Microbial Gene Ontology informed deep neural network for microbe functionality discovery in human diseases

Yunjie Liu [1,*], **Yao-zhong Zhang** [1] **and Seiya Imoto** [1]

[1] *Institute of Medical Science, The University of Tokyo, Tokyo, Japan*

Correspondence*:
Corresponding Author
liuyj@hgc.jp

ABSTRACT

The human microbiome plays a crucial role in human health and is associated with a number of human diseases. Determining microbiome functional roles in human diseases remains a biological challenge due to the high dimensionality of metagenome gene features. However, existing models were limited in providing biological interpretability, where the functional role of microbes in human diseases is unexplored. Here we propose to utilize a neural network-based model incorporating Gene Ontology (GO) relationship network to discover the microbe functionality in human diseases. We use four benchmark datasets, including diabetes, liver cirrhosis, inflammatory bowel disease, and colorectal cancer, to explore the microbe functionality in the human diseases. Our model discovered and visualized the novel candidates' important microbiome genes and their functions by calculating the important score of each gene and GO term in the network. Furthermore, we demonstrate that our model achieves a competitive performance in predicting the disease by comparison with other non-Gene Ontology informed models. The discovered candidates' important microbiome genes and their functions provide novel insights into microbe functional contribution.

Keywords: Gene Ontology, neural network, metagenome, microbe functionality

1 INTRODUCTION

With the development of metagenome sequencing technologies, a large number of studies were designed to find the association between the human microbiome and human disease Qin et al. (2012, 2014); Le Chatelier et al. (2013); Qin et al. (2010). The human microbiome has been shown to play an important role in diseases such as type II diabetes (T2D), liver cirrhosis, and obesity. Therefore, discovering the human microbiome's roles in human diseases will guide the researchers to explain these diseases from a metagenome aspect. However, describing the microbe's roles in human diseases remains a biological challenge due to the complexity of discovering and summarizing the microbe's roles with many microbes.

To address the problem, various machine learning and deep learning models were designed LaPierre et al. (2019); Pasolli et al. (2016); Reiman et al. (2020); Nguyen (2019); Rahman and Rangwala (2018). These models extracted different microbiome features and evaluated the significance of these features by predicting the diseases. LaPierre et al. LaPierre et al. (2019) compared the performance of different machine learning and deep learning models in predicting human diseases using different metagenome features. They extracted the taxa abundance feature using MetaPhlAn2 Truong et al. (2015) and k-mer abundance

feature using Jellyfish Marçais and Kingsford (2011) in their experiments and predicted the diseases using these features separately. They have shown the potential of deep learning models in disease predicting tasks. However, there are some limitations to these methods. On the one hand, the taxa abundance feature gives limited functional information and hampers people from understanding how the microbes affect the diseases. On the other hand, using k-mer abundance feature or deep learning models has limited biological interpretability. It is difficult to understand the mechanisms underlying the prediction results.

Instead of using the taxa abundance feature and k-mer abundance feature, functionality feature such as KEGG provides the functional aspect to explain the microbes' role in human diseases Kanehisa and Goto (2000). Traditionally, statistics-based models were used to identify the disease-associated function features Qin et al. (2012); Karlsson et al. (2013); Forslund et al. (2015). These statistics-based models identify the disease-associated functions that significantly differ between case and control groups. However, these models were typically based on the linear or independent assumption. These models will not detect features that have a complex relationship with diseases.

Recently, a novel interpretable deep learning model named P-NET was developed to predict treatment resistance in prostate cancer patients using the biologically informed hierarchical structure Elmarakeby et al. (2021). They demonstrated that P-NET could predict cancer state using molecular data with a performance superior to other machine learning models. However, whether a microbe functionality-informed hierarchical structure can effectively predict the diseases is still unknown. Furthermore, solving the problem is challenging due to the high dimensionality of metagenome genes and functionality features.

Another study named ParsVNN was designed to discover the cancer-specific, and drug-sensitive genes Huang et al. (2021). ParsVNN used GO hierarchical structure in building the visible neural network and pruned the edges in the network to remove the redundant features. However, the network is not applicable in handling metagenome data due to the high dimensionality of metagenome genes and functionality features. The network will be constructed with too many parameters before pruning the edges.

To address the above problems, we proposed a novel interpretable deep learning model utilizing GO hierarchical structure, which describes genes in molecular function, biological process, and cellular component Ashburner et al. (2000); Consortium (2020a), to interpret the microbiome functional roles in different human diseases. We performed our model on diabetes, liver cirrhosis, inflammatory bowel disease, and colorectal cancer, which showed the model's effectiveness. Furthermore, our model discovered the novel candidates' important microbiome genes and their functions by calculating the important score of each gene and GO term in the network in both diabetes and liver cirrhosis.

2 MATERIALS AND METHODS

2.1 Data preparation

The workflow of the pipeline is shown in Figure. 1. Raw metagenome sequencing data obtained in previous studies were used as input in the pipeline. AfterQC is used to perform quality control, including filtering low-quality reads and trimming adapters Chen et al. (2017). Human genomes are removed using Bowtie2, where the reference genome is Genome Reference Consortium Human Build 38 (GRCh38) Langmead and Salzberg (2012). Reads after quality control are aligned to the Unified Human Gastrointestinal Genome (UHGG) collection Almeida et al. (2021) by bowtie2, and quantification of genes is performed by using Salmon Patro et al. (2017). Transcript Per Million (TPM) is used as the gene quantification profile, which

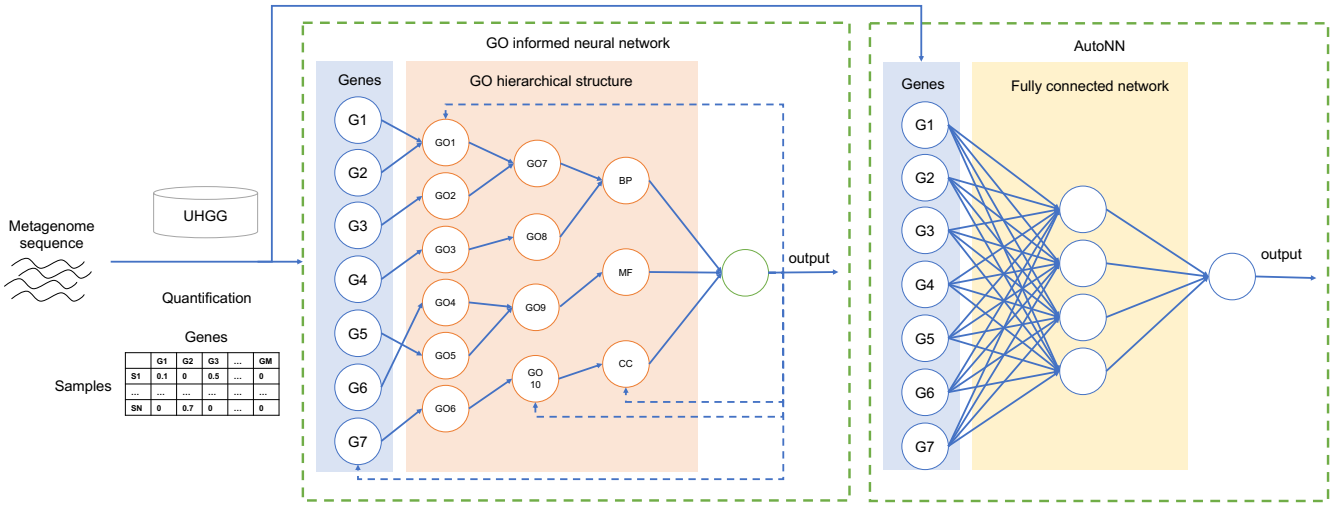


Figure 1. The workflow of the pipeline. Metagenome sequencing reads after quality control are aligned to the UHGG collection. After alignment, TPM is used as the gene quantification profile, which is used as the input of the GO-informed neural network and non-GO-informed neural network (AutoNN). In the GO-informed neural network, the solid arrow in the network is determined by Gene functional annotation information and GO hierarchical structure. The dashed arrows show the direction of calculating the importance score. The network output the disease prediction result and important genes and GO terms candidates. In AutoNN, the network is fully connected. Abbreviations: BP (biological process), MF (molecular function), and CC (cellular component)

is used as the GO network-informed deep neural network (DNN) input. The GO annotation information is obtained by referring to the UniProt database Consortium (2020b).

