GDmicro
classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data
Liao, Herui; Shang, Jiayu; Sun, Yanni

Published in:
Bioinformatics

Published: 01/12/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1093/bioinformatics/btad747

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Genome analysis

GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data

Herui Liao, Jiayu Shang, Yanni Sun

Abstract

Motivation: With advances in metagenomic sequencing technologies, there are accumulating studies revealing the associations between the human gut microbiome and many human diseases. These associations shed light on using gut microbiome data to distinguish case and control samples of a specific disease, which is also called host disease status classification. Importantly, using learning-based models to distinguish the disease and control samples is expected to identify important biomarkers more accurately than abundance-based statistical analysis. However, available tools have not fully addressed two challenges associated with this task: limited labeled microbiome data and decreased accuracy in cross-studies. The confounding factors, such as the diet, technical biases in sample collection/sequencing across different studies/cohorts often jeopardize the generalization of the learning model.

Results: To address these challenges, we develop a new tool GDmicro, which combines semi-supervised learning and domain adaptation to achieve a more generalized model using limited labeled samples. We evaluated GDmicro on human gut microbiome data from 11 cohorts covering 5 different diseases. The results show that GDmicro has better performance and robustness than state-of-the-art tools. In particular, it improves the AUC from 0.783 to 0.949 in identifying inflammatory bowel disease. Furthermore, GDmicro can identify potential biomarkers with greater accuracy than abundance-based statistical analysis methods. It also reveals the contribution of these biomarkers to the host’s disease status.

Availability and implementation: https://github.com/liaoherui/GDmicro.

1 Introduction

In recent years, many studies have shown strong associations between the human gut microbiome and several human diseases (Yu et al. 2017, Kwong et al. 2018, Gomaa 2020). For example, a meta-analysis of large-scale metagenomic samples shows that dozens of specific bacteria are enriched in colorectal cancer (CRC) patients across different countries (Wirbel et al. 2019). Another microbiome-related study found a reduced complexity of the bacterial phylum Firmicutes in inflammatory bowel disease (IBD) patients (Manichan et al. 2006). With the in-depth study of Firmicutes, anti-inflammatory properties of many species under this phylum have been revealed, implying their potential utilities in promoting gut health. These observations indicate that the composition of gut microbes may provide important information for distinguishing case and control samples of a particular disease. Given the accumulating evidence on the associations between diseases and the human gut microbiome, there is a need to develop a more accurate microbiome-based host disease status classification model. Such a model has the potential to identify more informative biomarkers for downstream analysis than abundance-based statistical analysis (Curry et al. 2021). The goal of this study is to create a more accurate host disease status classification model using data from the human gut microbiome and incorporating advanced functions for biomarker discovery.

Many computational methods have been developed to classify host disease status based on the human gut microbiome data. The abundance of gut microbes is a major feature used by these tools (Curry et al. 2021). Given the gut microbial composition abundance data, these methods apply traditional machine learning or deep-learning methods to distinguish case and control samples of a specific disease. The tools utilizing traditional machine learning include MetAML (Pasolli et al. 2016) and SIAMCAT (Wirbel et al. 2021). MetAML takes gut microbial compositional data as input and then applies random forest and support vector machines to classify host disease status. To improve the classification performance, it combines k-fold cross-validation and the grid search strategy to tune the best hyperparameters for models. SIAMCAT (Wirbel et al. 2021) is a toolbox that relies on various regression models, such as ridge regression, to classify host disease status using gut microbial compositional data.
In addition to traditional machine-learning methods, deep-
learning models have shown promising results in classifying
host disease status (Oh and Zhang 2020, Reiman et al. 2020).
One relevant work is DeepMicro (Oh and Zhang 2020),
which applies an autoencoder model to learn the deep repre-
sentation of input gut microbial compositional features.
Then, the multi-layer perceptron model is employed to clas-
sify host disease status with the learned latent features.
PopPhy-CNN (Reiman et al. 2020) is another popular deep-
learning-based tool. It takes gut microbial compositional data
and a phylogenetic tree as input and utilizes a novel convolu-
tional neural network learning framework for host disease
status classification.

