frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fcell.2021.700061

DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data

*Ningyi Zhang[1], Haoyan Wang[1], Chen Xu[2], Liyuan Zhang[1] and Tianyi Zang[1]**

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Center for Bioinformatics, Harbin Institute of Technology, Harbin, China

Endocrinology is the study focusing on hormones and their actions. Hormones are known as chemical messengers, released into the blood, that exert functions through receptors to make an influence in the target cell. The capacity of the mammalian organism to perform as a whole unit is made possible based on two principal control mechanisms, the nervous system and the endocrine system. The endocrine system is essential in regulating growth and development, tissue function, metabolism, and reproductive processes. Endocrine diseases such as diabetes mellitus, Grave's disease, polycystic ovary syndrome, and insulin-like growth factor I deficiency (IGFI deficiency) are classical endocrine diseases. Endocrine dysfunction is also an increasing factor of morbidity in cancer and other dangerous diseases in humans. Thus, it is essential to understand the diseases from their genetic level in order to recognize more pathogenic genes and make a great effort in understanding the pathologies of endocrine diseases. In this study, we proposed a deep learning method named DeepGP based on graph convolutional network and convolutional neural network for prioritizing susceptible genes of five endocrine diseases. To test the performance of our method, we performed 10-cross-validations on an integrated reported dataset; DeepGP obtained a performance of the area under the curve of ∼83% and area under the precision-recall curve of ∼65%. We found that type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM) share most of their associated genes; therefore, we should pay more attention to the rest of the genes related to T1DM and T2DM, respectively, which could help in understanding the pathogenesis and pathologies of these diseases.

Keywords: endocrine disease, Graves' disease, T2DM, PCOS, T1DM, IGF-I, deep learning methods

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Tao Huang,
Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, China
Xiaofang Zhao,
Institute of Computing Technology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Tianyi Zang
tianyi.zang@hit.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *25 April 2021*
Accepted: *31 May 2021*
Published: *06 July 2021*

Citation:
Zhang N, Wang H, Xu C, Zhang L and Zang T (2021) DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data. Front. Cell Dev. Biol. 9:700061. doi: 10.3389/fcell.2021.700061

INTRODUCTION

Endocrine diseases fall into broad categories of hormone over- or underproduction, modulate tissue response to hormones, or tumors caused by endocrine tissue (Belfiore and LeRoith, 2018). Hormones synthesized and released by the endocrine glands exert their functions by regulating the biological process of cells. There are several examples of common endocrine diseases: type I/II diabetes mellitus, Graves' disease (GD), polycystic ovary syndrome (PCOS), and insulin-like growth factor I (IGFI) deficiency, etc. To date, genome-wide association studies (GWAS) have reported numerous gene regions associated with different endocrine diseases. The aim of GWAS

analysis is to determine how the combined allele frequency of multiple susceptibility genes can affect autoimmunity and/or disease risk.

Graves' disease is an organ-specific autoimmune thyroid disease, resulting from excessive secretion of thyroid hormones by thyroid tissue (Dvornikova et al., 2020). The pathogenesis of GD is mediated by the production of antibodies to TSH receptors, which provide increased secretion of thyroid hormones and a rapid growth of the thyroid after stimulation (Smith et al., 2018; Soh and Aw, 2019). Since, GD is a hereditary and polygenic transmission disease (Perricone and Shoenfeld, 2019). It has been identified that associations between CTLA-4, FOXP3, TLR class polymorphism, and a number of pathological conditions develop in GD (Xiao et al., 2015; Fathima et al., 2019).

Diabetes mellitus, such as type 1 diabetes mellitus (TIDM) and type 2 diabetes mellitus (T2DM) are also a typical group of endocrine diseases. But basic pathogenic differences exist in these two types of diabetes mellitus. T1DM is immune mediated while T2DM is mediated by metabolic mechanisms (Eizirik et al., 2020). Glucagon secretion is observed to be reduced in patients with T1DM, with an increasing risk of insulin-induced hypoglycemia, but it is enhanced in T2DM, exacerbating the effects of reduced insulin release and action on glucose of blood levels (Gromada et al., 2018). Recent studies have detected several novel and promising TDM-susceptible genes, such as GCKR, SLC30A8, TLR4, and FTO (Ehrmann et al., 1999; Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004; Day et al., 2015).

Polycystic ovary syndrome is a common endocrinopathy among women, with symptoms including irregular menstrual cycles, hyperandrogenism and polycystic ovarian morphology (Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, 2004). It is also accompanied with obesity and insulin resistance, increasing the risk of diabetes, metabolic syndrome, and other cardiovascular diseases (Ehrmann et al., 1999). There are several published causal variants from GWAS studies associated with PCOS (Day et al., 2015; Hayes et al., 2015). Despite the detrimental impact of the disorder on women's health, the etiology remains poorly understood.

Insulin-like growth factor I production is mainly mediated by growth hormones (GH); both GH and IGF-I have an anabolic effect on skeletal muscle and bone. Despite the effects on growing process in childhood, GH also plays an important role in the regulation of metabolism, body composition, and mood, which persist into adult life (Gazzaruso et al., 2014). However, low serum levels of IGF-I have been detected in patients suffering from chronic liver disease and malnutrition despite normal or elevated GH secretion (Cuneo et al., 1995).