In this study, we prepare metagenome sequencing data for four diseases shown in Table 1. We select commonly existing genes from the UHGG gene catalog to reduce the feature dimension. We prepare a gene set where the genes exist in more than 1% of UHGG samples as commonly existing genes. The number of genes selected from the UHGG gene catalog is 22,927. To compare our model and other machine learning models, we filtered out the non-GO annotated genes and obtained the final selected gene set with 8,010 genes.

2.2 GO network informed DNN construction

GO describes genes in three aspects: molecular function, biological process, and cellular component, which are also three GO terms in the GO hierarchical structure. These three GO terms represent the roots of the ontologies, respectively. In our network, each node represents a GO term. There are different relations between GO terms where we choose the main relation: *is a, part of, has part, regulates (including positively regulates and negatively regulates)* as edges in our research. We define the root nodes as level 1, children nodes that directly relate to the root nodes as level 2, and so on. The distribution of GO terms in each level is shown in Figure. 2. The first six levels are used in constructing the neural network. Metagenome genes are annotated to level 6 by following rules: genes have annotation GO terms in level 1 to level 5 are not connected with level 6 GO terms; genes have annotation GO terms in level 6 to level 12 are connected with level 6 GO terms, where higher level GO terms are mapped to their ancestor GO terms in level 6.

In the whole neural network, the input layer x_{gene} represents the metagenome genes quantification profile. The second layer represents the level 6 GO terms nodes. The output of the input layer is calculated as $y = f[(M * W)^T x_{gene} + \mathbf{b}]$, where M is the mask matrix, W is the weights matrix, $\mathbf{b}$ is the bias vector,

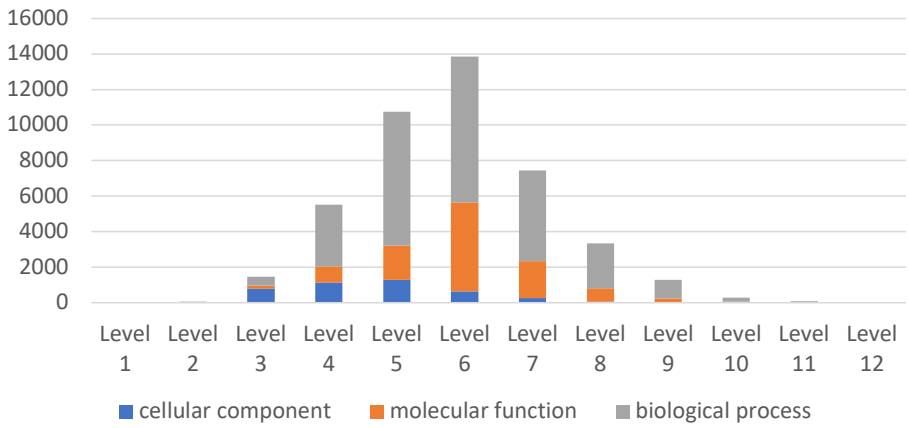


Figure 2. GO terms distribution in each level.

$*$ is Hadamard product, and the activation function f is $f = \tanh = (e^{2x} - 1)/(e^{2x} + 1)$. The mask matrix from genes to level 6 GO terms nodes is defined as a binary matrix $M \in \{0, 1\}^{D_x \cdot D_y}$ where D_x is gene number and D_y is the level 6 GO terms number. When the gene i is annotated to GO terms j, $M[i, j] = 1(1 \leq i \leq D_x, 1 \leq j \leq D_y)$; otherwise, $M[i, j] = 0$. The Hadamard product $*$ product each element of mask matrix M and weight matrix W to zero out all the connections that do not exist in the annotation. The second layer to the seventh layer represent the GO network where the output of each layer is calculated as $y = f[(M * W)^T x_{layer_i} + \mathbf{b}](i = 6, 5, 4, 3, 2, 1)$. The mask matrix zeros out the not connected GO terms in the network. We add a predictive layer with sigmoid activation $\sigma = 1/(1 + e^{-x})$ after each hidden layer to calculate the final prediction by taking the average of all the predictive elements in the network.

To obtain the important GO terms or important genes in the network, we use the DeepLIFT scheme as implemented in P-NET Elmarakeby et al. (2021); Shrikumar et al. (2017). The DeepLIFT solution calculates important scores via backpropagation and can find the example-specific explanations when given an example and output. In our case, the calculated important scores can show how the input genes affect the disease through the GO function network. Given a certain sample s, $n_1, ..., n_l$ to be the number of nodes in the certain layer, and the specific target y, DeepLIFT calculates an importance score $C_i^{l,s}$ for each node i based on the difference in the target activation $y - y_0$ fed by the certain sample s. The difference in target activation equals the sum of all node scores when fed by the certain sample s. That is,

$$\Delta y = y - y_0 = \sum_{i=1}^{n_l} C_i^{l,s} \tag{1}$$

We calculate the sample-level importance of all nodes in all layers using the 'Rescale rule' in DeepLIFT and calculate the total node-level importance score C_i^l by aggregating the sample-level importance score over all the n_s testing set samples.

$$C_i^l = |\sum_{i=1}^{n_s} C_i^{l,s}| \tag{2}$$

To reduce the bias introduced by over-annotation of certain nodes, we adjust the node important score by node degree d_i^l. The adjust node important score is calculated by:

$$adjusted_C_i^l = \begin{cases} \frac{C_i^l}{d_i^l}, d_i^l > \mu + 5\sigma \\ C_i^l \quad \text{(otherwise)}, \end{cases}$$

where μ is the mean of node degrees and σ is the standard deviation of node degrees.

2.3 Evaluation protocols

The evaluation protocols of the models were divided into two steps. In the first step, we divided our dataset into 90% of the training and validation set and 10% of the testing set for hyperparameter tuning. We performed 10-fold cross-validation on the training-validation set to obtain the best hyperparameter settings for each model. In the second step, we performed a 10-fold cross-validation on the whole dataset to obtain the final evaluation metrics. The prediction performance was measured using accuracy, precision, recall, AUC, AUPRC, and F1score.

2.4 Parameters settings

For GO informed model, we use hyperparameters grid search to find the best settings. The details of hyperparameters settings were: learning rate (0.01 / 0.005 / 0.001 / 0.0005 / 0.0001); regularizers (l1/l2/l1l2); reg_weight patterns (pattern1:[2,7,20,54,148,400](P-net default) / pattern2:[1,2,2,4,16,395](1/layer nodes number) / pattern3:[1,1,1,1,1,1] / pattern4:[1,2,4,8,16,32]). To compare the disease predicting performance with GO informed model and the non-GO informed model, we utilized the AutoNN model as baseline LaPierre et al. (2019). AutoNN is a fully connected neural network with a certain hidden layer, and the number of nodes in each hidden layer is determined by the number of input layer nodes and layer number. The details of hyperparameters settings of AutoNN were: hidden layer (1/2/3); drop rate (0/0.1); learning rate (0.01/0.001); and Adam optimizer. The number of learnable parameters of each model is shown in Table 2. AutoNN-hx represents for AutoNN model with x hidden layer.