Despite the promising results, existing tools have not
addressed two challenges well. The first challenge is the lim-
ited number of labeled data. Although there has been rapid
growth in microbiome data collection, a large number of sam-
ple still lack detailed metadata annotations. According to a
recent study, only 7.8% of the 444 829 human microbiome
samples had explicit information on host disease status
(Abdill et al. 2022), partly because of the high cost of obtain-
ing labels for microbiome data (Shen et al. 2022). As a result,
limited labeled data pose challenges for most learning models.
Second, many available methods ignore the domain discrep-
ency problem. For example, in our k-fold cross-validation ex-
periment on CRC, DeepMicro achieves a 0.803 area under
the curve (AUC) on the CRC-FR dataset. However, when ap-
piled to CRC-US dataset in the cross-study experiment, its
performance drops to 0.609. Specifically, data from different
studies have many differences due to confounding factors,
such as region, ethnicity, and diet, which can lead to changes
in the gut microbiome (He et al. 2018, Wirbel et al. 2019).
Thus, the domain difference between training and test data
can greatly affect the robustness of the classification model,
ultimately leading to poor performance in real applications.

1.1 Overview of our method
Given the limited labeled data and the rapid accumulation of
unlabeled data, we formulate the host disease status classifica-
tion problem as a semi-supervised learning task, which uses
both labeled and unlabeled data for feature learning (van
Engelen and Hoos 2020). It should be noted that, to prevent
test data leakage, classical semi-supervised learning methods,
such as the graph convolutional network (GCN), only utilize
the labeled training samples to optimize the loss function in
the training process (Zhu et al. 2003, Kipf and Welling 2017,
Zhou et al. 2020). Consequently, these methods exhibit im-
poved performance while effectively avoiding any data leak-
age concerns in many tasks, including text classification (Yao
et al. 2019), image classification (Gao et al. 2021), and pro-
tein function prediction (Gligorijevic et al. 2021). Here, we
present GDmicro, a GCN-based model to utilize the informa-
tion from both labeled and unlabeled samples for learning
and classification. First, an inter-host microbiome similarity
graph is built using the gut microbial compositional data. In
our graph, nodes represent the compositional abundance fea-
tures of the hosts’ microbiome (Fig. 1I), and edges represent
the similarity of learned latent features between two hosts’
microbiome (Fig. III and III). Then, GCN can take this graph
as input and incorporate the structural and node abundance
features for disease status classification. Because both labeled
and unlabeled samples can be integrated into the graph, graph
convolution in GCN can utilize local topology for feature
passing between two types of samples. Second, to overcome
the domain discrepancy problem, we apply a deep adapta-
tion network to learn transferable latent features from the micro-
bial compositional matrix across different domains (Fig. III).
We validated GDmicro on 10 cohorts covering 5 diseases and
compared GDmicro with alternative tools for disease status
classification. The results show that GDmicro has consistently
higher AUC than other methods. In addition, GDmicro allows
users to detect the most important species to disease status
classification and explore how these species affect the hosts’
disease status, which provides important information for bio-
marker discovery.

2 Materials and methods
We choose GCN (Kipf and Welling 2017) as the semi-
supervised learning engine for our host disease status classifi-
cation problem, which uses both labeled and unlabeled data for
feature learning. GCN has had several successful applica-
tions in identifying gene–gene interaction, disease–drug rela-
tionship, and interactions between phages and bacteria (Han
et al. 2019, Shang et al. 2021, Yu et al. 2021). In the host dis-
ease status classification problem, GCN has two distinct
advantages. First, it can utilize information from unlabeled
samples during graph convolution. Second, the GCN model
systematically integrates information from the nodes and
edges, which represent the abundance distribution of species
and their similarities between hosts. Then, we apply the deep
adaptation network to mitigate the impact of domain-specific
confounding factors on feature learning, leading to a more ro-
bus graph for microbiome data from different studies.

