A Drug-induced Liver Injury Prediction Model using Transcriptional Response Data with Graph Neural Network

Doyeong Hwang*‡, Minji Jeon*‡, Jaewoo Kang*†§
*Department of Computer Science and Engineering
†Interdisciplinary Graduate Program in Bioinformatics
Korea University
Seoul, Republic of Korea
Emails: {desertbeagle, mjjeon, kangj}@korea.ac.kr
‡Both authors contributed equally
§Corresponding author

***Abstract*—Drug-Induced Liver Injury (DILI) is a major cause of failed drug candidates in clinical trials and withdrawal of approved drugs from the market. Therefore, machine learning-based DILI prediction can be key in increasing the success rate of drug discovery because drug candidates that are predicted to potentially induce liver injury can be rejected before clinical trials. However, existing DILI prediction models mainly focus on the chemical structures of drugs. Since we cannot determine whether a drug will cause liver injury based solely on its structure, DILI prediction based on the transcriptional effect of a drug on a cell is necessary.**

In this paper, we propose GLIT which is a model that uses transcriptional response data and chemical structures and can be used for drug-induced liver injury prediction. GLIT learns the embedding vectors of drug structures and drug-induced gene expression profiles using graph attention networks in a biological knowledge graph for predicting DILI. GLIT outperformed a baseline model that uses only drug structure information by 7% and 19.2% in terms of correct classification rate (CCR) and Matthews correlation coefficient (MCC), respectively. In addition, we conducted a literature survey to confirm whether the class labels of drugs, in the unknown DILI class, predicted by GLIT are correct.

***Index Terms*—Drug-induced Liver Injury, Drug-induced Gene Expression Profiles, Drug Discovery, Graph Neural Network**

I. Introduction

It takes more than 10 years and 2.6 billion dollars on average to develop a novel drug but the success rate of drug development is less than 10% [1]. Drug-Induced Liver Injury (DILI) is a major cause of clinical trial failures or drug withdrawal from the market [2]. In particular, eight drugs that cause liver injury have been withdrawn from the market over the past 20 years [3]. If DILI can be predicted early, the time and cost of developing a novel drug can be reduced, and the chances of success in clinical trials can be improved. Thus, it is important to predict whether a drug candidate is likely to cause liver injury before clinical trials.

Identify applicable funding agency here. If none, delete this.

Several machine learning models trained on datasets such as Liver Toxicity Knowledge Base (LTKB) [4] and Open TG-GATEs [5] were previously proposed for DILI prediction. Most of the previously proposed machine learning models are trained on drug structure information for predicting DILI [6]–[9]. However, such models used for DILI prediction do not consider genetic information or the structures and complex biological mechanisms of drugs [10]. Therefore, these models cannot predict whether a drug will cause liver injury based solely on its structure.

Recently, the phenotype-based features of drugs which are helpful in predicting DILI have been identified. Several studies have shown that the level of drug-induced gene expression, which is one of the phenotype-based features, is helpful in predicting DILI [10]–[12]. The previous researches selected toxicity-related genes and showed that the drug-induced gene expression levels of the selected genes are useful in predicting drug toxicity. Moreover, recently released large-sized drug-induced gene expression profile datasets can be used for training models and thus help improve their DILI prediction performance. For example, the Broad Institute's Connectivity Map (CMap) [13] contains 1.6 million drug-induced gene expression profiles of more than 70 cell lines and 20,000 chemical compounds.

As deep neural networks have significantly improved performance in natural language process tasks and image recognition tasks, they have also been applied to drug discovery tasks such as predicting the binding affinity between drugs and proteins [14], generating a novel drug structure [15], [16], and predicting the properties of a compound [17]. Graph neural network is one of the neural network architectures which directly operates on a graph. Graph neural network enables models to utilize relational data. In the drug discovery field, the structure of a drug can be represented as a graph of atoms and their bonds. Graph neural network models that represent drug structures as graphs for embedding drug structures or predicting side effects have been proposed [18]–[20]. Moreover, previously proposed

2375-9356/20/$31.00 ©2020 IEEE
DOI 10.1109/BigComp48618.2020.00-54

graph neural networks in biological knowledge graphs such as protein-protein interaction network graphs or biological pathway graphs help to improve side effect prediction [21]. However, to the best of our knowledge, graph neural networks have not been applied for predicting DILI.