Additionally, we compare the disease predicting performance with different machine learning models: support vector machine (SVM), random forest (RF), and logistic regression (LR). We use hyperparameters grid search to find the best settings. The details of hyperparameters settings were: the type of kernel (linear/polynomial) and the error term penalty (0.25/0.5/0.75/1.0/1.25/1.5/1.75/2.0) for the SVM; the splitting criterion (entropy/gini), the maximum tree depth (2/6/10), and the number of estimators (10/50/100) for the RF; the penalty (l1/l2/elasticnet), solver (newton-cg/lbfgs/liblinear/sag/saga) and the inverse of regularization strength (0.25/0.5/0.75/1.0/1.25/1.5/1.75/2.0) for the logistic regression.

3 RESULTS

3.1 Disease predicting performance comparison

To evaluate the effectiveness of the GO-informed model, we compared the disease predicting performance between the GO-informed model and non-GO-informed models. The classification results in T2D, liver cirrhosis dataset, inflammatory bowel disease, and colorectal cancer is shown in Table 3. We found that GO-NN has a better performance in the diabetes dataset (AUC=0.778), inflammatory bowel disease dataset (F1score=0.876, AUPRC=0.979), and colorectal cancer dataset (F1score=0.841, AUPRC=0.937). RF has a better performance in the liver cirrhosis dataset (AUC=0.974). In addition, the GO-informed model performs better than the non-GO-informed neural network model (AutoNN).

3.2 GO informed neural network visualization

To understand how the microbes affect human diseases, we visualized the GO-informed neural network after training the diabetes dataset (Figure. 3) and inflammatory bowel disease dataset (Figure. 4). The first layer represents genes; the next layer represents GO terms in level 6, where genes are annotated; the next continues with level5 to level2 GO terms; the final layer represents the root GO terms: biological process, molecular function, and cellular component, which are directly connected with the outcome. We calculate the node's important score in the best fitting fold. We select the top 10 node important score genes in the input layer, and the top 10 nodes important score GO terms in each level except the last level with 3 GO terms. GO terms with an important score less than 1e-10 in each layer are not shown in the figure. Nodes with darker colors have a larger important score. The transparent nodes represent the undisplayed nodes in each layer. Links with darker colors have larger edge weights.

In diabetes classification, we detected 10 genes which exist in microbe species *Lachnospira* (GUT_GENOME 001023), *Bacteroides thetaiotaomicron* (GUT_GENO ME001120), *Prevotella stercorea* (GUT_GENOME001282), *Acetatifactor* (GUT_GE NOME000166), *Eubacterium_F* (GUT_GENOME001241), and *Agathobaculum butyriciproducens* (GUT_GENOME 001016) have important roles. In these species, *Lachnospira*, *Bacteroides thetaiotaomicron*, and *Prevotella stercorea* were reported to be associated with diabetes Das et al. (2021); Gurung et al. (2020); Díaz-Perdigones et al. (2022). In these species, *Lachnospira* contains four important genes that contribute to negative regulation of RNA biosynthetic process, negative regulation of cellular macromolecule biosynthetic process, and purine-containing compound biosynthetic process.

In inflammatory bowel disease classification, we detected 10 genes which exist in microbe species *Dorea longicatena* (GUT_GENOME000149), *Roseburia sp003470905* (GUT_ GENOME001506), *Gemmiger qucibialis* (GUT_GENOME251083), *Faecalibacterium sp900539885* (GUT_GENOME209802), *Angelakisella sp004557855* (GUT_GENOME222558), *Anaerobutyricum hallii* (GUT_GENOME001689), and *Blautia massiliensis* (GU T_GENOME000676) have important roles. In these species, *Dorea longicatena* was reported to be associated with inflammatory bowel diseaseDiederen et al. (2020). In these species, *Gemmiger qucibialis* contains three important genes that contribute to the purine nucleoside bisphosphate metabolic process, ribose phosphate biosynthetic process, and purine nucleobase metabolic process. Most of the selected gene ontology terms with a high important score in inflammatory bowel disease come from the biological process.

3.3 The relationship between the input gene number and the GO_NN model performance

Determining the input gene number of GO informed model is a crucial task. Selecting a large geneset will increase the number of parameters in the model. Simultaneously, genes that existed in a few samples which are considered as noises may misestimate as important genes. On the other hand, a small geneset will exclude the genes associated with the disease. To find the effect of input gene number in prediction result, we prepare a gene set where the genes exist more than 1%, 5%, and 10% of UHGG samples as commonly existing genes, which we named as Dataset-large, Dataset-median, Dataset-small separately. The number of genes selected in each gene set is 22,927, 7,663, and 3,451, separately. We compared the predicting performance of GO-informed NN in different gene sets in diabetes and liver cirrhosis, which is shown in Table 4. The result shows that the GO-informed model has higher performance (F1score, AUC, and AUPRC) in larger geneset in both diseases. Noted that there is a small performance gap between large geneset and medium geneset, which indicated that further increasing the input gene number has a little improvement in predicting performance.

