*Bioinformatics*, 36(12), 2020, 3818–3824
doi: 10.1093/bioinformatics/btaa203
Advance Access Publication Date: 24 March 2020
Original Paper



Systems biology

# Cancer subtype classification and modeling by pathway attention and propagation

**Sangseon Lee[1], Sangsoo Lim[2], Taeheon Lee[1], Inyoung Sung[3] and Sun Kim [1,2,3,*]**

[1]Department of Computer Science and Engineering, Institute of Engineering Research, [2]Bioinformatics Institute and [3]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 17, 2019; revised on January 13, 2020; editorial decision on March 9, 2020; accepted on March 19, 2020

## Abstract

**Motivation:** Biological pathway is an important curated knowledge of biological processes. Thus, cancer subtype classification based on pathways will be very useful to understand differences in biological mechanisms among cancer subtypes. However, pathways include only a fraction of the entire gene set, only one-third of human genes in KEGG, and pathways are fragmented. For this reason, there are few computational methods to use pathways for cancer subtype classification.

**Results:** We present an explainable deep-learning model with attention mechanism and network propagation for cancer subtype classification. Each pathway is modeled by a graph convolutional network. Then, a multi-attention-based ensemble model combines several hundreds of pathways in an explainable manner. Lastly, network propagation on pathway–gene network explains why gene expression profiles in subtypes are different. In experiments with five TCGA cancer datasets, our method achieved very good classification accuracies and, additionally, identified subtype-specific pathways and biological functions.

**Availability and implementation: The source code is available at http://biohealth.snu.ac.kr/software/GCN_MAE.:**

**Contact:** sunkim.bioinfo@snu.ac.kr

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Biological systems are too complex to understand as a whole. For this reason, biological systems are dissected to small subsystems that can be easily understood. The most widely used subsystems are biological pathway databases that are curated for years and these pathway databases, such as KEGG (Kanehisa and Goto, 2000), are widely used to analyze transcriptome data.

### 1.1 Motivation

Cancer subtypes are often classified based on gene expression profiles. For example, breast cancer is well characterized in terms of subtypes that are widely used for clinical applications (Grimm *et al.*, 2015; Hwang *et al.*, 2019). However, cancer subtype classification based on gene expression profiles showed poor stability on independent datasets (Alcaraz *et al.*, 2017; Kim *et al.*, 2012). Furthermore, it does not provide insightful biological information, such as subtype-specific activations of certain pathways (Gatza *et al.*, 2010; Segura-Lepe *et al.*, 2019). Thus, pathway-based cancer subtype classification is desirable since pathways can be an effective way to generate landscape of molecular functions of an organism as a collection of biological knowledge (Viswanathan *et al.*, 2008). While pathway databases contain static information in general, mapping transcriptome data to the pathways can enhance their usefulness by explaining the dynamics of cancer in terms of biological functions (Kunz *et al.*, 2019; Schadt *et al.*, 2005). Another view on why pathway-based cancer subtype classification is useful can be explained in terms of the number of dimensions or variables since the dimensionality from genes to pathways is two orders of magnitude smaller (20 000 versus 300), resulting in better interpretability on the feature space (Gatza *et al.*, 2010; Glaab *et al.*, 2010; Su *et al.*, 2009). However, there is a serious issue when pathways are used for cancer subtype classification. Only a fraction, one-third in the case of humans, is included in biological pathways and use of pathways is limited in predictive power for subtype classification, compared to use of the entire transcriptome. Thus, the main research question here is: *how can we use pathways for cancer subtype classification?* Most of pathway analysis tools are developed to measure pathway activation levels (Lim *et al.*, 2020). Given that each pathway is modeled as a single value of representing the activation status of the pathway, combining these values for the entire biological system is not straightforward, often resulting in poor performances of cancer subtype classification (Lim *et al.*, 2020).

In this study, we propose a deep-learning approach to investigate three important research questions:

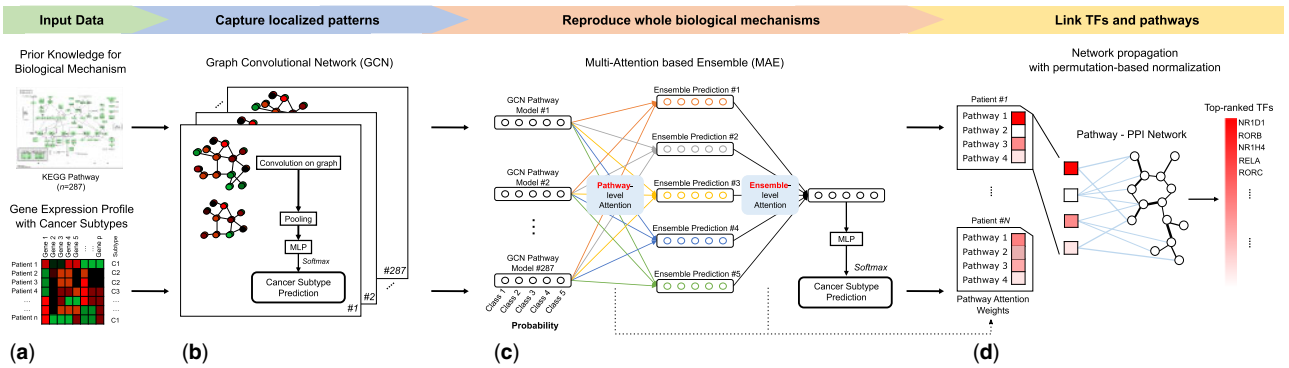



**Fig. 1.** The workflow of the proposed model. Given (**a**) gene expression data and pathway set, the proposed model consists of three major parts. (**b**) Using GCNs, each GCN pathway model captures localized gene expression patterns. Then, MLP is followed by GCN to encode extracted gene-level information into pathway-level. (**c**) MAE combines the outputs of GCN pathway models. To consider complex pathway combinations of cancer, two attention layers named pathway-level and ensemble-level are utilized. (**d**) To identify TFs related to highlighted pathways from (c), network propagation on a pathway–PPI network is performed. Seed values are determined by pathway attention weights from the attention model. To avoid that propagation is over-fitted on the high-degree nodes, permutation-based normalization of TFs is considered