In this paper, we propose GLIT which is a deep neural network model that employs graph neural network for predicting liver injury using transcriptional response data. GLIT predicts drug-induced liver injury using the structure information and transcriptional response profiles of drugs. GLIT utilizes a biological knowledge graph and pre-trained gene embedding vectors. We evaluated the prediction performance of GLIT on internal test sets, predicted the probability of liver injury indications caused by unlabeled drugs in an external test set, and conducted a literature survey to validate drugs with a high probability of causing liver injury. The main contributions of this work are as follows:

- GLIT employs drug transcriptional response and drug structure information for DILI prediction.
- A graph neural network with pre-trained gene embedding vectors is employed for using biological knowledge.
- Due to its deep neural network architecture, GLIT can generate embedding vectors of drugs and drug-induced gene expression profiles. The embedding vectors can be used for downstream analysis tasks such as drug-induced lung injury prediction tasks or personalized side effect prediction tasks.
- Our code and trained GLIT model are publicly available at: https://github.com/dmis-lab/GLIT.

II. METHODS

In this section, we describe our proposed model GLIT. Figure 1 illustrates the overall architecture of GLIT and Figure 2 illustrates the graph neural network layers of GLIT in detail. As shown in the figures, the structure information of a drug, the drug-induced gene expression level of a drug on a cell line, and the dosage and duration of drug administration are used as inputs for GLIT. A drug embedding vector is extracted from the last layer of a neural network with three layers in GLIT, and a drug-induced gene expression profile embedding vector is extracted from a graph neural network with three layers. The graph neural network uses a biological knowledge graph called OmniPath. The two vectors are concatenated with the dosage and duration of drug administration information, and fed to the prediction layers of GLIT to predict DILI.

Dataset description

GLIT was trained on the CMap dataset and the LTKB dataset. The CMap dataset contains 1.6 million drug-induced gene expression profiles of 77 cell lines and 22,209 chemical compounds. The profiles are measured using the L1000 assay which is used to measure the expression levels of 978 landmark genes. The expression levels of 11,350 additional genes are predicted using an inference algorithm. The CMap dataset obtained from Gene Expression Omnibus (GSE92742, GSE70138) contains data generated from each data preprocessing level and we use level 5 data among them. In the level 5 data, a raw drug-induced gene expression profile is normalized based on the other drug-induced gene expression profiles located in the same L1000 plate.

The LTKB dataset contains drugs that are annotated with the probability of causing liver injury. The dataset contains 1,036 FDA-approved drugs which are categorized into the following four classes: most-, less-, no-, and ambiguous-DILI concern. Note that drugs in the most-DILI concern class are known to cause liver injury and drugs in the less-DILI concern class can cause liver injury but rarely induce severe injury. To help our model capture strong data signals, we utilize drugs in the most-DILI concern class or in the no-DILI concern class for model training and evaluation. 197 drugs whose drug-induced gene expression profiles are available in the CMap dataset are selected among drugs in the most- class and drugs in the no-DILI concern class from the LTKB dataset.

We used a total of 12,348 samples. The structure information of a drug, the drug-induced gene expression profile of the drug on a cell line, dosage, and the duration of drug administration are used as input for GLIT. The structure information of a drug is represented using Extended Connectivity Fingerprint (ECFP), which is a method of representing the presence or absence of substructures of a drug. The drug-induced gene expression profile is represented using a knowledge graph.

We used OmniPath [22] which is a comprehensive collection of biological signaling pathways [22]. We used OmniPath as a biological knowledge graph because it contains a sufficient number of relations. OmniPath contains more relations between landmark genes than other biological knowledge graphs from databases such as KEGG [23] pathways and BioGRID [24].

To initialize the node representations of genes in OmniPath, we employed pre-trained gene embedding vectors generated using Gene2vec [25]. The embedding vectors generated using Gene2vec are distributed representations of genes. Gene2vec uses gene co-expression patterns in 984 data sets from the GEO [26] databases. Unlike random embedding vectors, the embedding vectors contain genomic information that can help train GLIT more efficiently. For this reason, the embedding vectors were employed for initializing representations of genes in GLIT.

We used 964 common genes in the entire OmniPath graph, the gene lists of Gene2vec, and L1000 landmark genes. A subnetwork of the 964 genes is utilized.

Problem statement

The notations $\mathcal{D} = \{D_1, D_2, \cdots, D_{197}\}$ and $\mathcal{E} = \{E_1, E_2, \cdots, E_{12348}\}$ denote a set of drugs in the combined dataset of LTKB and CMap datasets and a set of drug-induced gene expression profiles, respectively. Each drug $D \in \mathbf{R}^{2048}$ represents a 2048-bit ECFP generated by RDKit[1].

