and removing cases that could not be mapped, the final dataset contains annotated graphs for 3,549 combinations of 359 cell lines and 29 drugs.

Growth rate inhibition data

Proliferation is measured by GR$_{50}$ and GR$_{max}$ indices quantifying the value of growth rate inhibition (GR) based on time course and endpoint assays[37]. GR$_{50}$ is the concentration of a drug at which GR is 0.5, whereas GR$_{max}$ is the maximum measured GR value. In this study, we employ GR$_{max}$ values for the following 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[37]. The benchmarking dataset comprises 2124 positive instances of cell line-drug combinations with negative GR$_{max}$ values signifying a cytotoxic response, and 1425 negative instances with positive GR$_{max}$ values signifying a cytostatic response.

Graph partitioning

The latest release of the GO database contains 29,698 BP, 11,147 MF, and 4201 CC ontology terms[21]. Each DAG has its own hierarchy maintained by the domain-centric Gene Ontology (dcGO). According to the dcGO database, the numbers of level-1 nodes are 30, 15, and 22 for BP, MF, and CC, respectively[23]. Since our goal was to assign nodes to biological pathways, all-against-all semantic similarities between GO terms of PPI network nodes were calculated with the GOGO software based on the BP-DAG topology. The matrix of GOGO scores was then subjected to a hierarchical clustering analysis, which builds nested clusters by successively merging and splitting clusters[38]. There are two types of hierarchical clustering, agglomerative (bottom-up) and divisive (top-down). Agglomerative methods merge observations as moving up the hierarchy while divisive methods split the observations as moving down the hierarchy. Here, we employed the agglomerative HCA rather than other commonly used clustering methods, such as k-means[39], DBSCAN[40], and affinity propagation[41] because this technique is the most appropriate for GOGO semantic similarities and we found empirically that the clustering results represent well different levels of biological process in GO.

Graph statistics

Average degree of a graph is calculated by averaging the degrees of all nodes, defined as the mean number of connections to other nodes in the graph.

Density, ρ_G, is defined as the ratio of the number of edges and the number of possible edges. For an undirected graph, it is calculated as

$$\rho_G = \frac{2|E|}{|V|(|V|-1)} \tag{1}$$

where $|E|$ is the number of edges and $|V|$ is the number of nodes. Since the information is propagated more efficiently across dense graphs, increasing the graph density is generally beneficial for graph learning algorithms, such as GNN.

Diameter is the longest shortest path between any two nodes in a graph[19], i.e. the longest distance one must traverse from one node to reach another node. This is an important property because it directly informs the design of a graph model.

Clustering coefficient measures the tendency of nodes to form clusters. Most real-world graphs, such as social, powerline, citation, language, food, economic, and biological networks, tend to form communities, which are characterized by a high density of ties[42]. The higher the clustering coefficient, the more likely nodes are to form communities resulting in a shift of the information exchange from global to local.

Betweenness centrality of a node is the sum of the fraction of all-pair shortest paths passing the node[43]. Essentially, this property evaluates the amount of information flowing through nodes, i.e., a high betweenness centrality signifies more information flowing through a graph. Further, those nodes with a high information flow interact more frequently with other nodes, maintaining a significant control over the entire network.

Graph-feature entropy

A graph reduction basically compresses the information in a graph. However, as all lossy compression schemes, some information will be lost during the reduction. In order to measure how much meaningful information is lost or gained as opposed to the loss/gain of the redundant or irrelevant information, we introduce a concept of the graph-feature entropy. Briefly, the graph-feature entropy S of a graph G, S_G, is based on the Shannon entropy[40] defined as:

$$S_G \overset{\text{def}}{=} -\sum_j p(z_j)\log p(z_j) \tag{2}$$

where z_j is the column vector of the feature matrix filtered by the graph Laplacian, which corresponds to the j-th feature of all nodes in the graph. Using features filtered with the graph Laplacian effectively combines both feature and topological information of a graph providing a useful measure of the information content in featured graphs.

Next, we calculate the information gain/loss after the reduction, δ, using the Shannon entropy of features and the graph-feature-entropy:

$$\delta = \frac{S_{\text{reduced}} - S_{\text{original}}}{S_{\text{original}}} \tag{3}$$

where S_{original} is the entropy (either feature-only or graph-feature) of the original graph and S_{reduced} is the corresponding entropy of the reduced graph.

Matrix-based methods to predict drug response

The rows of input matrices are gene products, and the columns contain various features. The baseline matrix includes only the DGE data, and after flattening to a 19,144-element vector, it is concatenated to a 300-element vector of ligand embeddings (LE) computed by Mol2vec[41] for the input drug. The baseline classifier utilizing 19,444-element input feature vectors represents an archetypal approach to predict a drug response based solely on the gene expression changes and the drug chemical structure. The second method utilizes a full input matrix of 19,144 gene products × 3 features, DGE, KIP, and DGA, flattened to a 57,432-element vector. This approach is a matrix-based equivalent of the graph-based method utilizing exactly the same input features. Both matrix-based classifiers employ an MLP with an input layer of either 19,444 units for the baseline or 57,432 units for the graph-equivalent method, three hidden layers comprising 1024, 512, and 256 units, and a 2-unit output layer returning probabilities for the input drug to be effective and ineffective against the cell line.

Graph-based method to predict drug response

The graph-based approach devised in this study employs the WL graph kernel, also known as WL Tree classifier[10], to the reduced representations of cancer-specific PPI networks comprising DGE, KIP, and DGA data. WL Tree is a widely adopted method for learning from graphs with discrete node labels. A key component of this algorithm is a rapid feature extraction employing the WL test of isomorphism on graphs. Briefly, WL Tree iteratively maps the original node labels to new node labels based on the topological structure of the graph. The new node labels are then compressed and integrated with the original node labels in order to form a vector representation for the graph. The implementation of WL Tree used in this study was previously demonstrated to outperform other graph kernels on several graph classification benchmark datasets in terms of accuracy and runtime.