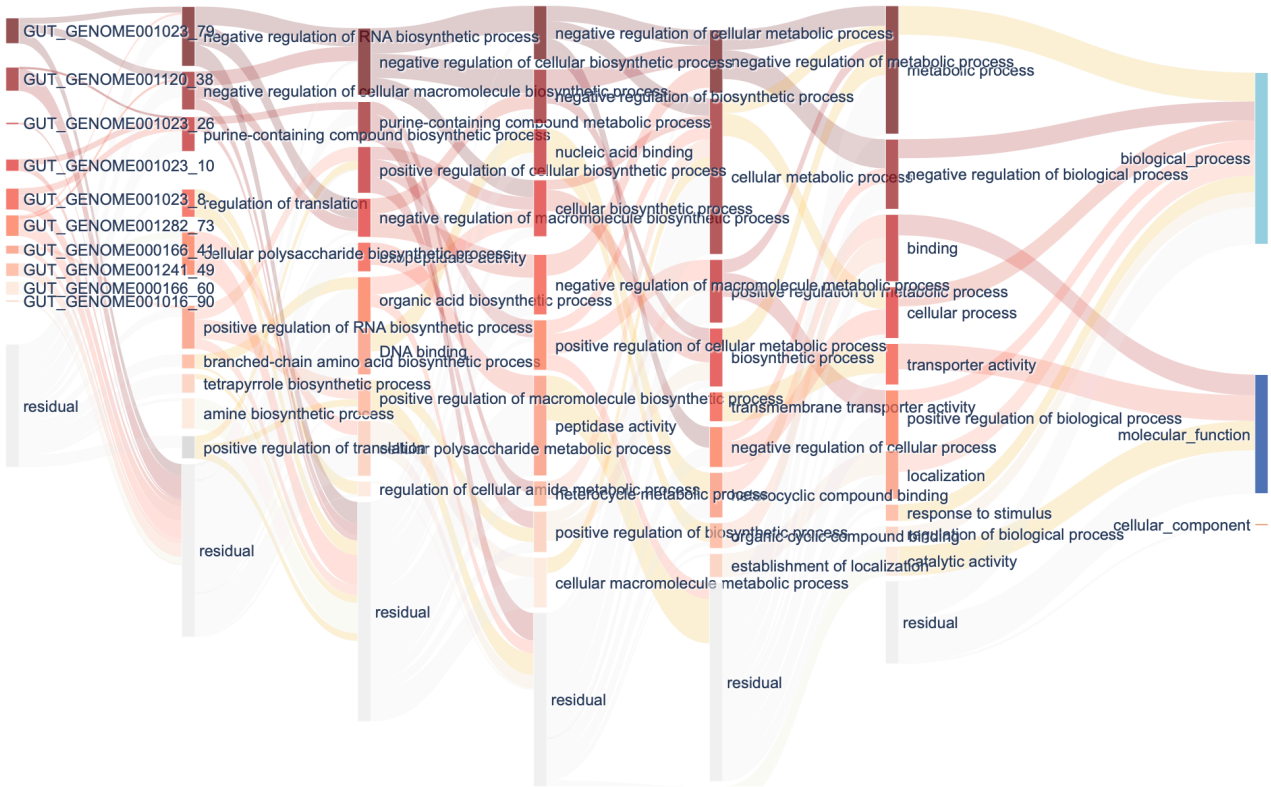


Figure 3. Interpretation of GO informed neural network in diabetes. The first layer shows the top 10 important genes. The following layers show the top 10 important GO terms in each level. The final layer shows the roots of each ontology. Nodes with darker colors have a larger important score. The transparent nodes represent the undisplayed nodes in each layer. Links with darker colors have larger edge weights.

3.4 Precision-recall curves comparison between GO_NN and RF

We noticed that the performance of random forest is competitive by comparing with other machine learning models in diseases such as liver cirrhosis. Therefore, we performed a further analysis by comparing GO_NN and RF results. The precision-recall curves comparison of GO_NN and RF in different diseases were shown in Fig. 5. From the precision-recall curves, there are less difference between the performance of two models in non-gastrointestinal disease including diabetes and liver cirrhosis datasets (Figure. 5.A, Figure. 5.B). and larger difference in gastrointestinal disease including inflammatory bowel disease and colorectal cancer (Figure. 5.C, Figure. 5.D). The overall performance of diabetes is lower than the other diseases, showing that machine learning models have difficulty in predicting diabetes. Limited information on the gene features results in difficulty in improving the performance, which is consistent with the previous study Qin et al. (2012); LaPierre et al. (2019).

4 DISCUSSION

The disease predicting result shows the effectiveness of the GO-informed neural network in predicting different diseases, especially gastrointestinal diseases. GO-NN gives a competitive result in different datasets and shows the importance of candidate genes and their functions. In addition, GO-informed neural network reduces the number of parameters for learning by utilizing genes GO annotation information and GO hierarchical structure. Compared with the non-GO-informed neural network, GO-informed neural network has fewer learnable parameters and overall disease prediction performance.

Figure 4. Interpretation of GO informed neural network in inflammatory bowel disease. The first layer shows the top 10 important genes. The following layers show the top 10 important GO terms in each level except the GO terms with an important score less than 1e-10. The final layer shows the roots of each ontology. Nodes with darker colors have a larger important score. The transparent nodes represent the undisplayed nodes in each layer. Links with darker colors have larger edge weights.

Furthermore, the visualization of GO informed model explains microbe functionality by integrating metagenome species, metagenome genes, and GO information. The network provides the functionality explanation of metagenome genes, which has the potential to discover novel species and functions that affect the disease. Specifically, GO informed model observes the important species not reported in previous research in both diabetes and liver cirrhosis datasets. These species have important functional roles in the disease, which cannot be discovered by a non-GO informed model.

Whereas GO informed model provides better performance than the non-GO-informed model, there are some issues with improving the performance of the GO-informed model. Firstly, we noted that the gene number affects the performance of GO informed model. Using a larger geneset helps improve the performance of GO informed model in both diseases. In addition, the sample size is still much smaller than the feature size. The GO-informed model performance may improve by using more qualified samples. Moreover, using heterogeneous data by combining GO with other biological priors, such as KEGG, may further guide model development and functional evaluation.

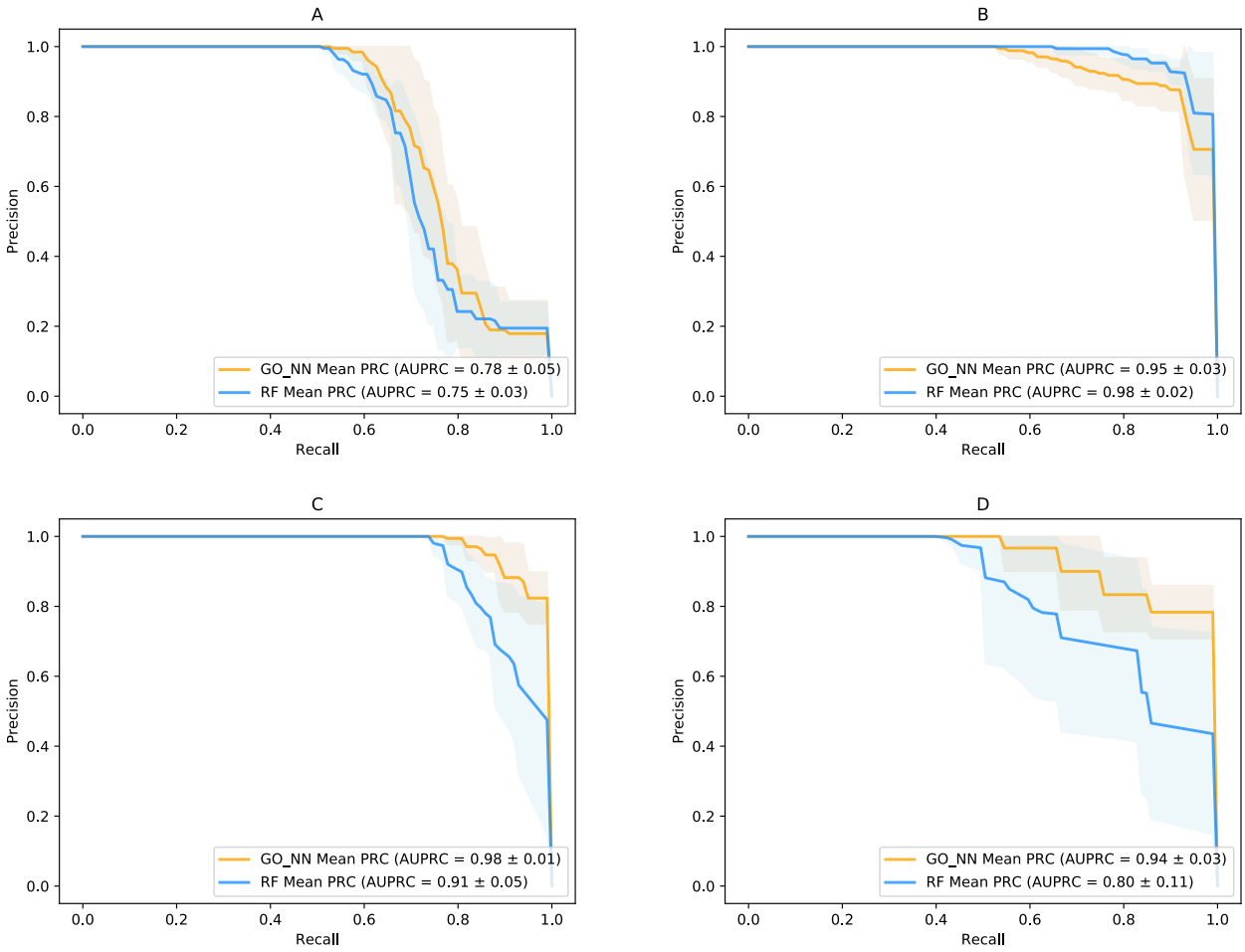


Figure 5. Precision-recall curves comparison of GO_NN and RF in different diseases. The solid line and the shadow represent the mean and standard deviation of 10-fold cross-validation results. **A** T2D precision-recall curves. **B** LC precision-recall curves. **C** IBD precision-recall curves. **D** CRC precision-recall curves.