[1] http://www.rdkit.org/

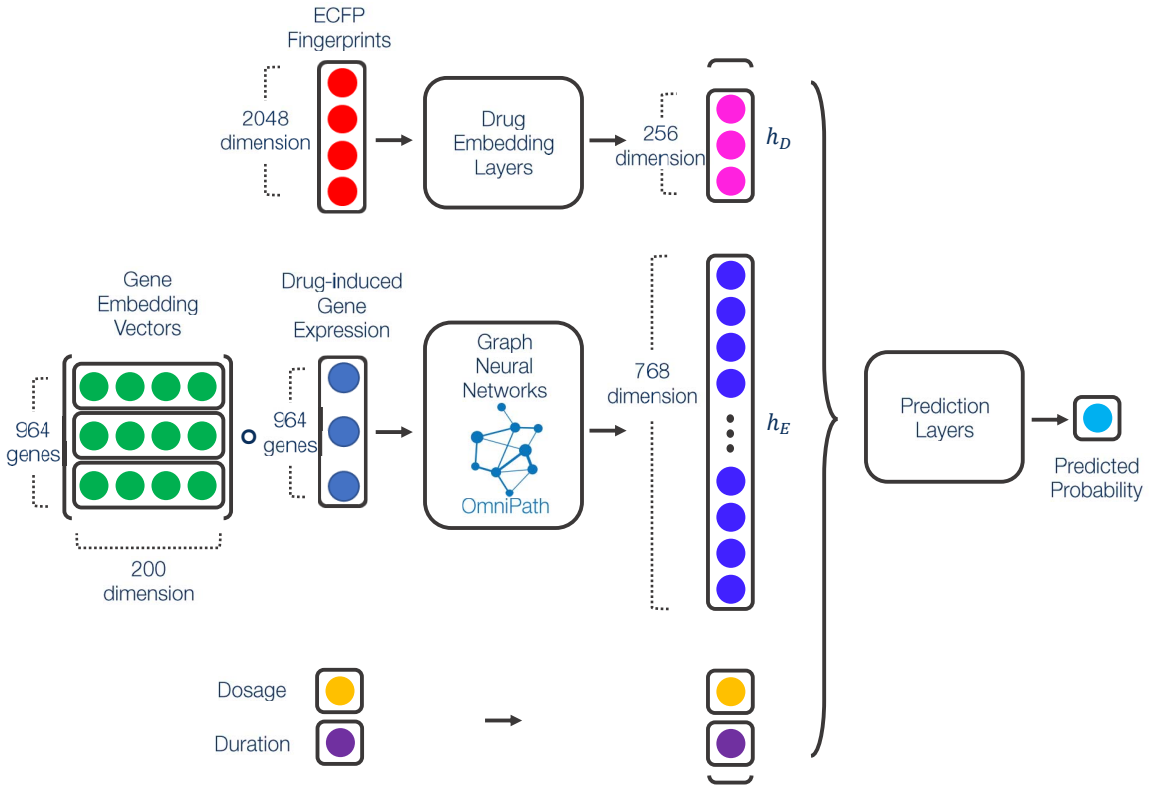

Fig. 1. The overall architecture of GLIT. The structure of a drug, the drug-induced gene expression level, and the dosage and duration of drug administration are used as inputs for GLIT. A drug embedding vector is extracted from the last layer of a neural network with three layers, and a drug-induced gene expression embedding vector is extracted from a graph neural network using a biological knowledge graph called OmniPath. The two embedding vectors are concatenated with the dosage and duration of drug administration information, and fed to the prediction layers of GLIT to predict DILI.

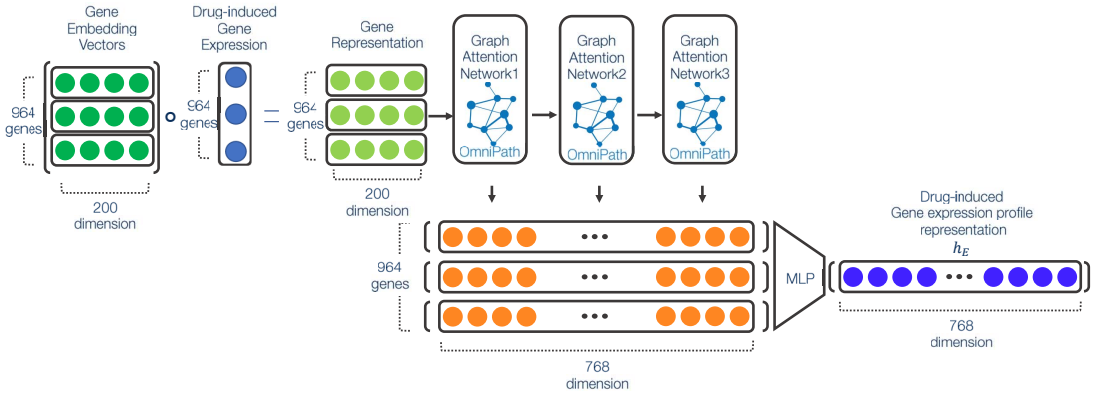

Fig. 2. Graph neural network architecture of GLIT. Pre-trained gene embedding vectors multiplied by drug-induced gene expression levels are initial node representations of genes in OmniPath. The three layers of a graph attention network (GAT) learn hidden representations of nodes using the given initialized node representations, and the learned hidden representations are integrated into a 768-dimensional drug-induced gene expression profile representation.