1. How can we improve accuracy in predicting cancer subtypes using transcriptome data in terms of pathways?
2. How different are pathway interactions among cancer subtypes?
3. Why are gene expression profiles different among cancer subtypes?

### 1.2 Backgrounds

To begin with, modeling individual pathways, we need an effective computational method that can consider interactions among genes. Due to recent advances in deep learning, *graph convolutional network* (*GCN*) can handle these interactions instead of traditional pathway analysis tools. Aggregation of node features (gene expression levels in this study) on the graph are performed in various ways, such as spectral graph convolution (Defferrard *et al.*, 2016; Dhillon *et al.*, 2007), layer-wise propagation (Kipf and Welling, 2017), diffusion process (Atwood and Towsley, 2016; Monti *et al.*, 2017) and graph embedding for sparse connections between nodes (Kong and Yu, 2018). Like convolutional neural network (CNN), these GCN can capture localized patterns in data, and unlike CNN, it can be used for non-grid structured data, such as graph. For these reasons, GCN has been successfully used in protein–protein interaction (PPI) network for the prediction of breast cancer subtype and drug side effects (Rhee *et al.*, 2018; Zitnik *et al.*, 2018).

Given a GCN model for each pathway, interactions between pathways can be considered as a network of pathways by combining several hundreds of pathway models again. For example, a condition-specific pathway network can be built from transcriptome data (Moon *et al.*, 2017) and we could use GCN again for combining several hundreds of pathway models. However, GCN is a deep-learning model which is a black-box model that cannot explain which input features are important and why the model performs well (Castelvecchi, 2016). To open up the black-box model, *attention mechanism* is frequently used (Vaswani *et al.*, 2017). The attention mechanism helps identify features that make the models achieve better performances (Choi *et al.*, 2016; Zheng *et al.*, 2017).

Another important question is how to explain the differences between gene expressions and interactions among subtypes in terms of pathways. The question then is how different *biological functions* among cancer subtypes by extending pathway-level information to gene-level (Jo *et al.*, 2016). In this regard, *network propagation* is also widely used in the network analysis for biological interpretation (Cowen *et al.*, 2017; Pearson, 1905). For example, network propagation has been successful in aggregating mutation profiles on the molecular interaction networks to detect significant gene modules (Hofree *et al.*, 2013; Leiserson *et al.*, 2015; Zhang *et al.*, 2018).

### 1.3 Our approach: pathway attention and propagation

In this study, we propose an *explainable deep-learning* model (Gunning, 2017) for cancer subtype classification and pathway modeling. The model consists of three steps (Fig. 1). To begin with, a pathway model is generated for each of the pathways by GCN to utilize biological prior knowledge respectively. Then, multiple GCN pathway models are integrated into a single model by *multi-attention-based ensemble* (*MAE*). The MAE model consists of two-level attentions to capture complex pathway combinations of cancer data. Finally, to show how different subtypes are in terms of biological functions, we propose a network propagation method with permutation-based normalization for identification of transcription factors (TFs) that influence gene expressions and pathways. In the following sections, Section 2 explains detailed implementation of the model. In Section 3, we demonstrate the power of our model in experiments with five cancer datasets.

## 2 Materials and methods

### 2.1 Encoding biological prior knowledge using GCN

Given a pathway $p$ as prior knowledge, a graph $G^p = (V^p, E^p)$ is determined, where $V^p$ is a set of nodes representing genes and $E^p$ is a set of edges representing molecular interactions between genes in the pathway $p$. Gene expression profile from RNA-seq is mapped to nodes $V^p$ that are represented as a vector $X_i^p = (x_{i,1}^p, x_{i,2}^p, \ldots, x_{i,m_p}^p)$, where $i$ is an index for each patient and $m_p$ is the number of genes in the pathway $p$ ($m_p = |V^p|$).

To capture localized gene expression patterns in $G^p$, a spectral convolutional approach is applied on the Laplacian matrix $L^p = D^p - A^p$ of a graph (Bruna *et al.*, 2014; Defferrard *et al.*, 2016). Here, $D^p$ is a weighted degree matrix of $G^p$ and $A^p$ is an adjacency matrix of $G^p$. Based on an eigenvalue decomposition of the graph Laplacian matrix $L^p = U\Lambda U^T|_p$, the spectral convolutional operator is defined as

$$L^p_{spectral} = U g_\theta(\Lambda) U^T X|_p, \tag{1}$$

where $g_\theta(\Lambda)|_p$ is a polynomial convolution filter that is applied on the diagonal matrix $\Lambda^p$.

$$g_\theta(\Lambda)|_p = \sum_{k=0}^{K-1} \theta_k \Lambda^k|_p. \tag{2}$$

$g_\theta(\Lambda)|_p$ is represented as a $K$-order polynomial function with weight parameters $\theta_k$ that works as a convolution filters reaching $K$-hop neighbors. This way, the spectral convolutional operator [Equation (1)] can capture localized expression patterns in $K$-hop neighbor nodes in a graph. Despite this advantage, it is difficult to use a polynomial convolution filter as is since it takes $O(n^2)$ time to calculate

the polynomial filter. In a recent study (Hammond *et al.*, 2011), an approximated polynomial function called Chebyshev expansion is proposed. Using the Chebyshev polynomial approximation, the spectral convolutional filter is re-defined as

$$L^{p}_{spectral} \approx U[\sum_{k=0}^{K-1} \theta'_k T_k(\tilde{\Lambda})]U^T X|_p, \quad (3)$$

where $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0 = 1$ and $T_1 = x$, and $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_{m_p}$.