In the following sections, we will first introduce how we en-
code and construct the inter-host microbiome similarity graph
with the gut microbiome data and deep adaptation network.
Then, we will describe the GCN model optimized for disease
identification and its application in biomarker discovery.

2.1 Inter-host microbiome similarity graph G
Increasing studies show that people with sclerosis, obesity,
and IBDs have similar gut microbial compositional abund-
ance (Chen et al. 2016, Lloyd-Price et al. 2019, Palmas et
al. 2021). One recent work constructed a patient relationship
graph based on the similarity of human multi-omics data for
disease diagnosis (Li et al. 2022). Inspired by these studies, we
construct an inter-host microbiome similarity graph G to in-
corporate microbial compositional similarity and composi-
tional abundance features. The major components of the
graph construction are sketched in Fig. 1I–III. Each node in
the graph represents a human microbial sample, and the edges
represent the similarity between the microbiome data. The node
features are the compositional abundance vectors of the sam-
ples. Because the compositional abundance data might contain
skewness or bias in some features, we apply log10-
transformation and z-score normalization on the features rather
than using the raw data. Specifically, the mean and standard de-
viation are calculated using the training data, and these values
are subsequently employed to normalize the test data following
log10-transformation. This standardized normalization ap-
proach is applied to all experiments. As a result, each node is
encoded with the normalized species abundance vectors.

As mentioned above, similar gut microbial compositional
data can indicate similar human disease status. Thus, we use
microbial compositional similarity to define the edge between
two nodes. Because there is no clear cutoff to determine whether two samples are significantly similar, we employ the k-nearest neighbor algorithm to generate the topological structure. Specifically, a node will be connected with its k closest nodes in G. Below we will detail how to compute the compositional similarity between different samples (Fig. III and III).

2.1.1 Using deep adaptation network to learn transferable features
To learn the most relevant features for edge construction from different studies, we applied a deep adaptation network (Long et al. 2015) to learn transferable latent features from compositional abundance matrices of training and test datasets.

The deep adaptation network is initially designed to solve the domain adaptation problem in the field of image processing (Long et al. 2015). The main idea behind this is to utilize the multiple kernel variants of maximum mean discrepancies (MK-MMD) (Gretton et al. 2012) to measure the difference between the source and the target domain and minimize the domain discrepancies during training. To apply MK-MMD to microbiome data, we added an MK-MMD-based adaptation regularizer to the loss function of a multi-layer fully connected network (Fig. III). Denote x as the compositional abundance vector of one sample, y as the disease status label of the sample, $D_s = \{ (x_{s,i}, y_{s,i}) \}_{i=1}^{n_s}$ as the source domain with $n_s$ labeled samples and $D_t = \{ (x_{t,i}) \}_{i=1}^{n_t}$ as the target domain with $n_t$ unlabeled samples. Then, the loss function of the network can be defined as:

$$L(\theta(x_{s,i}^l), y_{s,i}^l) + \lambda d^2_s(D_s^l, D_t^l),$$

where $L$ is the cross-entropy loss function for the disease status classification task, $\theta(x_{s,i}^l)$ is the conditional probability that the fully connected network in Fig. III assigns $x_{s,i}^l$ to label $y_{s,i}^l$, $\lambda > 0$ is a penalty parameter, and $l$ is the hidden layer index where the regularizer is effective. Additional details about the domain adaptation regularizer $d^2_s$ can be found in Supplementary Section S1.1. As a result, by minimizing the loss function with the MK-MMD-based adaptation regularizer, the model can learn the transferable latent features between data from different domains.

After training the deep adaptation network, we applied the trained model to convert input microbial compositional features to latent features from the fully connected network. The equation for the conversion is listed in Equation (2).

$$b^{(l+1)} = ReLU(b^{(l)} W^{(l)} + b^{(l)}), \; l \in \{0, 1, 2\},$$

where $b^{(l)}$ is the latent features captured from the $l$th hidden layer of the model, and $b^{(0)} = x$. $W^{(l)}$ and $b^{(l)}$ are the learnable weights and biases of the hidden layers. ReLU is the activation function. Finally, each input sample has corresponding latent features $b^{(2)}$ that contain robust statistic patterns from compositional abundance data.