Each profile $E \in \mathbf{R}^{964}$ consists of drug-induced gene expression levels of 964 genes. GLIT predicts as follows:

$$\hat{Y} = f(D, E, B, T|\theta), \tag{1}$$

where $\hat{Y}$, B, T, and θ denote a binary DILI class label, dosage, duration, and the parameters of the model, respectively.

Drug embedding layers

Drug structure information is represented as a 2048-bit vector D which is fed to a multi-layer perceptron (MLP) to obtain a hidden representation. Three layers of MLP are stacked to produce a 256-dimensional hidden vector:

$$h_D = \mathbf{W_L}(\sigma(\mathbf{W_{L-1}}(...(\sigma(\mathbf{W_1}D + b_1))...) + b_{L-1})) + b_L \tag{2}$$

where $\mathbf{W_i}$, $\mathbf{b_i}$, σ, and L are the weight matrix, the bias term of the i-th layer, the ReLU activation function, the number of layers, respectively.

Graph neural network layers for drug-induced gene expression profile representation

For embedding drug-induced gene expression profiles, we use graph attention network (GAT) [27] which is one of the most effective graph neural network architectures that learns hidden representations of each node using the attention scores of a given node and its neighbors in a graph. Note that in the GAT layers of GLIT, nodes are genes and the neighboring nodes are the genes that have relationships with the given gene in OmniPath.

As mentioned above, we employ pre-trained gene embedding vectors rather than randomly initialized vectors as input node representations to help GLIT easily capture genomic information. Gene embedding vectors generated by Gene2vec [25]. Then the drug-induced gene expression level of a gene, is multiplied with its pre-trained gene embedding vector from Gene2vec. The gene embedding vectors generated by this process contain genomic and drug-induced gene expression level information. If a gene is up-regulated by a given drug, the length of the embedding vector of the gene will be longer based on the gene's differential expression level, increasing the influence of the gene information represented as the embedding vectors next to the neighboring nodes. These gene embedding vectors are the node representations which are inputted to GAT layers. Note that gene embedding vectors are optimized while GLIT is trained. Each gene embedding vector is expected to obtain useful information for a given task.

To be specific, a set of node feature vectors $\mathbf{h} = \{h_1, h_2, ..., h_N\}, h_i \in \mathbf{R}^F$ is used as input to a GAT layer. N is the number of nodes in a graph; in this study, N is 964. F is the dimension of node vectors. In this task, F is set to 200 for the first layer, and to 256 for the other layers. Each node vector in $\mathbf{h}$ is projected through the shared weight matrix parameter $\overline{\mathbf{W}} \in \mathbf{R}^{\mathbf{F}' \times \mathbf{F}}$, where F' is the output dimension of a GAT layer and is fixed to 64. For each projected node vector, the attention coefficient of a given node h_i and one of its neighboring nodes $h_j, j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of nodes neighboring node i is calculated. The calculated attention coefficients of $\mathcal{N}_i$ are normalized using the softmax function. In this study, multi-head attention [28] is used to calculate k number of different attention coefficients α^k as follows:

$$\alpha_{ij}^k = \frac{exp(LeakyReLU(\mathbf{a}^{kT}\left[\overline{\mathbf{W}}^k h_i \| \overline{\mathbf{W}}^k h_j\right]))}{\sum_{m \in \mathcal{N}_i} exp(LeakyReLU(\mathbf{a}^{kT}\left[\overline{\mathbf{W}}^k h_i \| \overline{\mathbf{W}}^k h_m\right]))} \quad (3)$$

where $\|$ denotes concatenation. A representation of a node is obtained from its neighbors' node representations using multi-head attention coefficients as follows:

$$h_i{}' = \|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \overline{\mathbf{W}}^k h_j\right) \quad (4)$$

where $h_i{}' \in \mathbf{R}^{KF'}$ is a hidden representation of a node, K is the number of heads in multi-head attention and σ is an activation function. In this study, K is 4 and σ is ReLU. For each GAT layer, the node representations are extracted and concatenated to make the final output matrix of size $\mathbf{R}^{MKF' \times 964}$, where M is the number of GAT layers. The final output matrix of GAT layers is then fed to 3 layers of MLP, and the MLP layers return a drug-induced gene expression profile embedding vector $h_E \in \mathbf{R}^{MKF'}$.