5 CONCLUSIONS

In conclusion, we propose to utilize a GO-informed neural network to discover the microbe functionality in human diseases, which existing models cannot obtain. The GO-informed neural network model has effectiveness in disease prediction in diabetes and liver cirrhosis datasets. Our model discovered the important microbiome function, genes, and microbe species by calculating the important score of each gene and GO term in the network. We visualized the network's important genes and GO terms and provided insights into microbe contribution in functional aspects, which has the potential for clinical translation in disease-specified microbe-involved functions.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

YL obtained the necessary data, performed the experiments, and wrote the main manuscript. YZ and SI supervised the project. All authors reviewed the manuscript.

FUNDING

Not applicable.

ACKNOWLEDGMENTS

Not applicable.

SUPPLEMENTAL DATA

Not applicable.

DATA AVAILABILITY STATEMENT

GO informed neural network is available at https://github.com/YunjieLiu-HGC/GO_NN. The datasets generated and analysed during the current study are publicly available in the Science Data Bank repository at https://www.scidb.cn/anonymous/dkl2ZUlm.

REFERENCES

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* 39, 105–114. doi:10.1038/s41587-020-0603-3

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29. doi:10.1038/75556

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., and Gu, J. (2017). Afterqc: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18, 80. doi:10.1186/s12859-017-1469-3

Consortium, T. G. O. (2020a). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* 49, D325–D334. doi:10.1093/nar/gkaa1113

Consortium, T. U. (2020b). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* 49, D480–D489. doi:10.1093/nar/gkaa1100

Das, T., Jayasudha, R., Chakravarthy, S., Prashanthi, G. S., Bhargava, A., Tyagi, M., et al. (2021). Alterations in the gut bacterial microbiome in people with type 2 diabetes mellitus and diabetic retinopathy. *Scientific Reports* 11, 2738. doi:10.1038/s41598-021-82538-0

Diederen, K., Li, J. V., Donachie, G. E., de Meij, T. G., de Waart, D. R., Hakvoort, T. B. M., et al. (2020). Exclusive enteral nutrition mediates gut microbial and metabolic changes that are associated with remission in children with crohn's disease. *Scientific Reports* 10, 18879. doi:10.1038/s41598-020-75306-z

Díaz-Perdigones, C. M., Muñoz-Garach, A., Álvarez Bermúdez, M. D., Moreno-Indias, I., and Tinahones, F. J. (2022). Gut microbiota of patients with type 2 diabetes and gastrointestinal intolerance to metformin

differs in composition and functionality from tolerant patients. *Biomedicine Pharmacotherapy* 145, 112448. doi:https://doi.org/10.1016/j.biopha.2021.112448

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature* 598, 348–352. doi:10.1038/s41586-021-03922-4

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528, 262–266. doi:10.1038/nature15766

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* 4, 293–305. doi:10.1038/s41564-018-0306-4

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., et al. (2020). engRole of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51, 102590–102590. doi:10.1016/j.ebiom.2019.11.051. 31901868[pmid]

Huang, X., Huang, K., Johnson, T., Radovich, M., Zhang, J., Ma, J., et al. (2021). ParsVNN: parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways. *NAR Genomics and Bioinformatics* 3. doi:10.1093/nargab/lqab097. Lqab097

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28, 27–30. doi:10.1093/nar/28.1.27

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi:10.1038/nature12198

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods* 9, 357–359. doi:10.1038/nmeth.1923

LaPierre, N., Ju, C. J.-T., Zhou, G., and Wang, W. (2019). Metapheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* 166, 74–82. doi:https://doi.org/10.1016/j.ymeth.2019.03.003. Deep Learning in Bioinformatics

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546. doi:10.1038/nature12506

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi:10.1093/bioinformatics/btr011

Nguyen, T. H. (2019). Metagenome-based disease classification with deep learning and visualizations based on self-organizing maps. In *Future Data and Security Engineering*, eds. T. K. Dang, J. Küng, M. Takizawa, and S. H. Bui (Cham: Springer International Publishing), 307–319

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology* 12, 1–26. doi:10.1371/journal.pcbi.1004977

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14, 417–419. doi:10.1038/nmeth.4197

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi:10.1038/nature08821

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi:10.1038/nature11450

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi:10.1038/nature13568

Rahman, M. A. and Rangwala, H. (2018). Regmil: Phenotype classification from metagenomic data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (New York, NY, USA: Association for Computing Machinery), BCB '18, 145–154. doi:10.1145/3233547.3233585

Reiman, D., Metwally, A. A., Sun, J., and Dai, Y. (2020). Popphy-cnn: A phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE Journal of Biomedical and Health Informatics* 24, 2993–3001. doi:10.1109/JBHI.2020.2993761

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. *CoRR* abs/1704.02685

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12, 902–903. doi:10.1038/nmeth.3589

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). enPotential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 10, 766

Table 1. Summary of the datasets covered in this study.

Dataset	Case samples number	Control samples number	Citation
T2D	187	172	Qin et al. (2012)
Liver cirrhosis	162	145	Qin et al. (2014)
Inflammatory bowel disease	163	56	Franzosa et al. (2019)
Colorectal cancer	53	88	Zeller et al. (2014)

Table 2. Learnable parameters number in different neural network models.

GO-NN	AutoNN-h1	AutoNN-h2	AutoNN-h3
131,619	788,448,067	992,856,283	1,182,689,296