We use the above filter as a convolutional filter. Then, extracted patterns are pooled with neighboring nodes using the Graclus algorithm. We found empirically that max pooling performed better than average pooling for the pathway models, thus, max pooling was used. After convolution and pooling, gene expression profiles are dimensionally reduced into pathway-level vectors. In turn, these vectors are given to a multilayer perceptron (MLP) and then they concerted to subtype-wise probability vectors.

The entire structure of GCN pathway models is shown in Supplementary Figure S1. To deal with high-dimension low sample characteristics, over 20 000 dimensions of genes and typically <1000 samples, of transcriptome data, dropout and shallow networks for GCN and MLP are used to avoid over-fitting. Cross-entropy loss is used as a cost function.

### 2.2 Re-producing comprehensive biological process by MAE

As described in Section 2.1, a GCN pathway model is built for each pathway. Using these GCN pathway models, gene expression profile $X_i$ for each patient is converted into $P$ number of encoded vectors $h^p(X_i)$ (as shown in the Supplementary Fig. S1), where $p = 1, 2, \ldots, P$ (total number of pathways). To combine encoded vectors of hundreds of pathways, attention mechanism is used. Each encoded vector $h^p(X_i) \in \mathbb{R}^d$, where $d$ corresponds to the number of cancer subtypes, is concatenated, resulting in a large matrix form $h(X_i) \in \mathbb{R}^{P\times d}$. Attention scores are calculated on the concatenated matrix $h(X_i)$ and attention-based combination of pathway vectors $\tilde{h}(X_i)$ is generated below as in Equation (4).

$$\begin{aligned} W &\in \mathbb{R}^{d\times a}, b \in \mathbb{R}^a, u \in \mathbb{R}^a \\ Y &= \tanh(h(X_i)W + b) \in \mathbb{R}^{P\times a} \\ \alpha &= \mathrm{Softmax}(Yu) \in \mathbb{R}^P \\ \tilde{h}(X_i)_j &= \sum_{k=1}^{P} h^k(X_i)_j \alpha_k \\ &\text{where } \tilde{h}(X_i) \in \mathbb{R}^d \end{aligned} \quad (4)$$

Our MAE model operates at two hierarchical levels (Fig. 1c): pathway-level attention and ensemble-level attention. The basic mechanism for the pathway-level attention is the same as in Equation (4), but multiple attention mechanisms are used to capture various combinations of pathway encoded vectors. Each attention mechanism generates $\tilde{h}(X_i)|_l \in \mathbb{R}^d$, where $l$ is $l$-th attention mechanism. As an ensemble-level attention, multiple pathway-level attention encoded vectors are concatenated, resulting in a form of $\tilde{h}_{MGD}(X_i) = (\tilde{h}(X_i)|_1; \tilde{h}(X_i)|_2; \ldots; \tilde{h}(X_i)|_L) \in \mathbb{R}^{L\times d}$, where $L$ is the number of pathway-level attention. Then, as in Equation (4), an ensemble-level attention vector $\tilde{h}_{fin}(X_i) \in \mathbb{R}^d$ is computed. After the MAE step, $\tilde{h}_{fin}(X_i)$ is used as input to two-layer fully connected MLP for the cancer subtype classification. Cross-entropy loss is used as an objective function.

### 2.3 Linking pathways and TFs by network propagation with permutation-based normalization

Our approach of combining hundreds of pathways using multi-attention models does provide some insights on how pathways interact differentially among cancer subtypes. Investigation on the difference in gene interaction among subtypes is much more complicated because the number of genes is almost two orders of magnitude larger than the number of pathways. Here, we propose an effective approach of investigating gene interactions using transcriptome data by linking pathways and TFs using network propagation with permutation-based normalization.

Network propagation is typically done by performing random walks on a network. A random work starts with seed nodes that are pre-selected and the seeds have certain amount of information to be propagated. However, performing a random walk on a long path will dilute the information too much, especially when hub nodes with many edges are involved. To avoid this dilution problem, a random walk with restart algorithm (Köhler *et al.*, 2008) is used.

The random walk with restart is calculated as below:

$$p^{(t+1)} = (1 - r)Wp^{(t)} + rp^{(0)}, \quad (5)$$

where $W$ is a column-wise normalized weighted adjacency matrix of a network and $p^{(t)}$ is a vector that contains the propagated values of each node at time step $t$. The seed vector $p^{(0)}$ is a normalized vector of initial values and $r$ is a restart parameter.