Prediction layer

Finally, a drug embedding vector h_D, a drug-induced gene expression profile embedding vector h_E, dosage B, and duration T are concatenated into a single vector. The vector is then fed to three layers of MLP to predict binary classification probabilities as follows:

$$\begin{aligned} y' = Softmax(\mathbf{W'_{L'}}(\sigma(\mathbf{W'_{L'-1}}(...(\sigma(\mathbf{W'_1}[h_D \| h_E \| B \| T] \\ + b'_1))...) + b'_{L'-1})) + b'_{L'}) \end{aligned} \quad (5)$$

where $\mathbf{W'_i}$, $\mathbf{b'_i}$, σ, L' are the weight matrix, bias term of the i-th layer, ReLU activation function, and the number of layers, respectively.

III. Results

Dataset preparation

All the drugs in the combined dataset of LTKB and CMap datasets were randomly divided into the training set (70%), validation set (10%), and test set (20%). The drug-induced expression profiles of the drugs in each set are used as input. If the training set contains 151 drugs and each drug has 10 drug-induced gene expression profiles, the total number of samples in the training set will be 1510. The entire dataset is randomly divided into training, validation, and test sets five times to prevent overfitting. The average performance of our model on the five test sets is reported.

Optimization

To optimize GLIT, cross-entropy loss is calculated based on the predicted probabilities. Cross-entropy loss is defined as follows:

$$CrossEntropyLoss = -\sum_i (y_i log(y'_i) + (1 - y_i) log(1 - y'_i)) \quad (6)$$

where y' is the predicted value, and y is the true value.

Moreover, ReLU activation and dropout were used for all the MLP layers except for the final prediction layer. The dropout rate p was set to 0.5. Batch normalization was applied to the following input: drug structure information, drug-induced gene expression profile, dosage, and the duration of drug administration. Hyper-parameters were searched to maximize performance on validation sets. The hyper-parameters of GLIT such as the number of epochs, learning rate, and weight decay were determined by a grid search.. For GLIT, 20 epochs, a learning rate of 5e-04, and a weight decay of 1e-05 were used. Range of hyperparamters searched are provided as table I.

TABLE I
RANGES OF HYPERPARAMETERS SEARCHED FOR OPTIMIZATION OF PERFORMANCES

Hyperparameters	Range
Number of layers for GNN(M)	{ 2, 3, 4 }
Hidden dimension(F), Number of heads(K) in GAT	{ (16, 4), (32, 4), (64, 4), (128, 2), (256, 1) }
Number of layers in Drug embedding layers(L)	{ 2, 3 }
Number of layers in Prediction layers(L')	{ 2, 3 }
Learning rate	{ 5e-4, 1e-4, 5e-5, 1e-5 }
Weight decay	{ 1e-5, 1e-7, 1e-9 }
Number of epochs	{ 10, 20, 30 }

Baseline models

We assessed the performance of GLIT through ablation tests and comparison with baseline models. We tested the following models: an MLP model that is trained on profiles by using only drug-induced gene expression levels as input, an MLP model that is trained on drugs by using only drug fingerprints as input, and an MLP model that is trained on profiles and drugs by using both drug-induced gene expression levels and drug structure information as input. Moreover, to show that Gene2vec improves performance, we compared the performance of the MLP model trained on profiles and drugs with the MLP model trained on profiles using Gene2vec and drugs. The MLP model trained on profiles and drugs using Gene2vec receives gene embedding vectors multiplied by drug-induced gene expression levels. The embedding vectors of genes are fed to MLP layers to obtain a drug-induced gene expression profile embedding vector. The MLP model trained on profiles and drugs using Gene2vec and GLIT were compared to evaluate the effectiveness of GATs.

In addition, we used the following three baseline models and measured their performance. The random forest classifier of Kim et al. [8] was trained on weighted ECFP for predicting DILI class labels. Wang et al. [11] employed an MLP classifier that uses the drug-induced gene expressions of genes in toxicity-relevant gene sets for DILI class label prediction. Feng et al. [10] selected 1574 toxicity-related genes, trained an MLP classifier on the selected genes, and conducted a gene ontology analysis which showed that the selected genes are related to DILI. However, the expression levels of some of the selected genes in the studies of Wang et al. and Feng et al. were not measured or inferred from the L1000 assay. 963 genes in the published gene set by Wang et al. and 1015 genes in the set by Feng et al., which are commonly included in the L1000 assay gene list, are used. The performance of two MLP classifiers trained on the two common gene lists is measured. The hyperparameters of the MLP classifiers were searched by a grid search.