Though there have been numerous susceptible loci identified by GWAS that are associated with endocrine diseases, etiology and pathology is still unclear. Analogous to other complex traits, common susceptibility loci identified by GWAS account for only a small proportion of the genetic heritability of the traits. As GWAS were designed to detect the common allelic variants with a minor allele frequencies of 2 to 5%, the variants occurred less frequently but with greater effect sizes being ignored which may account for the observed deficit in heritability (Manolio et al., 2009). Since SNPs detected by GWAS may not be the real causative regions, the SNPs related to them may be the real causative genes of complex diseases due to the theory of linkage disequilibrium (LD). Therefore, we take the expression level of genes regulated by SNPs into account to reduce the impact of LD. eQTL research plays an important role in prioritizing SNP loci in GWAS susceptible regions (Barral et al., 2012). Previous studies usually investigate susceptible genes of complex diseases based on the regulation function of SNPs on gene expression. MR analysis, for example, is proposed to explain the causative effect on a phenotype of gene expression based on the regulation of genetic variants (Freeman et al., 2013). However, it is rarely available in practice to obtain such a large sample size of different types of data, such as phenotype, genome-wide SNP genotype, and gene expression data, to perform a MR analysis. To overcome this, a SMR method was proposed which integrates summary-level data from independent GWAS with eQTL data to identify disease susceptible genes (Zhu et al., 2016).

Machine learning methods have been widely used in prediction problems, such as support vector machine (SVM), network embedding algorithms (such as node2vec), network diffusion algorithms (such as Laplacian heat diffusion, random walk with restart), etc. Combined with multiple biomolecular features, novel biomarkers, genes, and proteins can be predicted (Chen et al., 2018, 2019; Zhang et al., 2020). Nowadays, deep learning methods have also been utilized in bioinformatics. Liu et al. (2017) applied convolutional neural network (CNN) model to identify cell cycle-regulated genes. Graph convolutional network (GCN) was utilized to predict disease-related metabolites in the study by Zhao et al. (2020). Most machine learning and deep learning methods focused on feature extraction and selection. However, no computational method has been developed to predict the susceptible genes of endocrine diseases based on integrated omics data to eliminate the LD disequilibrium bias.

In this study, we developed "DeepGP," a method to prioritize susceptible genes of endocrine diseases based on deep learning approaches. First, we obtained curated disease-gene associations of five endocrine diseases from disGeNET database; susceptible regions and expression level data were downloaded from GWAS catalog and GTEx database, respectively. After mapping the genes to susceptible loci based on position information of genes, the feature vector of each gene was composed of two types of features, a phenotype-based feature derived from GWAS dataset and a transcriptome-based feature derived from eQTL data. Disease similarity network can be obtained from our previous work, which can represent disease features. GCN was then utilized to decipher the integrated feature representations of the gene. Finally, the classification of candidate genes was performed by CNN.

MATERIALS AND METHODS

Work Frame

DeepGP contains three main parts, data preprocessing (feature extraction), feature reconstruction based on GCN, and endocrine disease-related gene prediction based on CNN. In the feature extraction process, we obtained endocrine disease-related gene

information from DisGeNET (Bauer-Mehren et al., 2010), GWAS Catalog (Buniello et al., 2019), and GTEx Portal databases (Carithers and Moore, 2015). After extracting the features of genes and diseases, we built a heterogeneous network composed of genes and diseases. We then utilized the GCN method to reconstruct the integrated gene features to obtain a more precise feature representation of each gene. In the disease gene prediction part, CNN is used to prioritize the causative genes related to endocrine diseases based on a comprehensive feature representation of disease-gene pairs. The workflow of DeepGP is shown in **Figure 1**.

Data Collection and Preprocessing

Genome-wide association studies have identified thousands of genetic variants that are associated with diseases and traits of medical importance in humans. However, the genes detected from the SNPs identified by GWAS which are pathogenic on diseases remain largely unknown due to the complicated LD between SNPs. Intuitively, genes closest to the top associated variants in position are the most likely causative genes. However, there have been studies reporting that causal genes are distinct from the nearest genes (Zhu et al., 2016). Studies have verified that gene expression may be influenced by different genetic variants among genes with different genotypes of the genetic variants, which means the phenotypes can be influenced by genetic variants through regulating the expression of their target genes. Therefore, we used a GCN network embedding method based on two types of omics data to obtain the comprehensive feature vector of each gene.

We first collected endocrine diseases from the "Endocrine diseases" chapter according to the International Classification of Diseases 11th Revision released by the World Health Organization. The major glands of the endocrine system include the pineal gland, pituitary gland, pancreas, ovaries, testes, thyroid gland, parathyroid gland, hypothalamus, and adrenal glands. Considering that some of the endocrine diseases have not been previously widely investigated, we chose five of the endocrine diseases, GD (which is mainly related to thyroid gland), T1DM/T2DM (which is mainly related to pancreas), PCOS (which is mainly related to ovaries), and IGFI deficiency (which is mainly related to pituitary gland). We then collected curated disease genes from the DisGeNET database (Piñero et al., 2016), causative loci from the GWAS database (Buniello et al., 2019), and expression level data from the GTEx database (Carithers and Moore, 2015). In addition, gene–gene interaction network was downloaded from HumanNet v2.0, where the correlation scores between a pair of genes were calculated. We then utilized an R package named biomaRt to obtain detailed information, such as gene location, chromosome number, and start and end position of each gene. BiomaRt can also be used to transform different gene IDs, such as gene symbol, Ensembl ID, and entrez ID, from different databases. After mapping the genes to the susceptible loci identified by GWAS and eQTL, genes with at least one susceptible SNP were kept. Finally, we obtained 7,406 genes, including 4,212 known causal genes and 3,194 candidate genes obtained from HumanNet v2.0.