Table 3. Classification result in four different datasets

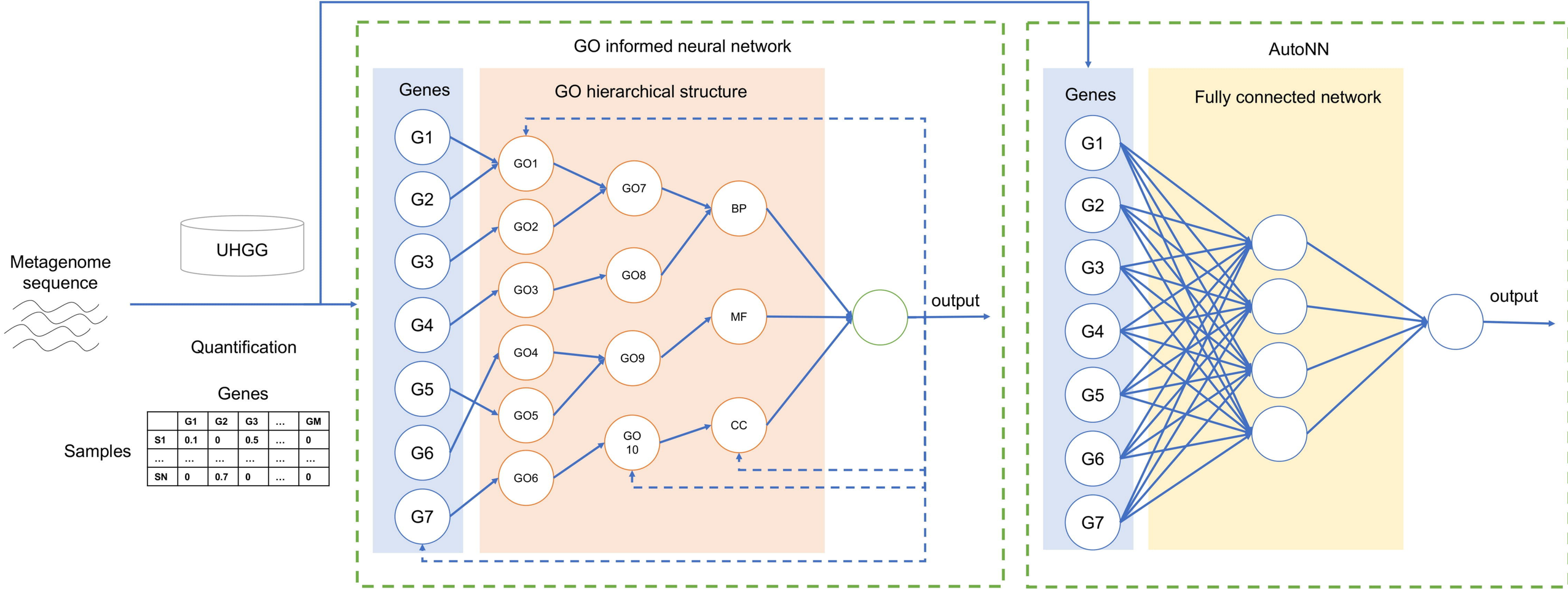
Dataset	Method	Accuracy	Precision	Reacll	F1Score	AUC	AUPRC
T2D	GO_NN	**0.708±0.055**	0.647±0.041	**0.958±0.042**	**0.772±0.040**	**0.778±0.051**	**0.776±0.048**
	AutoNN	0.565±0.068	0.578±0.069	0.547±0.120	0.558±0.091	0.604±0.065	0.637±0.062
	SVM	0.581±0.018	0.555±0.010	0.926±0.026	0.694±0.015	0.495±0.023	0.552±0.010
	RF	0.681±0.036	**0.698±0.033**	0.668±0.062	0.682±0.042	0.733±0.022	0.751±0.031
	LR	0.511±0.056	0.522±0.051	0.600±0.071	0.557±0.053	0.545±0.058	0.521±0.033
LC	GO_NN	0.884±0.044	0.912±0.066	0.871±0.037	0.890±0.040	0.931±0.038	0.952±0.028
	AutoNN	0.791±0.059	0.815±0.069	0.788±0.060	0.800±0.056	0.862±0.062	0.868±0.064
	SVM	0.750±0.050	0.742±0.043	0.812±0.051	0.775±0.045	0.849±0.041	0.704±0.048
	RF	**0.934±0.033**	**0.958±0.037**	**0.918±0.039**	**0.937±0.032**	**0.974±0.013**	**0.978±0.015**
	LR	0.822±0.044	0.826±0.050	0.847±0.084	0.834±0.048	0.901±0.029	0.781±0.048
IBD	GO_NN	**0.791±0.018**	0.780±0.016	**1.000±0.000**	**0.876±0.010**	**0.936±0.037**	**0.979±0.013**
	AutoNN	0.781±0.086	0.809±0.089	0.917±0.093	0.853±0.060	0.821±0.057	0.910±0.069
	SVM	0.770±0.020	**0.817±0.020**	0.888±0.018	0.851±0.012	0.728±0.021	0.808±0.018
	RF	0.783±0.051	0.804±0.035	0.935±0.041	0.864±0.032	0.806±0.083	0.915±0.046
	LR	0.687±0.033	0.771±0.032	0.824±0.000	0.796±0.017	0.640±0.045	0.765±0.027
CRC	GO_NN	0.873±0.066	0.892±0.148	**0.817±0.095**	**0.841±0.069**	**0.939±0.030**	**0.937±0.027**
	AutoNN	0.787±0.065	0.888±0.139	0.550±0.150	0.665±0.122	0.874±0.060	0.849±0.075
	SVM	**0.887±0.043**	**0.980±0.060**	0.733±0.082	0.836±0.063	0.893±0.031	0.827±0.071
	RF	0.780±0.123	0.824±0.193	0.567±0.249	0.648±0.224	0.844±0.086	0.801±0.112
	LR	0.767±0.061	0.813±0.128	0.550±0.107	0.651±0.102	0.852±0.038	0.633±0.099

Comparison of the predicting performance of GO-informed NN (GO_NN), non-GO-informed NN (AutoNN, LaPierre et al.), and other machine learning models in different diseases. The mean and the standard deviation are recorded for different evaluation metrics after a 10-fold cross-validation. The best performances of each metric are highlighted in bold separately.

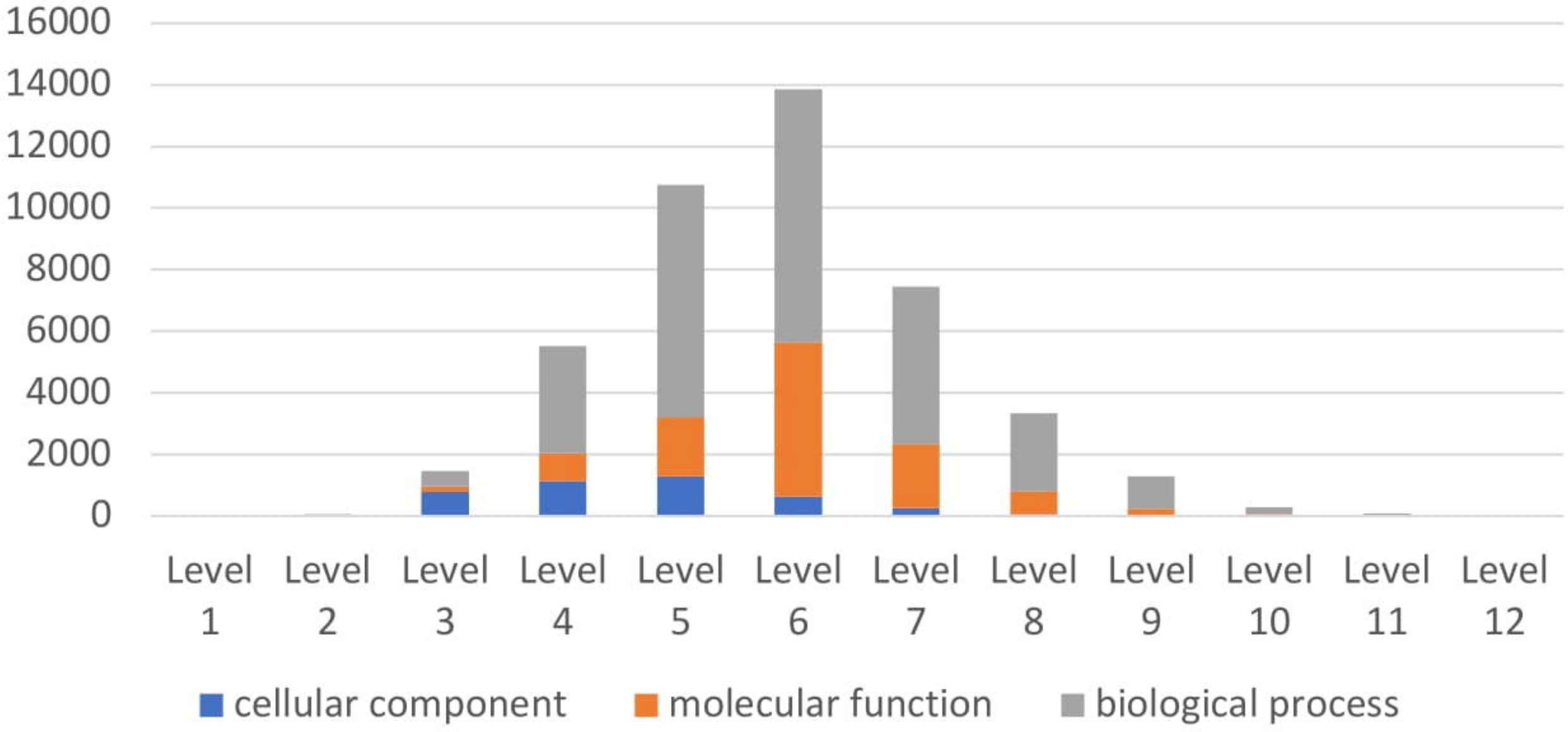
Table 4. Classification result in different genesets