We also compared the performance of GLIT using graph convolutional network (GCN) [29], GraphSAGE [30], and GAT. GCN is a graph neural network which utilizes spectral graph convolution. The node vectors of graphs are transformed into a spectral domain through Fourier transform. A convolution operation is then applied to the node vectors. GraphSAGE is another graph neural network model that represents a node vector by aggregating the node representations of a subset of its neighbors. The nodes in the subset are randomly sampled from uniform distribution.

TABLE II
EVALUATION RESULTS OF THE BASELINE MODELS AND GLIT. THE AVERAGE PERFORMANCE AND STANDARD DEVIATION ON THE FIVE TEST SETS ARE REPORTED.

Model \ Metric	CCR	MCC	F1	Accuracy
Kim et al. [8]	0.7204 (0.0577)	0.4637 (0.1647)	0.7662 (0.0651)	0.7264 (0.077)
Wang et al. [11]	0.5162 (0.0498)	0.0466 (0.2067)	0.7115 (0.0311)	0.569 (0.0478)
Feng et al. [10]	0.5075 (0.0599)	0.0297 (0.2255)	0.7121 (0.0316)	0.5642 (0.0485)
MLP trained on profiles	0.5436 (0.0534)	0.1068 (0.1352)	0.6514 (0.0487)	0.5548 (0.0535)
MLP trained on drugs	0.6951 (0.0358)	0.3926 (0.0716)	0.7057 (0.0365)	0.6954 (0.0352)
MLP trained on profiles and drugs	0.7248 (0.0451)	0.467 (0.08)	0.7532 (0.0226)	0.7281 (0.0412)
MLP trained on profiles and drugs using Gene2vec	0.7451 (0.0555)	0.4986 (0.1055)	0.7709 (0.0332)	0.7479 (0.0522)
Graph convolutional network	0.7472 (0.052)	0.5052 (0.1045)	0.7711 (0.0431)	0.7498 (0.0545)
GraphSAGE	0.747 (0.0615)	0.5029 (0.1233)	0.7703 (0.0556)	0.7498 (0.0611)
GLIT	**0.7711** (0.0450)	**0.5525** (0.1032)	**0.7883** (0.0501)	**0.7732** (0.0470)

Evaluation results

The final prediction label of a drug is the averaged predicted probability of the samples in the test set regarding a drug. Model evaluation is based on the final predictions of drugs. Corrected classification rate (CCR), Matthews correlation coefficient (MCC), Accuracy, and F1 score are used as the evaluation metrics:

$$CCR = 0.5 * (\frac{T_N}{T_N + F_P} + \frac{T_P}{T_P + F_N}) \tag{7}$$

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P+F_P)(T_P+F_N)(T_N+F_P)(T_N+F_N)}} \tag{8}$$

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{9}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{10}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{11}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

where T_P, T_N, F_P, F_N are the number of true positive, true negative, false positive, and false negative samples, respectively.

Table II shows the performance of the three baseline models and the ablation test results of GLIT. The averages and the standard deviations of performance on the five test sets are reported. As shown in the table, GLIT achieves the highest scores in terms of all the evaluation metrics among the baseline models and the ablations. Among the three baseline models,

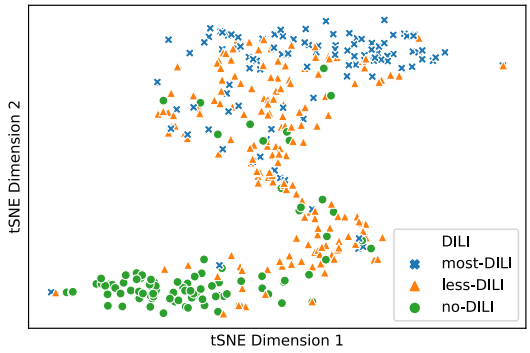

Fig. 3. Visualization of drug embedding vectors with DILI class labels

the model by Kim et al. obtained the best performance. However, GLIT outperforms the model of Kim et al. by 7% and 19.2% in terms of CCR and MCC, respectively. The results of the MLP model trained on profiles, the MLP model trained on drugs, and the MLP model trained on profiles and drugs show that the MLP model trained on profiles and drugs outperforms by 33.3% and 4.3% in terms of CCR and MCC, respectively. Moreover, using pre-trained gene embedding vectors from Gene2vec improve the CCR scores of MLP models by 2.8%, and graph neural networks using a biological knowledge graph improve the CCR scores by 3.5%. The performance of models with GCN and GraphSAGE are similar to that of the MLP model trained on drug data and profiles generated using pre-trained gene embedding vectors from Gene2vec. The results of models using GCN and GraphSAGE show that GAT helps improve performance. We presumed that multi-head attention helps GLIT learn the relations between genes. As these results demonstrate, using drug structure information, drug-induced gene expression profiles, pre-trained gene embedding vectors, and GAT with a biological knowledge graph is effective in predicting DILI.