Therefore, each gene has a feature of 2 × 25D based on GWAS and eQTL summary data according to five diseases. For the phenotype-based feature of each gene, we used the *p*-value of top 5 related SNPs of each disease to indicate the gene feature vector, for genes with less than five related SNPs, the feature vector is filled with 1. Thus, the phenotype-based feature of each gene could be denoted as:

$$G_S^i = [P_S^1 P_S^2, \cdots, P_S^{25}] \tag{1}$$

where P_S^i denotes the *p*-value of SNPs mapped by gene locations. Transcriptome-based feature of each gene can be extracted by the same method:

$$G_S^T = [P_T^1, P_T^2, \cdots, P_T^{25}] \tag{2}$$

where P_T^i denotes the *p*-value of susceptible loci consistent with that from the phenotype-based feature vector of each gene obtained from eQTL data. Thus, an initial integrated 2 × 25D feature vector of each gene is constructed.

Feature Reconstruction by GCN

In this section, we introduced a network-embedding algorithm based on GCN in order to present a new representation of gene features. GCN is a graph deep learning method based on node features and network architecture to classify the nodes

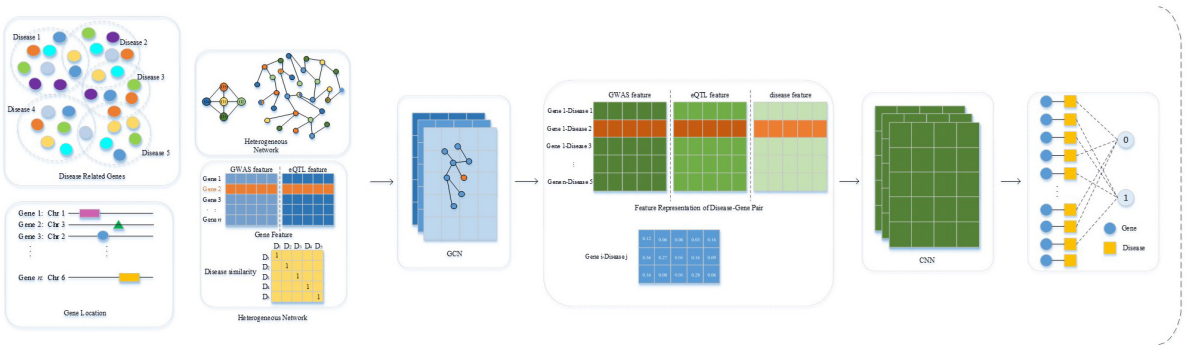

FIGURE 1 | Workflow of DeepGP.

of a network. Although GCNs have been successfully applied in other domains, to our knowledge, this is the first time that GCN was utilized to represent latent gene features from several omics datasets and network properties, while also being capable of disentangling the underlying molecular mechanisms driving the etiology of endocrine diseases.

Considering a graph $G = (V, E, W)$, where V denotes the nodes of the network, E denotes the edges of the network, and W the weight matrix encoding the interacting weight between nodes, which is obtained from HumanNet v2.0 as the gene interactions between gene pairs, the feature matrix of the nodes can be denoted as $X \in R^{N\times F}$. Thus, the eigenvectors of the graph Laplacian L can be denoted as:

$$\mathrm{L} = D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}} \tag{3}$$

where the adjacency matrix $\tilde{A} = A + I$ has added self-connections since gene nodes should contain both gene interaction and gene itself information and D is the degree matrix.

Finally, we can define a propagation rule for each layer:

$$H^{l+1} = (LH^{l}XW^{l}) \tag{4}$$

$$\mathrm{ReLU}\,(x) = \begin{cases} x, & x>0 \\ 0, & x\leq 0 \end{cases} \tag{5}$$

where σ denotes a nonlinearity, such as the Rectified Linear Unit activation function. The input of the first layer is X, which includes the gene expression feature and gene variation feature of five interested traits, so H^0=X. Therefore, the feature of the gene network could be extracted by formula (4). Since we combined two omics types of data, each gene feature could be represented as a 2 × 25D vector. Thus, we have a feature representation of each gene; the feature can be denoted as: $g_i = \begin{bmatrix} P^1_{i,s}, P^2_{i,s}, P^3_{i,s}, \cdots, P^{25}_{i,s} \\ P^1_{i,T}, P^2_{i,T}, P^3_{i,T}, \cdots, P^{25}_{i,T} \end{bmatrix}$.

Causal Gene Prediction With CNN

After obtaining the best combination of initial feature representation by GCN, we constructed disease features based on disease similarity matrix calculated by the method ImpAESim. Each disease feature can be denoted as a 1 × 5D vector:

$$D_i = \{S_{i,j}\}, \mathrm{j} = 1, 2, 3, 4, 5 \tag{6}$$

$S_{i,j}$ denotes the similarity between D_i and D_j, $S_{i,j}$ is 1 if $i = j$. We then combined the gene feature and disease feature as a 3 × 5D feature of disease–gene pair.