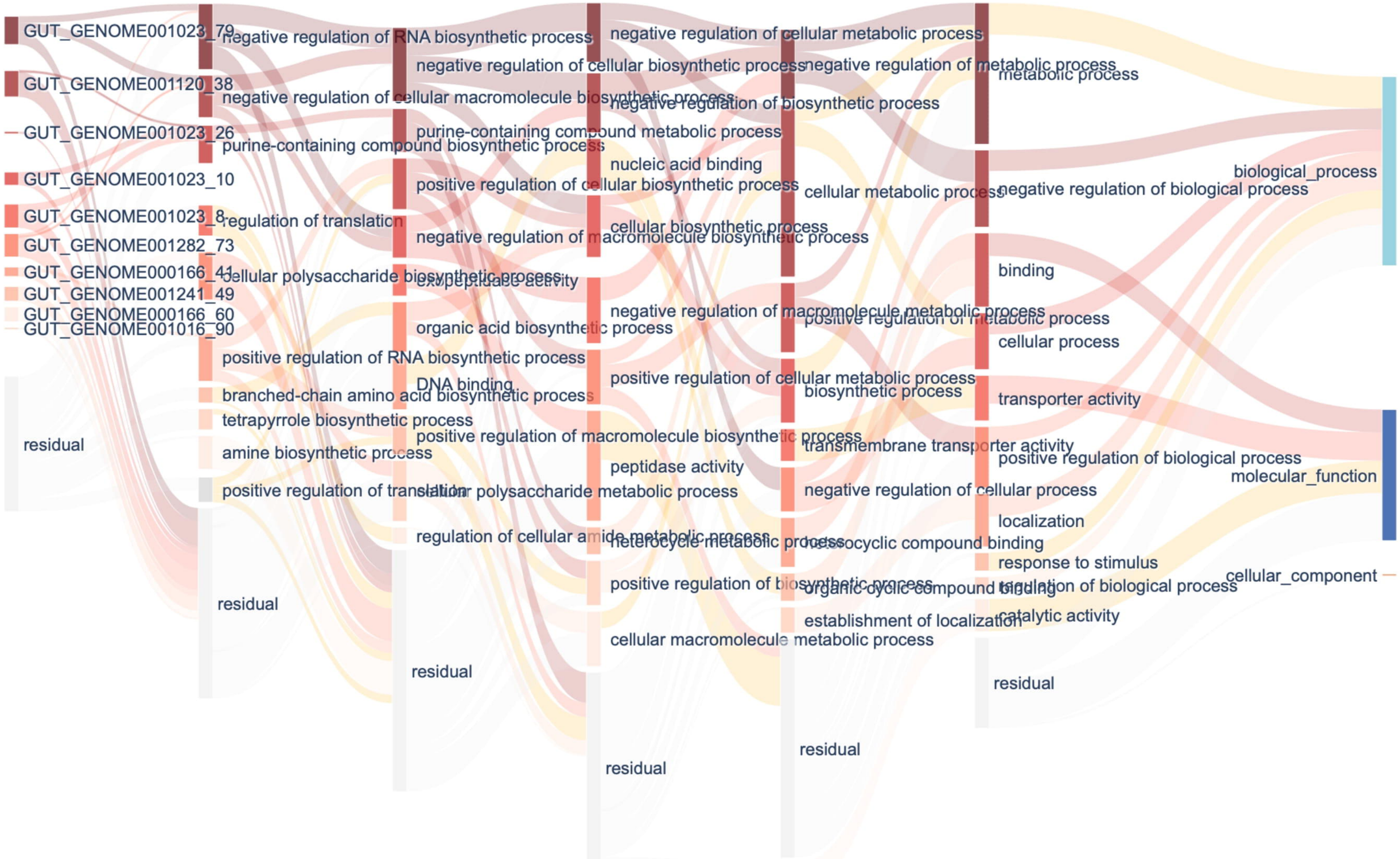
Dataset	Accuracy	Precision	Recall	F1score	AUC	AUPRC
T2D-Large	**0.708±0.055**	**0.647±0.041**	**0.958±0.042**	**0.772±0.040**	**0.778±0.051**	**0.776±0.048**
T2D-Medium	0.667±0.067	0.631±0.054	0.858±0.079	0.726±0.055	0.708±0.044	0.708±0.060
T2D-Small	0.562±0.082	0.557±0.066	0.795±0.063	0.652±0.053	0.639±0.061	0.709±0.054
LC-Large	**0.884±0.044**	**0.912±0.066**	0.871±0.037	**0.890±0.040**	**0.931±0.038**	**0.952±0.028**
LC-Medium	**0.884±0.021**	0.905±0.037	**0.876±0.019**	**0.890±0.018**	0.924±0.028	0.943±0.028
LC-Small	0.794±0.051	0.821±0.076	0.794±0.042	0.805±0.039	0.857±0.047	0.888±0.047
IBD-Large	**0.791±0.018**	0.780±0.016	**1.000±0.000**	**0.876±0.010**	**0.936±0.037**	**0.979±0.013**
IBD-Medium	0.778±0.038	**0.794±0.030**	0.947±0.043	0.863±0.024	0.807±0.042	0.931±0.019
IBD-Small	0.770±0.054	0.792±0.034	0.935±0.052	0.857±0.035	0.795±0.070	0.927±0.029
CRC-Large	**0.873±0.066**	**0.892±0.148**	**0.817±0.095**	**0.841±0.069**	**0.939±0.030**	**0.937±0.027**
CRC-Medium	0.760±0.072	0.714±0.132	0.733±0.225	0.695±0.124	0.907±0.050	0.884±0.057
CRC-Small	0.700±0.085	0.650±0.158	0.717±0.273	0.637±0.128	0.887±0.068	0.844±0.100

Comparation of the predicting performance of GO informed NN in different gene sets in diabetes (T2D), liver cirrhosis (LC), inflammatory bowel disease (IBD), and colorectal cancer (CRC). The mean and the standard deviation are recorded for different evaluation metrics after a 10-fold cross-validation. The best performances of each metric are highlighted in bold separately.