When performing the network propagation, constructing network topology and selecting seeds are two most important issues. In the case of the network topology, biological prior knowledge and gene expression data are utilized (Supplementary Fig. S2a). Based on a PPI from BIOGRID database (Stark *et al.*, 2006), an absolute value of Pearson's correlation is mapped on each edge in the network. To link pathways and the weighted PPI, pathway nodes are added in the network. Edges between a pathway node and genes in the pathway are also added with constant weight 1. On the pathway-PPI network, seed nodes are selected as the pathway nodes and values are assigned by attention weights from the GCN pathway models with multi-attention. The pathway attention weights are calculated using pathway-level attention values and ensemble-level attention values. The attention weight corresponding to one pathway is extracted from multiple pathway-level attentions, and the extracted values are summed by considering weights through the ensemble-level attention.

As a result of network propagation, all nodes in a network have propagated values. The propagated values are determined by not only the initial values of seed nodes but also the topology of a network. For example, if a node has a high degree, it may have a larger propagated value than other nodes regardless of seed values. To address this problem, a null distribution of propagated values is computed by a permutation-based approach (Supplementary Fig. S2b). Given pathway attention weights of patient samples, a pathway attention weight is randomly selected from the samples on each pathway and a new random patient is generated. By repeating this procedure 1000 times and performing network propagation on the random patient samples, a permutation-based network propagation values are generated. From the permutation result, each node in the network is ranked in terms of propagated values and a mean permutation rank is computed by averaging the ranks of all random patients. On a real patient, each node is transformed into ranks which are normalized by the mean permutation rank. Remember that our goal in this step is linking highlighted pathways and TFs. We used TFs that are curated in the literature (Lambert *et al.*, 2018) and TFs are ranked as a result of network propagation and normalization.

## 3 Results

### 3.1 Pathway database and cancer dataset

We used the KEGG pathway database. Since our goal is to construct a pathway-based model of transcriptome, we excluded some pathways that are not directly relevant. Specifically, pathways related to drug development were removed. Pathways that are physical clusters of genes were also excluded from the analysis. In addition, pathways with <5 genes were excluded since deep-learning models of these small pathways are not feasible. As a result, we used 287 pathways for modeling cancer subtypes. In total, 5515 genes were included in the 287 pathways, thus, the 5515 genes were used for our analysis. Graph representations of the pathways were extracted using KEGGgraph (Zhang and Wiemann, 2009) library in R.

The Cancer Genome Atlas (TCGA) RNA-seq datasets were used as gene expression profiles. RNA-seq datasets were downloaded from Firebrowse (http://firebrowse.org/). Five cancers with subtypes defined, BLCA, BRCA, COAD, PRAD and STAD, were analyzed. Subtype information for these cancer datasets except BRCA was from the original research papers of each cancer type. In the case of BRCA, the original article (Cancer Genome Atlas Network, 2012) classified subtypes based on microarray data and the number of samples was small since it was one of the early TCGA papers. Thus, BRCA subtypes were re-generated by the PAM50 classification method (Parker *et al.*, 2009) on log2-transformed RNA-seq data. Detailed information and references about cancer datasets are described in Supplementary Table S1.

## 3.2 Evaluation of individual GCN pathway models

Before constructing one unified model of transcriptome data, each pathway was modeled as a component. We performed experiments to see how well each GCN model classified cancer subtypes. Hyper-parameters of each GCN model were determined in 3-fold cross-validation (CV) within training data (Supplementary Fig. S3). Classification performances were measured in terms of weighted *F*1 score with 10-fold CV (Supplementary Tables S2 and S3).

The average classification accuracies were 76.39% for BLCA, 66.91% for BRCA, 71.54% for COAD, 70.12% for PRAD and 78.13% for STAD. There was a huge variation in performances. For example, in the case of BLCA, the maximum value was 90.98% whereas the worst-case value was 46.78%. Because these pathways reflect only a small part of the biological process, some of these pathways were highly correlated with cancer subtypes, but some did not. However, no pathways were commonly singled out in achieving the best performances in all five cancers. Thus, our goal of combining all pathways in one model is well supported by these experiments.

## 3.3 Performance of ensemble of GCN pathway models with multi-attention

### 3.3.1 Effectiveness of the MAE model of GCN pathway models

The first performance evaluation is to see how accurate it can be and how much performance gain can be achieved by combining all pathway models into one model with multi-attention. Hyper-parameters for the MAE model were determined in the same manner as described in Section 3.2.

*Performance gain of attention mechanism*: By combining all GCN pathway models to single GCN+MAE models, the performance gain was significant (Table 1). In COAD data, the GCN+MAE model with 11 attentions showed the best *F*1 score of 87.01% with 4.22% improvement over the best of single GCN pathway model. The other cancer data except STAD were also achieved over 2.7% improvements. Besides, ensemble of multi-attentions notably affected the performance gain. Most of the GCN+MAE models were showed over 2.0% performance gain than single attention without an ensemble-level attention (GCN+Single Att). Even with the GCN+Single Att models, the performances were also better than single GCN pathway model. The performance gains were smaller than 1.0% in three cancers, but over 2.0% performance gains were also observed in the other cancers (BRCA and COAD). Thus, these experiments show the effectiveness of attention mechanisms that combine hundreds of pathway models.