We then trained a CNN model to predict causal genes based on the gene features derived from GCN. Analogous to other machine learning methods, CNN consists of a training step where the estimation of network parameters from a given training dataset is learned, and the testing step utilized the well-trained network to predict outputs of new testing dataset (Min et al., 2017). Since our feature format of each gene–disease pair is 3 × 5D, which can be regarded as an image with three channels, in this work, the structure of the CNN section is shown in **Figure 2**. The CNN section includes four parts as follows: convolution layer, max-pooling layer, fully connected layer, and an output layer. Convolution layer is responsible for extracting the subspace features of the input. Max-pooling layer is used for dimension reduction to discard the redundant information. The final fully connected layer connects all the nodes and the output layer applying sigmoid as the activation function to solve the binary classification problem.

The tanh function is the activation function in each convolutional layer.

$$\tanh\,(x) = \frac{\sinh x}{\cosh x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{7}$$

The sigmoid function is used as the activation function in the output layer.

$$\delta\,(x) = 1/(1 + e^{-x}) \tag{8}$$

Since our disease gene prediction can be treated as a binary classification problem, we chose the binary cross-entropy function as the loss function to assess the probability of the output.

$$\mathrm{loss} = -\sum_{i\ 1}^{n} y_i' \log\,(y_i) + (1 - y_i')\log(1 - y_i') \tag{9}$$

$$\frac{\partial \mathrm{loss}}{\partial y} = -\sum_{i=1}^{n} \frac{y_i'}{y_i} - \frac{1 - y_i'}{1 - y_i'} \tag{10}$$

According to formulas (7, 8), loss is 0 as long as y_i' is equal to y_i .

As a result, each disease–gene pair was assigned a correlation score with a range of [0, 1], where 1 denotes the pair having the strongest association and 0 means no association.

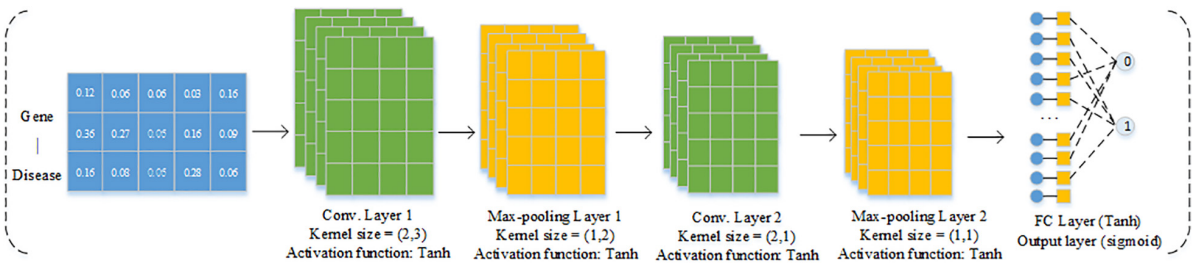

FIGURE 2 | Structure of CNN model.

Training Steps

According to the 4,212 curated disease-related genes to five endocrine diseases, there are 6,258 positive disease–gene pairs derived from the curated disease–gene associations, and 14,802 (4,212 × 5–6,258) pairs are not reported to be associated; we randomly selected 6,258 pairs to construct the negative samples. However, the sample size of different diseases are extremely unbalanced, as shown in **Table 1**; the sample size ranges from 28 IGFI–gene pairs to 3,058 T2D–gene pairs due to the insufficiency of studies related to these diseases, which may have a serious negative impact on the classification performance. From the analysis for summary data derived from GWAS and eQTL, T1D, and T2D share as much as 950 genes in total, with 679/1,629 genes merely related to T1D and 2,108/3,058 genes merely related to T2D.

After obtaining the new dataset that consisted of 6,258 positive samples and 6,258 negative samples, we conducted a 10-cross-validation on this new dataset to test the performance. First, the dataset is randomly divided into 10 groups, then 10 times of iterations were performed based on nine of 10 groups as training set and one group as test set, which made sure that each group can be used as an independent test set.

RESULTS

Performance Evaluation on Predicting Disease–Gene Associations

The area under the curve (AUC) and the area under the precision-recall curve (AUPR) are used to assess DeepGP. The AUC and AUPR of each iteration in 10-cross-validation process are shown in **Table 2**. As a result, DeepGP achieved a mean AUC of 0.845 and a mean AUPR of 0.833, which have shown better and stable in disease–gene prediction.

TABLE 1 | Number of curated disease genes.

Disease	T2D	T1D	PCOS	GD	IGFI deficiency
No. of samples	3,058	1,629	974	568	28

TABLE 2 | AUC and AUPR of DeepGP in 10 times 10 cross-validation.