DATA AVAILABILITY

Data generated in this study are freely available to the academic community at https://osf.io/dzx7b/.

Received: 29 September 2021; Accepted: 4 April 2022;
Published online: 29 April 2022

REFERENCES

1. Gyurko, D. M. et al. Adaptation and learning of molecular networks as a description of cancer development at the systems-level: potential use in anti-cancer therapies. *Semin. Cancer Biol.* **23**, 262–269 (2013).
2. Klinke, D. J. 2nd Signal transduction networks in cancer: quantitative parameters influence network topology. *Cancer Res.* **70**, 1773–1782 (2010).
3. Liu, Y. et al. A multiscale computational approach to dissect early events in the Erb family receptor mediated activation, differential signaling, and relevance to oncogenic transformations. *Ann. Biomed. Eng.* **35**, 1012–1025 (2007).
4. Chen, C. et al. Construction and analysis of protein-protein interaction networks based on proteomics data of prostate cancer. *Int J. Mol. Med.* **37**, 1576–1586 (2016).
5. Guda, P., Chittur, S. V. & Guda, C. Comparative analysis of protein-protein interactions in cancer-associated genes. *Genomics Proteom. Bioinform.* **7**, 25–36 (2009).
6. Kanhaiya, K. et al. Controlling directed protein interaction networks in cancer. *Sci. Rep.* **7**, 10327 (2017).
7. Erten, S., Bebek, G. & Koyuturk, M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J. Comput. Biol.* **18**, 1561–1574 (2011).
8. Sun, J. et al. A systematic analysis of FDA-approved anticancer drugs. *BMC Syst. Biol.* **11**, 87 (2017). Suppl 5.
9. Bronstein, M. M. et al. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
10. Shervashidze, N. et al. Weisfeiler-Lehman graph kernels. *J. Mach. Learn Res.* **12**, 2539–2561 (2011).
11. W, H. L. Graph representation learning. *Synth. Lectures Artif. Intell. Mach. Learn.* **14**, 1–159 (2020).
12. Google. *Google*. https://www.google.com/search/howsearchworks/. Available from: https://www.google.com/search/howsearchworks/.
13. Zheng, Y. & Wu, Z. A machine learning-based biological drug-target interaction prediction method for a tripartite heterogeneous network. *ACS Omega* **6**, 3037–3045 (2021).
14. Kong, J. et al. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nat. Commun.* **11**, 5485 (2020).
15. Lee, T. Y. et al. RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res.* **39**, D777–D787 (2011).
16. Golub, G. H. & Van Loan, C. F. *Matrix Computations* 3rd edn. *Johns Hopkins Studies In The Mathematical Sciences* xxvii, 694pp (Johns Hopkins University Press, 1996).
17. Rosen, K. H., *Discrete Mathematics and its Applications*. 7th edn. (McGraw-Hill, 2012).
18. Gross, J. L. & Yellen, J. *Graph Theory and its Applications. The CRC Press Series on Discrete Mathematics and its Ap plications* 585pp (CRC Press, 1999).
19. West, D. B., *Introduction to Graph Theory* 2nd edn, xix, 588pp (Prentice-Hall, 2001).
20. Zhao, C. & Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Sci. Rep.* **8**, 15107 (2018).
21. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* **25**, 25–29 (2000).
22. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
23. Fang, H. & Gough, J. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.* **41**, D536–D544 (2013).
24. Chang, Y. et al. Cancer Drug Response Profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.* **8**, 8857 (2018).
25. Chen, B. et al. Relating chemical structure to cellular response: an integrative analysis of gene expression, bioactivity, and structural data across 11,000 compounds. *CPT Pharmacomet. Syst. Pharm.* **4**, 576–584 (2015).
26. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
27. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
28. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
29. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
30. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* **2016**, baw100 (2016).
31. Sorgenfrei, F. A., Fulle, S. & Merget, B. Kinome-wide profiling prediction of small molecules. *ChemMedChem* **13**, 495–499 (2018).
32. Pletscher-Frankild, S. et al. DISEASES: text mining and data integration of disease-gene associations. *Methods* **74**, 83–89 (2015).
33. Pinero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
34. Altschul, S. F. et al. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
35. Manning, G. et al. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
36. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech.* **29**, 25–38 (2018).
37. Hafner, M. et al. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* **13**, 521–527 (2016).
38. Maimon, O. & Rokach L., *Data Mining and Knowledge Discovery Handbook* 2nd edn, xx, 1285pp. (Springer, 2010).
39. MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In: Le Cam, L. M. & Neyman, J. (eds) *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, 1967).
40. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
41. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. modeling* **58**, 27–35 (2018).
42. P, H. W. & L., S. Transitivity in structural models of small groups. *Comp. Group Stud.* **2**, 107–124 (1971).
43. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**, pp.35–41 (1977).

ACKNOWLEDGEMENTS

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. Portions of this research were conducted with computing resources provided by Louisiana State University.

AUTHOR CONTRIBUTIONS

L.P. conceptualized graph-based machine learning, developed machine-learning algorithms, performed computations, and drafted the manuscript. M.S. conceptualized biological data utilization, performed data acquisition and integration, and drafted the manuscript. H.C.W. and C.B. conceptualized graph algorithms. J.R. conceptualized machine learning and acquired funding. M.B. conceptualized graph algorithms and machine learning, supervised the project, acquired funding, and prepared the final manuscript. All authors reviewed and approved the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Michal Brylinski.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.