*Comparison with existing methods*: We compared our GCN+MAE model with other classification methods: SAS (Lim *et al.*, 2016), the pathway activity inference tool, and RAW that used all available 20 531 gene expression. The results are summarized in Table 1. A recent study (Lim *et al.*, 2020) compared 13 pathway activity inference tools and SAS showed reasonably good performances on various tests including cancer subtype classification. Thus, we chose SAS for performance comparison. As pointed out in the study (Lim *et al.*, 2020), a large portion, about 2/3, of gene expression information is lost. Thus, we also chose RAW to compare classification performances when 'all' genes are used. To classify cancer subtypes with SAS and RAW, we used support vector machine (SVM) and random forest (RF) classifier. In addition, pathway activities measured by SAS were trained by MLP classifier. In the experiments of comparing GCN+MAE model with SAS and RAW classifiers, the GCN+MAE model performed significantly better, at least above 2.9%, than both SAS and RAW classifiers. For example, in the case of COAD, SAS+SVM achieved best performance at 77.54%, which was significantly lower than the GCN model (82.79%) and the GCN+MAE model (87.01%). In fact, performance of the SAS classifier was worse than the RAW classifier, which showed that information loss of genes in pathways is substantial. To further explore the relationships between the pathway activity inference tools and gene-level gene sets, we tested 10 additional pathway inference tools and another gene set of cancer hallmarks (Hanahan and Weinberg, 2011) (Supplementary Fig. S5). Interestingly, in these experiments, the gene-level classification models performed better

**Table 1.** Performance comparison of models

| | BLCA | BRCA | COAD | PRAD | STAD |
|---|---|---|---|---|---|
| GCN+MAE (best) | **93.74** (9-Att) | **85.52** (14-Att) | **87.01** (11-Att) | **89.62** (9-Att) | **91.49** (7-Att) |
| GCN+MAE (#class-Att) | 93.48 | 85.22 | 86.25 | 88.52 | 90.8 |
| GCN+Single Att[a] | 91.08 | 85.03 | 84.97 | 86.55 | 90.96 |
| GCN best | 90.98 (hsa04151)[d] | 82.72 (hsa05206)[e] | 82.79 (hsa04151)[d] | 86.13 (hsa05200)[f] | 90.79 (hsa04151)[d] |
| SAS+SVM[b] | 81.51 | 74.41 | 77.54 | 79.25 | 76.08 |
| SAS+RF | 79.12 | 73.54 | 69.44 | 67.02 | 67.00 |
| SAS+MLP | 83.27 | 48.51 | 76.40 | 77.52 | 76.82 |
| RAW+SVM[c] | 89.18 | 82.62 | 78.41 | 82.58 | 86.39 |
| RAW+RF | 79.83 | 77.11 | 74.69 | 68.36 | 76.17 |

*Note*: Our proposed model (GCN+MAE) was compared with other models including the GCN pathway model. The best F1 values were represented as bold texts. The ensemble models using attention mechanism showed better performance than the other classifiers. 'GCN + MAE (best)' indicated how many attention mechanisms were used in parentheses. In the parentheses of 'GCN best', the ID of the pathway showing the performance was described.

[a]Instead of multi-attention, the GCN pathway models are combined with single attention mechanism.

[b]The pathway activity inference tool from Lim *et al.* (2016).

[c]A total of 20531 genes are used as input features.

[d]hsa04151: PI3K-Akt signaling pathway.

[e]hsa05206: MicroRNAs in cancer.

[f]hsa05200: Pathways in cancer.

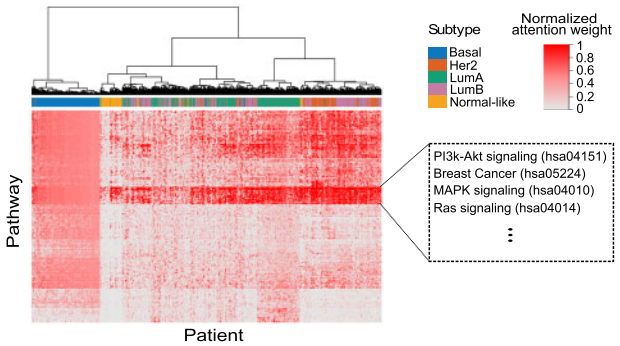



**Fig. 2.** Heatmap of the attention weight of GCN+MAE model on BRCA data. On the BRCA data, pathway attention weights of each patient were extracted from the best GCN+MAE model in Table 1. To better visualization, the whole attention weights were divided at 90th percentile and values of higher than 1 were forced to 1. Patients and pathways were then reorganized by similarity measure using the Manhattan distance and ward D.2 clustering

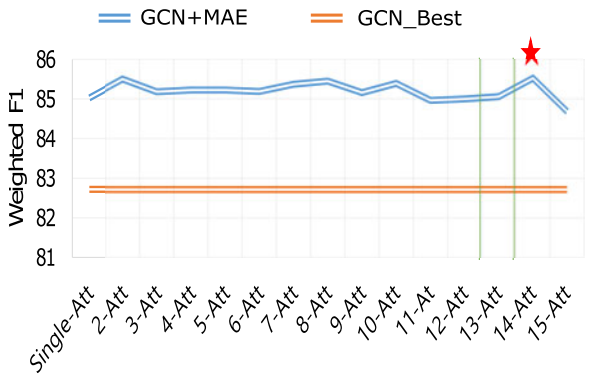



**Fig. 3.** Performance of GCN+MAE according to the number of attention mechanisms on BRCA. The *x*-axis of the figure represents the number of attentions from single attention to 15 attentions. The *y*-axis represents weighted *f*1 score. The blue line is for GCN+MAE and the orange line is for a single GCN pathway model's best result. Red stars indicate the best classification result point. All results in the figure were calculated in 10-fold CV tests. (Color version of this figure is available at *Bioinformatics* online.)

than the pathway activity inference tool-based models. These experimental results suggested that proper modeling and aggregation of pathway information is important for the pathway-based modeling of transcriptome data.