	1	2	3	4	5	6	7	8	9	10	Average
AUC	0.832	0.845	0.821	0.854	0.864	0.855	0.831	0.861	0.856	0.831	0.845
AUPR	0.827	0.837	0.816	0.845	0.838	0.825	0.816	0.858	0.842	0.826	0.833

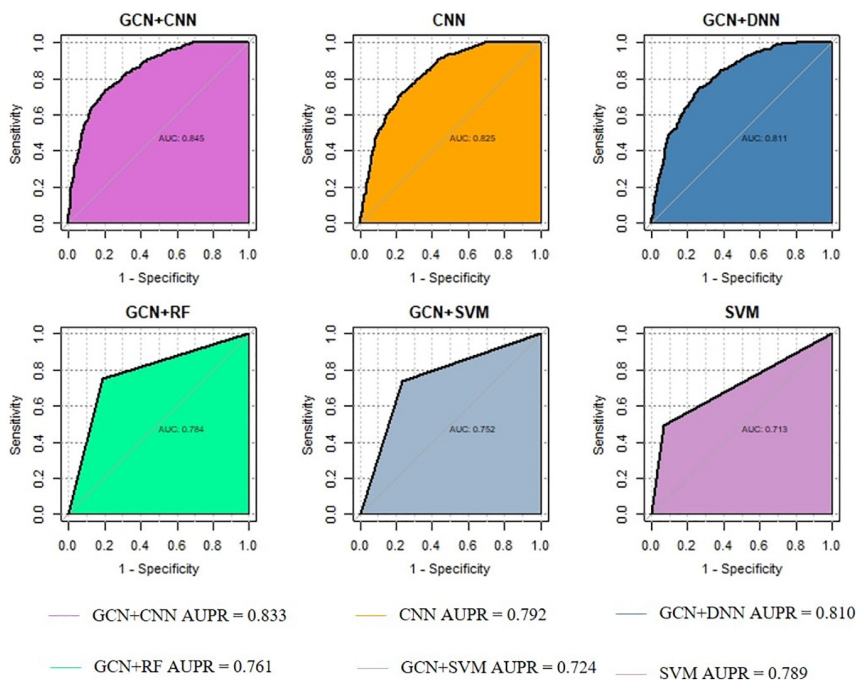

FIGURE 3 | Performance comparison among different methods.

Comparison Experiments With Classic Methods

We evaluated the performance of DeepGP and classic machine learning methods, such as SVM, random forest (RF), Naïve Bayes, and deep neural network (DNN) for predicting disease–gene associations. To validate the performance of extracting best combination of gene features by GCN, we alternatively only used CNN for feature extraction and prediction. Then, to assess the effect of convolution layers in CNN, we applied a typical deep learning method DNN and two classic machine learning methods SVM and RF. As a contrast, we also only used a SVM model for the classification task. Therefore, we compared the performance of five methods with DeepGP: CNN, GCN-DNN, GCN-SVM, GCN-RF, and SVM.

The method of training and testing was performed the same as DeepGP. As shown in **Figure 3**, comparing with other classic machine learning methods, DeepGP achieves the highest performance according to both AUC and AUPR. CNN and DNN achieved the second and third highest AUROC which infers deep learning methods are better than classic machine learning methods in this disease–gene prediction task; however, it can also be inferred that convolution layers are essential. In addition, the performance was improved after feature encoding by GCN.

It has been shown that the depth of CNN models can affect the classification performance. To assess the influence of the depth of CNN models, we compared DeepGP with a shallower CNN model, denoted as GCN+CNN (B), consisting of one set of one convolution layer and one max-pooling layer. Since the feature dimension of each disease–gene pair is 3×5, we also enlarged the kernel size to detect the effect of dimension reduction. The performance is shown in **Figure 4**. GCN+CNN (A) denotes the structure of DeepGP, GCN+CNN (B) denotes the shallower CNN model, and GCN+CNN (C) denotes the structure of CNN with enlarged kernel size. As a result, DeepGP achieved the best performance comparing with other methods even though the feature dimension of each sample is low. Hence, according to the above results, we conclude that our proposed DeepGP is competitive against other methods.

Validation of Prediction Results

After verifying the effectiveness of DeepGP based on the comparison experiments. We conducted the disease–gene prediction process among all the unknown disease–gene pairs. We set a threshold of 0.5 as default to screen the scores and identified 7,702 of 14,903 pairs to be true. Among the predicted associations, 971 novel genes were identified to be associated with GD, 3,142 novel genes associated with PCOS, and 2,437 and 1,154 novel genes associated with T1DM and T2DM, respectively. Due to deficiency of studies focused on IGFI deficiency, only 28 of 4,212 were reported as curated genes related to the trait. The highest score of IGFI gene pairs was 0.33, which is under the threshold, and was predicted to be have no

TABLE 3 | Top 5 related genes with four diseases.

GD	PCOS	T1D	T2D
CCL27	RBM14	ADM2 (+)	UPK3B (+)
CXCL16	miR-1307	Enho (+)	miR-592
BECN1	CMKLR1	FUT6 (+)	NELFCD (+)
PROX1	AKT3	FUT7	miR-589 (+)
PTX3	GCGR	ATRNL1 (+)	Linc00641 (+)