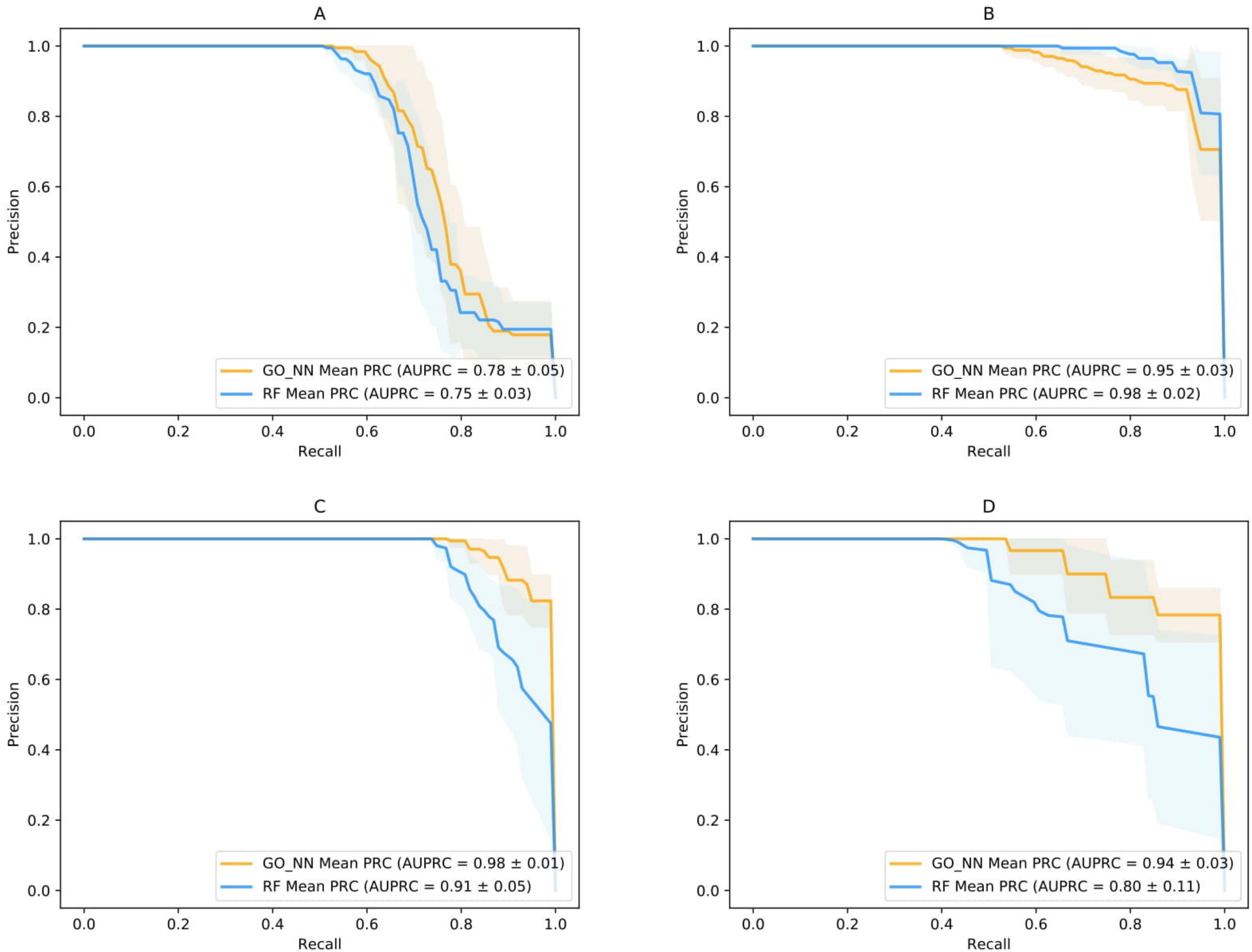


	G1	G2	G3	…	GM
S1	0.1	0	0.5	…	0
…	…	…	…	…	…
SN	0	0.7	0	…	0


16000
14000
12000
10000
8000
6000
4000
2000
0
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Level 7
Level 8
Level 9
Level 10
Level 11
Level 12
cellular component
molecular function
biological process


GUT_GENOME001023_79
GUT_GENOME001120_38
GUT_GENOME001023_26
GUT_GENOME001023_10
GUT_GENOME001023_8
GUT_GENOME001282_73
GUT_GENOME000166_41
GUT_GENOME001241_49
GUT_GENOME000166_60
GUT_GENOME001016_90
residual
negative regulation of RNA biosynthetic process
negative regulation of cellular macromolecule biosynthetic process
purine-containing compound biosynthetic process
regulation of translation
cellular polysaccharide biosynthetic process
positive regulation of RNA biosynthetic process
branched-chain amino acid biosynthetic process
tetrapyrrole biosynthetic process
amine biosynthetic process
positive regulation of translation
residual
negative regulation of cellular biosynthetic process
purine-containing compound metabolic process
nucleic acid binding
positive regulation of cellular biosynthetic process
negative regulation of macromolecule biosynthetic process
exopeptidase activity
organic acid biosynthetic process
DNA binding
peptidase activity
cellular polysaccharide metabolic process
regulation of cellular amide metabolic process
positive regulation of biosynthetic process
cellular macromolecule metabolic process
residual
negative regulation of cellular metabolic process
negative regulation of biosynthetic process
cellular metabolic process
cellular biosynthetic process
negative regulation of macromolecule metabolic process
positive regulation of cellular metabolic process
biosynthetic process
positive regulation of macromolecule biosynthetic process
heterocycle metabolic process
organic cyclic compound binding
establishment of localization
residual
negative regulation of metabolic process
metabolic process
negative regulation of biological process
binding
positive regulation of metabolic process
cellular process
transporter activity
transmembrane transporter activity
positive regulation of biological process
negative regulation of cellular process
localization
heterocyclic compound binding
response to stimulus
regulation of biological process
catalytic activity
residual
biological_process
molecular_function
cellular_component

GUT_GENOME000149_3
GUT_GENOME001506_71
GUT_GENOME251083_40
GUT_GENOME251083_51
GUT_GENOME209802_19
GUT_GENOME251083_214
GUT_GENOME143522_65
GUT_GENOME222558_46
GUT_GENOME001689_24
GUT_GENOME000676_3
residual
tetrapyrrole biosynthetic process
purine nucleoside bisphosphate metabolic process
branched-chain amino acid biosynthetic process
purine nucleoside metabolic process
ribose phosphate biosynthetic process
purine nucleobase metabolic process
nucleotide-sugar biosynthetic process
porphyrin-containing compound metabolic process
alpha-amino acid biosynthetic process
cobalamin metabolic process
residual
purine-containing compound metabolic process
tetrapyrrole metabolic process
carbohydrate derivative biosynthetic process
organic acid biosynthetic process
organophosphate biosynthetic process
residual
organic substance metabolic process
organonitrogen compound metabolic process
organic substance biosynthetic process
residual
residual
nitrogen compound metabolic process
biosynthetic process
residual
metabolic process
residual
biological_process
cellular_component
molecular_function


A
Precision
Recall
GO_NN Mean PRC (AUPRC = 0.78 ± 0.05)
RF Mean PRC (AUPRC = 0.75 ± 0.03)
B
Precision
Recall
GO_NN Mean PRC (AUPRC = 0.95 ± 0.03)
RF Mean PRC (AUPRC = 0.98 ± 0.02)
C
Precision
Recall
GO_NN Mean PRC (AUPRC = 0.98 ± 0.01)
RF Mean PRC (AUPRC = 0.91 ± 0.05)
D
Precision
Recall
GO_NN Mean PRC (AUPRC = 0.94 ± 0.03)
RF Mean PRC (AUPRC = 0.80 ± 0.11)