*Highlighted pathways in breast cancer*: Until now, the usefulness of GCN+MAE model was analyzed in terms of classification performances. We further investigated pathway attention weights across subtypes. Figure 2 and Supplementary Figure S6 show clustering results of the attention weights in BRCA and the other cancers, respectively. In addition, Supplementary Figures S7 and S8 show consistencies of the attention weights on each subtype from 10-fold CV results. Since the number of data is not sufficient, the number of samples of a particular subtype belonging to 1-fold is small. A small number of samples are difficult to reflect the characteristics of the subtypes. To address this issue, 4-fold results were averaged to calculate importance of pathways on consistency comparisons.

Pathway attention weights for each patient were determined by a weighted sum of pathway-level attention vectors, and these weights were extracted from the ensemble-level attention vector. For BRCA, our GCN+MAE model was able to highlight pathways that are known to be important in breast cancer. For example, the highlighted pathways were PI3k-Akt signaling (hsa04151) (Paplomata and O'Regan, 2014) and MAPK signaling (hsa04010) pathways (Santen *et al.*, 2002). Overall, patients were well clustered in the heatmap of attention weights in Figure 2. In particular, patients of Basal subtype formed a distinct cluster, which could explain the aggressiveness of Basal subtype breast cancer in terms of dysregulated or over-activated pathways.

#### 3.3.2 Effects of the number of multi-attention mechanism

Note that, we used multi-attention, thus, the number of attentions used to combine pathways would result in performance differences. Thus, we investigated how the classification performance varied concerning the number of attentions. As shown in Figure 3 and Supplementary Figure S9a–e, MAE models outperformed the single GCN pathway model for all cancer datasets, except STAD cancer data. Performance differences, when different numbers of attentions were used, were <1% in most cases.

Another experiment that we performed was clustering analysis of patients and, interestingly, we found that the optimal number of attentions was quite similar to the number of patient clusters. Since input to the GCN+MAE model was a combination of outputs of GCN pathway models, we concatenated outputs of GCN pathway models into a single vector to represent a patient. Then, we performed *X*-Means clustering on the concatenated vectors (Supplementary Fig. S9f). The number of clusters was quite similar to the number of attentions, which could be an explanation of why the GCN+MAE model achieved good performances in subtype classification. For further details, refer to the Supplementary Material.

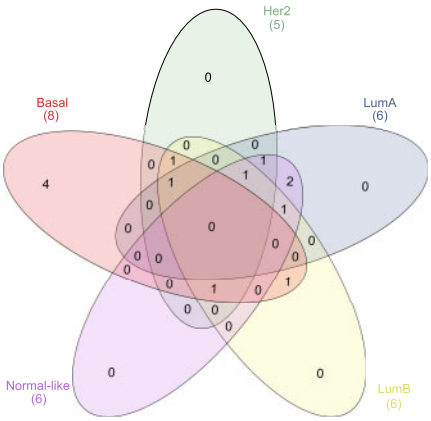
### 3.4 Identification of TFs as regulator of pathways and Gene Ontology term analysis of TF target genes

*Subtype-specific TFs*: Multi-attention mechanisms produce which pathways are highlighted while making predictions of cancer subtypes. Though some of top 25 highlighted pathways were relatively different among cancer subtypes, most of them were overlapped significantly (Supplementary Table S5). Thus, highlighted pathways were not successful in explaining differences in biological functions among cancer subtypes. For this reason, further investigation was performed to focus on differences of biological functions among subtypes as described in Section 2.3. By normalized rank from network propagation analysis with permutation-based normalization, TFs were ranked by the propagation scores. A majority of ranks were common in consistency comparisons on 10-fold CV experiments as in pathway attention weights tests in Section 3.3.1 (Supplementary Figs S10 and S11). While the TFs in the top 25 highlighted pathways were overlapped among subtypes, TFs that were selected by the network propagation were quite distinct among subtypes (Supplementary Table S6).

Top 25 TFs in each subtype of cancers were selected based on the network propagation (Supplementary Table S7). For example, in the case of BRCA, nuclear hormone receptor-related TFs, such as *NR1D1, NR1H4, RORA, RORB* and *RORC*, were ranked high in Basal subtype. On the other hand, tumor suppressor genes, such as *TP53* and *FOXO4*, ranked top in Luminal A subtype. Then, target genes of these top-ranked TFs were determined from a curated database (Han *et al.*, 2018), and then biological functions of these target genes were identified by Gene Ontology (GO) term analysis by Enrichr (Kuleshov *et al.*, 2016). Remember that about two-third genes, including many TFs, are not included in the KEGG pathway database. Our analysis scheme below is an effective way to investigate differences in biological functions of cancer subtypes from subtype-specific pathways: subtype-specific pathways → TFs → biological functions.

*Subtype-specific biological functions*: Enriched GO biological processes (GO_BPs) of target genes of TFs were analyzed, varying the number of top TFs from 5 to 25 in a step of 5. Then, consistently detected GO_BP terms regardless of the number of TFs were collected as subtype-specific biological functions (Fig. 4 for BRCA; Supplementary Fig. S12 for other cancers). GO_BPs enriched in each subtype were represented as Venn diagram using InteractiVenn (Heberle *et al.*, 2015). As shown in Figure 4, GO_BP terms enriched in Basal subtype are most distinct compared to GO_BP terms enriched in other subtypes, which could be a good explanation of why the Basal subtype is most aggressive. On the other hand, LumA and normal-like subtypes shared almost the same GO_BP terms including positive regulation of nucleic acid-template transcription (GO: 1903508) and positive regulation of gene expression (GO: 0010628), which also explains better prognosis of LumA compared to other subtypes. In addition, three subtypes, such as Her2, LumB and Basal, shared cellular response to cytokine stimulus (GO: 0071345) that was known as important factors of breast cancer development and metastasis (Eichbaum *et al.*, 2011; Esquivel-Velázquez *et al.*, 2015).