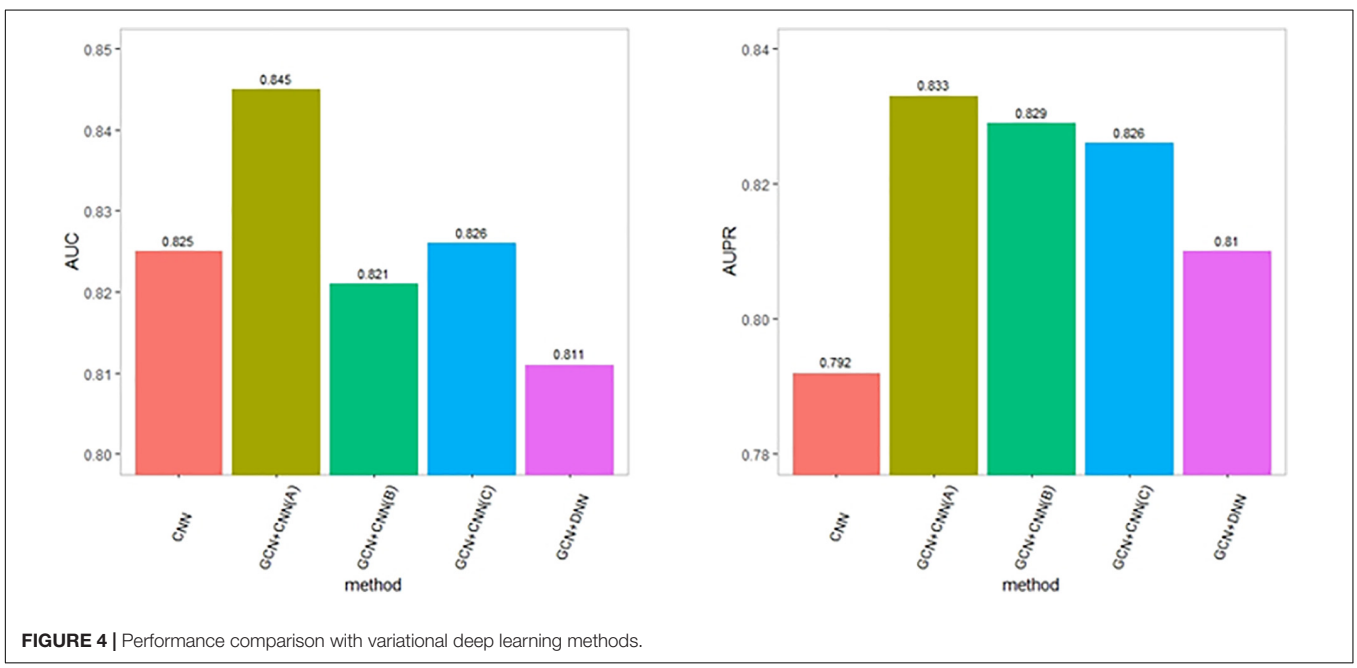

FIGURE 4 | Performance comparison with variational deep learning methods.

associations by DeepGP. For the rest of the diseases, 772 of 3,778 genes were shared.

Case Study

Graves' Disease

Graves' disease is mainly mediated by T cells which produce cytokines and chemokines in abnormal amounts. CCL27 is reported to be associated with serum chemokine concentrations detected in GD, which might be a good biomarker for GD (Hiratsuka et al., 2015). CXCL16 have been demonstrated to bind to the unique receptor CXCR6, which is expressed on a subset of multiple types of T cells, and it has been implicated in the pathogenesis of atherosclerosis and GD (Günther et al., 2012). It has been identified that loci at BECN1 can be denoted as a significant extract differentially methylated region for Graves' orbitopathy under the context of GD (Shi et al., 2019).

Polycystic Ovary Syndrome

RBM14 and miR-1307 have been reported to be up-regulated in PCOS patients in the study of Xu et al. (2015) and Che et al. (2020). Besides, CMKLR1, known as chemerin chemokine-like receptor 1, is the receptor of chemerin which is expressed at both mRNA and protein levels in human granulosa cells, and it has been reported to vary in women with PCOS (Bongrani et al., 2019).

Diabetes Mellitus

Due to the pathophysiological characteristics and many other potential etiopathogenesis factors T1DM and T2DM share, there are genes linked to both diseases such as *GLIS3* (Mahajan et al., 2014), *EIL2AK3* (also named as *PERK*; Delépine et al., 2000), etc. According to the result obtained from our method, we identified 2,582 genes related to T1D and 1,153 genes related to T2D, with a number of 474 shared genes. We then searched the top-ranked predicted genes related to T1D and T2D. According to the work of Ahmed et al., ADM2 is found to be preferentially up-regulated by bacteroides dorei (BD), which is a bacteria increased significantly at the time of onset of T1D. Enho is an energy homeostasis-associated gene that can produce a regulatory peptides named adropin, which has been identified to be strongly associated with type 1 diabetes in children (Polkowska et al., 2019). *FUT7* gene has been demonstrated to be linked with an antigen termed bile salt-dependent lipase which is reported to be associated with type 1 diabetes (Panicot et al., 1999). UPK3B is regarded as a mesothelial-like cell marker of a major adipocyte progenitor cell subpopulations which may induce adipocyte dysfunction in visceral adipose tissue in type 2 diabetes (Strieder-Barboza et al., 2020). miR-592 is reported to be associated with T2D due to its background of insulin resistance by Song et al. (2019). NELFCD has been identified to be matched with risk haplotypes across five FDM-risk haplotype, which is further identified in a common T2D gene, *ANK1*.

Therefore, we can illustrate that the predicted disease-gene pairs are reliable from the case studies mentioned above; top related genes with diseases are shown in **Table 3**. The "+" means the gene identified with T1D (T2D) is also related to T2D (T1D) according to the DisGeNET database.