Qualitative analysis: drug embedding vector analysis

We visualized embedding vectors of drugs and observed whether the drugs with the same DILI class labels are located closely to each other. The embedding vectors h_D of drugs in the most-, less-, and no-DILI concern classes were extracted from our trained GLIT model. The embedding vectors are visualized using t-SNE (Figure 3). The visualization shows a cluster of drugs with most-DILI concern class labels and a cluster of drugs with no-DILI concern class labels. Although the data of drugs in the less-DILI concern class was not used for training GLIT, which was trained on either the data of drugs in the most-DILI concern class or the data of drugs in the no-DILI concern class, drugs in the less-DILI class are located between drugs in the most- class and drugs in the no-DILI concern class. Thus, the visualization shows that drug embedding vectors extracted from GLIT contain DILI information.

Qualitative analysis: evaluation on the external dataset

To confirm whether drugs selected by GLIT cause liver injury and validate the effectiveness of GLIT as a DILI prediction model in the practical drug development process, we conducted a literature survey. To do this, we used 2,023 drugs that have drug-induced gene expression profiles in the CMap dataset but have no DILI class labels. GLIT predicted the probability of each drug inducing liver injury. For some of the drugs that are the most likely to cause liver injury, we could find several studies which reported that the drugs may induce liver injury. For example, Amodiaquine is an antimalarial drug that has a 0.99 probability of causing liver injury, which was predicted by GLIT. In general, it is not recommended for malaria patients due to its side effects, including severe liver injury [31], [32]. The antibiotic Flucloxacillin is another drug that has a 0.978 probability of causing liver injury, which was calculated by GLIT. It is reported to have severe hepatic adverse effects [33]. The results show that GLIT is an effective DILI prediction model that can be used for drug discovery.

IV. Discussion

Drug structure information may be considered as redundant since drug-induced gene expression profiles already include drug information such as as transcriptional effects. However, Table II shows that drug structure information helps to improve model performance. Additional drug structure information is helpful because the amount of available data is insufficient for training models to capture DILI-related features. The number of drugs annotated with DILI labels is small and the drug-induced gene expression profiles of most of the drugs annotated with DILI labels are not included in the CMap dataset. As more drug-induced gene expression profiles of drugs with DILI classes become available, the performance of GLIT will continue to improve. Moreover, when a larger dataset is available, we can use more complex deep neural network techniques such as applying a cross attention mechanism between drug features and drug-induced gene expression features.

In addition, instead of representing drugs using the ECFP method, we represented drugs as graphs of their atoms and bonds. Then, the three MLP layers of GLIT were replaced with graph neural network layers used for embedding drug structure information. However, the performance of the model that uses graph neural networks for embedding drug structure information was lower than that of GLIT. The current method for drug graph representation does not consider chirality which is important drug structure information. Therefore, we represent drugs using the ECFP method which considers chirality.

In future work, we can further improve GLIT and train it to predict the liver injury probability of drug combinations. Moreover, drug embedding vectors and drug-induced gene expression profile embedding vectors that are extracted from GLIT could be used for downstream analysis tasks such as

drug-induced lung injury prediction tasks or personalized side effect prediction tasks.

V. Conclusions

In this paper, we proposed GLIT which is a graph neural network-based drug-induced liver injury prediction model trained on transcriptional response data. GLIT achieves a CCR of 0.771 and MCC of 0.553, which are 7% and 19.2% higher, respectively, than the best performing baseline model. We believe that GLIT will prove to be a useful DILI prediction model in the drug development process. Our code and pre-trained model are publicly available at: https://github.com/dmis-lab/GLIT.

Acknowledgment

This research was supported by the National Research Foundation of Korea [NRF-2016M3A9A7916996, NRF-2017R1A2A1A17069645]; and by the National IT Industry Promotion Agency grant funded by the Ministry of Science and ICT and Ministry of Health and Welfare [NO. C1202-18-1001, Development Project of the Precision Medicine Hospital Information System (P-HIS)].