GO_BP terms that were enriched only in Basal subtype could explain its aggressive phenotype. Positive regulation of cell proliferation (GO: 0008284) and negative regulation of apoptotic process (GO: 0043066) were closely related to cancer cell proliferation and they were known as one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Basal subtype cancer had a high rate of proliferation than other subtypes. This dysregulated cell proliferation resulted in worse prognosis and treatment options are difficult to choose for patients (Cakir *et al.*, 2012; Castelvecchi, 2016). Two more enriched GO_BP terms were inflammatory response (GO: 0006954) and positive regulation of cytokine production (GO: 0001819). Cytokines were a family of proteins related to immune systems. Inflammation and escape of immune destruction were also




| GO Biological Process | Basal (8) | Her2 (5) | LumA (6) | LumB (6) | Normal-like (6) |
|---|---|---|---|---|---|
| positive regulation of cell proliferation (GO:0008284) | ✓ | | | | |
| negative regulation of apoptotic process (GO:0043066) | ✓ | | | | |
| inflammatory response (GO:0006954) | ✓ | | | | |
| positive regulation of cytokine production (GO:0001819) | ✓ | | | | |
| cytokine-mediated signaling pathway (GO:0019221) | ✓ | | | ✓ | |
| cellular response to cytokine stimulus (GO:0071345) | ✓ | ✓ | | ✓ | |
| regulation of cell proliferation (GO:0042127) | ✓ | ✓ | | ✓ | ✓ |
| regulation of apoptotic process (GO:0042981) | ✓ | ✓ | ✓ | ✓ | |
| regulation of transcription from RNA polymerase II promoter (GO:0006357) | | ✓ | ✓ | | ✓ |
| positive regulation of transcription, DNA-templated (GO:0045893) | | ✓ | ✓ | ✓ | ✓ |
| positive regulation of nucleic acid-templated transcription (GO:1903508) | | | ✓ | | ✓ |
| positive regulation of gene expression (GO:0010628) | | | ✓ | | ✓ |
| negative regulation of transcription, DNA-templated (GO:0045892) | | | ✓ | ✓ | ✓ |

**Fig. 4.** GO_BPs enriched in each subtype of BRCA. To determine biological functions of top-ranked TFs by network propagation with permutation-based normalization, GO enrichment tests were performed using target genes of a different number of those TFs from 5 to 25. Among those tests, consistently detected GO_BPs were illustrated as a Venn diagram and listed as a table. The number of detected GO_BPs was denoted in the parentheses. Similar to the clinical prognosis, almost the same GO_BPs were observed in LumA and normal-like subtypes, and the other three subtypes showed also similar results. Unlike other subtypes, Basal subtype contained unique four GO_BP terms that were related to aggressiveness and metastasis of cancer

members of hallmarks of cancer (Hanahan and Weinberg, 2011). Many studies reported that Basal subtype breast cancer exhibited differential expression of inflammation related genes and stronger immunogenicity than the other breast cancer subtypes (Hartman *et al.*, 2013; Liu *et al.*, 2018). For these reasons, inflammatory-related cytokines were considered to be immunotherapeutic targets of Basal or triple-negative breast cancer (Fabre *et al.*, 2018; Kim *et al.*, 2015).

## 4 Conclusion

For cancer subtype prediction using pathways, we proposed an explainable ensemble of deep pathway models. Using GCN and multi-attention, our model captured localized gene expression patterns and aggregated information spread out various pathways. On the TCGA five cancer data, we showed that the proposed method outperformed the existing pathway activity inference methods and single GCN models. In addition, unlike other methods, the proposed model used the multi-attention to obtain a list of pathways that can effectively classify and explain the characteristics of cancer subtypes from the deep-learning model. Biological functions of these pathways were identified by connecting pathways and TFs by network propagation algorithm. Although our proposed model showed good performance for the interpretation of deep-learning models on cancer subtype classification tasks, there are still some limitations that can be further improved. For example, GCN could extract meaningful information from the pathway, but it was difficult to know what part of the pathway was useful for characterizing the cancers. Recently, the method called class activation mapping was developed (Zhou *et al.*, 2016). It highlighted areas of the image that are most informative and relevant to the prediction of class. If this method is used for GCN models on pathway analysis, it may be able to extract which genes are useful in modeling characteristics of cancers.

## Funding

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT [NRF-2017M3C4A7065887]; the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) [NRF2014M3C9A3063541]; and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea [HI15C3224].

*Conflict of Interest*: none declared.

## References

Alcaraz,N. *et al.* (2017) De novo pathway-based biomarker identification. *Nucleic Acids Res.*, **45**, e151.

Atwood,J. and Towsley,D. (2016) Diffusion-convolutional neural networks. In: *Advances in Neural Information Processing Systems, Barcelona, Spain*. pp. 1993–2001.

Bruna,J. *et al.* (2014) Spectral networks and locally connected networks on graphs. In: ICLR 2014. *Banff, Canada*.

Cakir,A. *et al.* (2012) A comprehensive morphological study for basal-like breast carcinomas with comparison to nonbasal-like carcinomas. *Diagn. Pathol.*, **7**, 145.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

Castelvecchi,D. (2016) Can we open the black box of AI? *Nature*, **538**, 20–23.