DISCUSSION

In summary, we proposed a disease gene prediction method based on integrated deep learning models. We construct gene features considering both biological process and "linkage disequilibrium" theory. Identifying disease genes merely based on disease-susceptible loci identified by GWAS studies may be inaccurate due to LD. It has been indicated that genetic variants can affect the phenotype by regulating the gene expression level. In this study, eQTL data are also utilized to extract gene features. To fully use the underlying information contained in a gene interaction network, GCN was applied to extract comprehensive gene features based on the constructed gene network. Therefore, our method exploits the predictive power derived from complementary data types and the underlying network simultaneously. Disease features are derived based on disease similarity, which is calculated by the method named ImpAESim. Finally, we combined gene and disease features as a disease–gene pair feature. CNN was then used to classify the disease–gene pairs as a binary classification task.

As a result, DeepGP achieved an average AUC of 0.845 and an average AUPR of 0.833 after a 10 times 10-cross-validation based on the constructed training set, which is superior to other classic machine learning and deep learning methods. We then used the well-trained model to predict the novel disease-gene associations based on the disease–gene pairs which have not been reported to be associated before. We verified the prediction results based on a case study. Most of the top disease-related genes have reported evidences to illustrate that the genes may have associations with the diseases. In addition, we also identified 474 genes shared by T1DM and T2DM which may be helpful in designing therapeutic methods for diabetes mellitus patients. Therefore, the novel disease genes identified by DeepGP provide a strong support for the feasibility of extracting diagnostic markers for future validation and shed light on new strategies for the diagnosis and treatment of endocrine diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

NZ did most of the work and wrote the manuscript. HW, CX, and LZ assisted in completing the experiments. TZ proposed the idea of the manuscript. HW made a great contribution in modifying the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Barral, S., Bird, T., Goate, A., Farlow, M. R., Diaz-Arrastia, R., Bennett, D. A., et al. (2012). Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* 78, 1464–1471. doi: 10.1212/wnl.0b013e3182553c48

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Belfiore, A., and LeRoith, D. (2018). *Principles of Endocrinology and Hormone Action*. New York, NY: Springer.

Bongrani, A., Mellouk, N., Rame, C., Cornuau, M., Guérif, F., Froment, P., et al. (2019). Ovarian expression of adipokines in polycystic ovary syndrome: a role for chemerin, omentin, and apelin in follicular growth arrest and ovulatory dysfunction? *Int. J. Mol. Sci.* 20:3778. doi: 10.3390/ijms20153778

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Res.* 47, D1005–D1012.

Carithers, L. J., and Moore, H. M. (2015). *The Genotype-Tissue Expression (GTEx) Project*. New Rochelle, NY: Mary Ann Liebert, Inc.

Che, Q., Liu, M., Zhang, D., Lu, Y., Xu, J., Lu, X., et al. (2020). Long noncoding RNA HUPCOS promotes follicular fluid androgen excess in PCOS patients via aromatase inhibition. *J. Clin. Endocrinol. Metab.* 105, 1086–1097. doi: 10.1210/clinem/dgaa060

Chen, L., Zhang, Y.-H., Huang, G., Pan, X., Huang, T., and Cai, Y. D. (2019). Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms. *Gene Ther.* 26, 465–478. doi: 10.1038/s41434-019-0099-y

Chen, L., Zhang, Y.-H., Zhang, Z., Huang, T., and Cai, Y. D. (2018). Inferring novel tumor suppressor genes with a protein-protein interaction network and network diffusion algorithms. *Mol. Ther. Methods Clin. Dev.* 10, 57–67. doi: 10.1016/j.omtm.2018.06.007

Cuneo, R. C., Hickman, P. E., Wallace, J. D., Teh, B. T., Ward, G., Veldhuis, J. D., et al. (1995). Altered endogenous growth hormone secretory kinetics and diurnal GH—binding protein profiles in adults with chronic liver disease. *Clin. Endocrinol.* 43, 265–275. doi: 10.1111/j.1365-2265.1995.tb02031.x

Day, F. R., Hinds, D. A., Tung, J. Y., Stolk, L., Styrkarsdottir, U., Saxena, R., et al. (2015). Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* 6, 1–7.

Delépine, M., Nicolino, M., Barrett, T., Golamaully, M., Lathrop, G. M., and Julier, C. (2000). EIF2AK3, encoding translation initiation factor 2-α kinase 3, is mutated in patients with Wolcott-Rallison syndrome. *Nat. Genet.* 25, 406–409. doi: 10.1038/78085

Dvornikova, K. A., Bystrova, E. Y., Platonova, O. N., and Churilov, L. P. (2020). Polymorphism of toll-like receptor genes and autoimmune endocrine diseases. *Autoimmun. Rev.* 19:102496. doi: 10.1016/j.autrev.2020.102496

Ehrmann, D. A., Barnes, R. B., Rosenfield, R. L., Cavaghan, M. K., and Imperial, J. (1999). Prevalence of impaired glucose tolerance and diabetes in women with polycystic ovary syndrome. *Diabetes Care* 22, 141–146. doi: 10.2337/diacare.22.1.141

Eizirik, D. L., Pasquali, L., and Cnop, M. (2020). Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* 16, 349–362. doi: 10.1038/s41574-020-0355-7

Fathima, N., Narne, P., and Ishaq, M. (2019). Association and gene–gene interaction analyses for polymorphic variants in CTLA-4 and FOXP3 genes: role in susceptibility to autoimmune thyroid disease. *Endocrine* 64, 591–604. doi: 10.1007/s12020-019-01859-3

Freeman, G., Cowling, B. J., and Schooling, C. M. (2013). Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int. J. Epidemiol.* 42, 1157–1163. doi: 10.1093/ije/dyt110

Gazzaruso, C., Gola, M., Karamouzis, I., Giubbini, R., and Giustina, A. (2014). Cardiovascular risk in adult patients with growth hormone (GH) deficiency and following substitution with GH—an update. *J. Clin. Endocrinol. Metab.m* 99, 18–29. doi: 10.1210/jc.2013-2394

Gromada, J., Chabosseau, P., and Rutter, G. A. (2018). The α-cell in diabetes mellitus. *Nat. Rev. Endocrinol.* 14, 694–704.