References

[1] A. Mullard, "New drugs cost us $2.6 billion to develop," 2014.
[2] F. Ballet, "Hepatotoxicity in drug development: detection, significance and solutions," *Journal of hepatology*, vol. 26, pp. 26–36, 1997.
[3] S. Babai, L. Auclert, and H. Le-Louët, "Safety data and withdrawal of hepatotoxic drugs," *Therapie*, 2018.
[4] M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong, "Fda-approved drug labeling for the study of drug-induced liver injury," *Drug discovery today*, vol. 16, no. 15-16, pp. 697–703, 2011.
[5] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada, "Open tg-gates: a large-scale toxicogenomics database," *Nucleic acids research*, vol. 43, no. D1, pp. D921–D927, 2014.
[6] F. Hammann, V. Schöning, and J. Drewe, "Prediction of clinically relevant drug-induced liver injury from structure using machine learning," *Journal of Applied Toxicology*, vol. 39, no. 3, pp. 412–419, 2019.
[7] H. Hong, S. Thakkar, M. Chen, and W. Tong, "Development of decision forest models for prediction of drug-induced liver injury in humans using a large set of fda-approved drugs," *Scientific reports*, vol. 7, no. 1, p. 17311, 2017.
[8] E. Kim and H. Nam, "Prediction models for drug-induced hepatotoxicity by using weighted molecular fingerprints," *BMC bioinformatics*, vol. 18, no. 7, p. 227, 2017.
[9] C. Zhang, F. Cheng, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silico prediction of drug induced liver toxicity using substructure pattern recognition method," *Molecular informatics*, vol. 35, no. 3-4, pp. 136–144, 2016.
[10] C. Feng, H. Chen, X. Yuan, M. Sun, K. Chu, H. Liu, and M. Rui, "Gene expression data based deep learning model for accurate prediction of drug-induced liver injury in advance," *Journal of chemical information and modeling*, 2019.
[11] H. Wang, R. Liu, P. Schyman, and A. Wallqvist, "Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses," *Frontiers in pharmacology*, vol. 10, p. 42, 2019.
[12] P. Kohonen, J. A. Parkkinen, E. L. Willighagen, R. Ceder, K. Wennerberg, S. Kaski, and R. C. Grafström, "A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury," *Nature communications*, vol. 8, p. 15932, 2017.
[13] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu *et al.*, "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles," *Cell*, vol. 171, no. 6, pp. 1437–1452, 2017.
[14] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, 2018.
[15] A. Gupta, A. T. Müller, B. J. Huisman, J. A. Fuchs, P. Schneider, and G. Schneider, "Generative recurrent networks for de novo drug design," *Molecular informatics*, vol. 37, no. 1-2, p. 1700111, 2018.
[16] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design," *Science advances*, vol. 4, no. 7, p. eaap7885, 2018.
[17] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data," *Molecular pharmaceutics*, vol. 13, no. 7, pp. 2524–2530, 2016.
[18] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.
[19] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[20] K. Liu, X. Sun, L. Jia, J. Ma, H. Xing, J. Wu, H. Gao, Y. Sun, F. Boulnois, and J. Fan, "Chemi-net: a molecular graph convolutional network for accurate drug property prediction," *International journal of molecular sciences*, vol. 20, no. 14, p. 3389, 2019.
[21] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.
[22] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, "Omnipath: guidelines and gateway for literature-curated signaling pathway resources," *Nature methods*, vol. 13, no. 12, p. 966, 2016.
[23] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research*, vol. 45, no. D1, pp. D353–D361, 2016.
[24] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D535–D539, 2006.
[25] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, and D. Zhi, "Gene2vec: distributed representation of genes based on co-expression," *BMC genomics*, vol. 20, no. 1, p. 82, 2019.
[26] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "Ncbi geo: mining millions of expression profiles—database and tools," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D562–D566, 2005.
[27] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[30] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.
[31] A. Mak, R. Kato, K. Weston, A. Hayes, and J. Uetrecht, "Editor's highlight: an impaired immune tolerance animal model distinguishes the potential of troglitazone/pioglitazone and tolcapone/entacapone to cause idili," *Toxicological Sciences*, vol. 161, no. 2, pp. 412–420, 2017.
[32] A. Mak and J. Uetrecht, "Immunization with amodiaquine-modified hepatic proteins prevents amodiaquine-induced liver injury," *Journal of immunotoxicology*, vol. 12, no. 4, pp. 361–367, 2015.
[33] S. Takai, S. Higuchi, A. Yano, K. Tsuneyama, T. Fukami, M. Nakajima, and T. Yokoi, "Involvement of immune-and inflammatory-related factors in flucloxacillin-induced liver injury in mice," *Journal of Applied Toxicology*, vol. 35, no. 2, pp. 142–151, 2015.