Choi,E. *et al.* (2016) Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: *Advances in Neural Information Processing Systems, Barcelona, Spain*. pp. 3504–3512.

Cowen,L. *et al.* (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.

Defferrard,M. *et al.* (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: *Advances in Neural Information Processing Systems, Barcelona, Spain*. pp. 3844–3852.

Dhillon,I.S. *et al.* (2007) Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**, 1944–1957.

Eichbaum,C. *et al.* (2011) Breast cancer cell-derived cytokines, macrophages and cell adhesion: implications for metastasis. *Anticancer Res.*, **31**, 3219–3227.

Esquivel-Velázquez,M. *et al.* (2015) The role of cytokines in breast cancer development and progression. *J. Interferon Cytokine Res.*, **35**, 1–16.

Fabre,J. *et al.* (2018) The interleukin-17 family of cytokines in breast cancer. *Int. J. Mol. Sci.*, **19**, 3880.

Gatza,M.L. *et al.* (2010) A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. USA*, **107**, 6994–6999.

Glaab,E. *et al.* (2010) Learning pathway-based decision rules to classify microarray cancer samples. In: *German Conference on Bioinformatics 2010, Braunschweig, Germany*. Gesellschaft für Informatik eV.

Grimm,L.J. *et al.* (2015) Can breast cancer molecular subtype help to select patients for preoperative MR imaging? *Radiology*, **274**, 352–358.

Gunning,D. (2017) Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*. nd web, 2.

Hammond,D.K. *et al.* (2011) Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.*, **30**, 129–150.

Han,H. *et al.* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, **46**, D380–D386.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *cell*, **144**, 646–674.

Hartman,Z.C. *et al.* (2013) Growth of triple-negative breast cancer cells relies upon coordinate autocrine expression of the proinflammatory cytokines IL-6 and IL-8. *Cancer Res.*, **73**, 3470–3480.

Heberle,H. *et al*. (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics*, **16**, 169.
Hofree,M. *et al*. (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.
Hwang,K.-T. *et al*. (2019) Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: a population-based study using seer database. *Clin. Cancer Res*., **25**, 1970–1979.
Jo,K. *et al*. (2016) Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. *Bioinformatics*, **32**, i128–i136.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
Kim,G. *et al*. (2015) SOCS3-mediated regulation of inflammatory cytokines in PTEN and p53 inactivated triple negative breast cancer model. *Oncogene*, **34**, 671–680.
Kim,S. *et al*. (2012) Pathway-based classification of cancer subtypes. *Biol. Direct*, **7**, 21.
Kipf,T.N. and Welling,M. (2017) Semi-supervised classification with graph convolutional networks. In: *ICLR 2017, Toulon, France*.
Köhler,S. *et al*. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet*., **82**, 949–958.
Kong,Y. and Yu,T. (2018) A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics*, **34**, 3727–3737.
Kuleshov,M.V. *et al*. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*., **44**, W90–W97.
Kunz,M. *et al*. (2019) In silico signaling modeling to understand cancer pathways and treatment responses. *Brief. Bioinform*.
Lambert,S.A. *et al*. (2018) The human transcription factors. *Cell*, **172**, 650–665.
Leiserson,M.D. *et al*. (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet*., **47**, 106–114.
Lim,S. *et al*. (2016) Protein interaction network (PIN)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods*, **110**, 81–89.
Lim,S. *et al*. (2020) Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Brief. Bioinform*., **21**, 36–46.
Liu,Z. *et al*. (2018) A comprehensive immunologic portrait of triple-negative breast cancer. *Transl. Oncol*., **11**, 311–329.
Monti,F. *et al*. (2017) Geometric deep learning on graphs and manifolds using mixture model CNNs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii*. pp. 5115–5124.
Moon,J.H. *et al*. (2017) PINTnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI. *BMC Syst. Biol*., **11**, 15.
Paplomata,E. and O'Regan,R. (2014) The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Ther. Adv. Med. Oncol*., **6**, 154–166.
Parker,J.S. *et al*. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*., **27**, 1160–1167.
Pearson,K. (1905) The problem of the random walk. *Nature*, **72**, 342.
Rhee,S. *et al*. (2018) Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden*. pp. 3527–3534. AAAI Press.
Santen,R.J. *et al*. (2002) The role of mitogen-activated protein (MAP) kinase in breast cancer. *J. Steroid Biochem. Mol. Biol*., **80**, 239–256.
Schadt,E.E. *et al*. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet*., **37**, 710–717.
Segura-Lepe,M.P. *et al*. (2019) Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC Bioinformatics*, **20**, 543.
Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535–D539.
Su,J. *et al*. (2009) Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*, **4**, e8161.
Vaswani,A. *et al*. (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems, California, USA*. pp. 5998–6008.
Viswanathan,G.A. *et al*. (2008) Getting started in biological pathway construction and analysis. *PLoS Comput. Biol*., **4**, e16.
Zhang,J.D. and Wiemann,S. (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, **25**, 1470–1471.
Zhang,W. *et al*. (2018) Classifying tumors by supervised network propagation. *Bioinformatics*, **34**, i484–i493.
Zheng,W. *et al*. (2017) An attention-based effective neural model for drug-drug interactions extraction. *BMC Bioinformatics*, **18**, 445.
Zhou,B. *et al*. (2016) Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada*. pp. 2921–2929.
Zitnik,M. *et al*. (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**, i457–i466.