Günther, C., Carballido-Perrig, N., Kaesler, S., Carballido, J. M., and Biedermann, T. (2012). CXCL16 and CXCR6 are upregulated in psoriasis and mediate cutaneous recruitment of human CD8+ T cells. *J. Invest. Dermatol.* 132, 626–634. doi: 10.1038/jid.2011.371

Hayes, M. G., Urbanek, M., Ehrmann, D. A., Armstrong, L. L., Lee, J. Y., Sisk, R., et al. (2015). Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat. Commun.* 6, 1–13.

Hiratsuka, I., Itoh, M., Yamada, H., Yamamoto, K., Tomatsu, E., Makino, M., et al. (2015). Simultaneous measurement of serum chemokines in autoimmune thyroid diseases: possible role of IP-10 in the inflammatory response. *Endocr. J.* 62, EJ15–EJ0448.

Liu, C., Cui, P., and Huang, T. (2017). Identification of cell cycle-regulated genes by convolutional neural network. *Comb. Chem. High Throughput Screen.* 20, 603–611.

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46, 234–244. doi: 10.1038/ng.2897

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Min, S., Lee, B., and Yoon Deep, S. (2017). learning in bioinformatics. *Brief. Bioinformatics* 18, 851–869.

Panicot, L., Mas, E., Thivolet, C., and Lombardo, D. (1999). Circulating antibodies against an exocrine pancreatic enzyme in type 1 diabetes. *Diabetes* 48, 2316–2323. doi: 10.2337/diabetes.48.12.2316

Perricone, C., and Shoenfeld, Y. (2019). *Mosaic of Autoimmunity: the Novel Factors of Autoimmune Diseases*. Cambridge, MA: Academic Press.

Piñero, J., Bravo, À, Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45:gkw943.

Polkowska, A., Pasierowska, I. E., Pasławska, M., Pawluczuk, E., and Bossowski, A. (2019). Assessment of serum concentrations of adropin, afamin, and neudesin in children with type 1 diabetes. *BioMed. Res. Int.* 2019:6128410.

Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* 81, 19–25. doi: 10.1016/j.fertnstert.2003.10.004

Soh, S-B, and Aw, T-C (2019). Laboratory testing in thyroid conditions-pitfalls and clinical utility. *Ann. Lab. Med.* 39:3. doi: 10.3343/alm.2019.39.1.3

Shi, T.-T., Hua, L., Xin, Z., Li, Y., Liu, W., and Yang, Y. L. (2019). Identifying and validating genes with DNA methylation data in the context of biological network for Chinese patients with Graves' orbitopathy. *Int. J. Endocrinol.* 2019:6212681.

Smith, M. J., Rihanek, M., Coleman, B. M., Gottlieb, P. A., Sarapura, V. D., and Cambier, J. C. (2018). Activation of thyroid antigen-reactive B cells in recent onset autoimmune thyroid disease patients. *J. Autoimmun.* 89, 82–89. doi: 10.1016/j.jaut.2017.12.001

Song, Y., Wu, L., Li, M., Xiong, X., Fang, Z., Zhou, J., et al. (2019). Down-regulation of MicroRNA-592 in obesity contributes to hyperglycemia and insulin resistance. *EBioMedicine* 42, 494–503. doi: 10.1016/j.ebiom.2019.03.041

Strieder-Barboza, C., Flesher, C. G., Geletka, L. M., Orourke, RW., and Lumeng, CN. (2020). 1973-P: single-Nuclei transcriptomics of human adipose tissue identify distinct adipocyte progenitor subpopulations in type 2 diabetes. *Am. Diabetes Assoc.* 69(Suppl. 1).

Xiao, W., Liu, Z., Lin, J., Li, J., Wu, K., Ma, Y., et al. (2015). Polymorphisms in TLR1, TLR6 and TLR10 genes and the risk of Graves' disease. *Autoimmunity* 48, 13–18.

Xu, B., Zhang, Y.-W., Tong, X.-H., and Liu, Y. S. (2015). Characterization of microRNA profile in human cumulus granulosa cells: identification of microRNAs that regulate Notch signaling and are associated with PCOS. *Mol. Cell. Endocrinol.* 404, 26–36. doi: 10.1016/j.mce.2015.01.030

Zhang, H., Wang, S., and Huang, T. (2020). Identification of chronic hypersensitivity pneumonitis biomarkers with machine learning and differential co-expression analysis. *Curr. Gene Ther.* 21.

Zhao, T., Hu, Y., and Cheng, L. (2020). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinformatics* 20:bbaa212.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487. doi: 10.1038/ng.3538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.