2.1.2 Edge construction
Given the latent features of all samples, a sample–sample Euclidean distance matrix can be calculated. This matrix effectively captures the microbial compositional similarity across all samples. Then, we employ the k-nearest neighbor algorithm to determine whether two samples have a connection. In our algorithm, we will connect the k closest samples for each sample according to the distance matrix.

2.2 The GCN model
After constructing the graph, we train a GCN model to classify disease status for input samples. The most important component of GCN is the graph convolution layer, which can take advantage of the topological structure for feature...
learning. Because both labeled and unlabeled samples are connected in the graph, GCN can also utilize node features from unlabeled nodes. If there are $n$ samples in the graph, the graph convolution layer is defined as:

$$H^{(l+1)} = ReLU(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}),$$

(3)

$$\tilde{A} = A + I_n,$$

(4)

where $A$ is the $R^{n \times n}$ adjacency matrix of the inter-host microbiome similarity graph, $I_n$ is an $R^{n \times n}$ identity matrix, and $D$ is the $R^{n \times n}$ degree matrix of $\tilde{A}$. $H^{(0)}$ refers to the output of the first hidden layer and $H^{(l)}$ is the node feature matrix. $W^{(l)}$ represents the weight matrix of the hidden layers. As shown in Equation (3), in each convolution layer, the model considers the 1-hop neighborhood of the nodes to calculate new node features. Then, the successive convolution will be applied in the $l$ layers, and thus, the model can learn features based on the topological structure to enlarge the receptive field from unlabeled nodes and make a classification. All the nodes will take part in the convolution as shown in Equation (3). But only the labeled nodes are used in maximizing the loss function, which is a cross-entropy loss function for the disease status classification. This operation ensures that the GCN model does not incorporate the label information of test samples into its training process. Finally, the model will assign disease status to unlabeled samples in the classification step.

### 2.3 Discover biomarkers and analyze their contribution to the host disease status with GDmicro

In microbiome-related studies, biomarkers typically refer to species that are highly correlated with diseases and are often considered signatures of certain health conditions (Wirbel et al. 2019). Biomarker discovery is an essential task that can help reveal the underlying biological mechanisms of diseases. However, identifying biomarkers using deep-learning models remains challenging. Although methods, such as the ablation approach, are available for interpreting deep-learning models (Wang et al. 2021), they cannot be employed in this study due to computational efficiency and performance issues. Specifically, the sheer number of features (over 800 in CRC datasets) makes these methods computationally expensive. In addition, in our specific case, methods like the ablation approach, which requires the removal of one feature from hundreds, cannot produce noticeable differences to the model, making feature selection highly difficult. To tackle these issues, we propose a straightforward yet highly effective performance-based method that utilizes both the input data and the GCN model to extract significant species. Furthermore, based on the identified biomarkers, we design a function that utilizes the GCN model to analyze the biomarkers' contributions to the host disease status. We will discuss these two methods in the following paragraphs.

First, we utilize the node features in the GCN model for biomarker discovery. Once the graph is given, the performance of the GCN can vary when feeding it with different species. Therefore, we run GDmicro multiple times using a single species as the node feature each time. As a result, each species obtains a corresponding AUC on the training data, allowing for species ranking according to their respective AUCs. Higher-ranked species can be crucial biomarkers for classifying host disease status.

Second, given the identified biomarkers, we will further characterize the contribution of these biomarkers to the hosts’ disease status. To solve this problem, we first define four kinds of contributions of biomarkers to the disease status: “Increase2Disease,” “Increase2Health,” “Decrease2Disease,” and “Decrease2Health.” There are four kinds of contributions because the microbes can have diverse effects on the disease status of the host. For instance, potential biomarkers usually exhibit a consistent change in abundance across patient samples in multiple studies, while other microbes demonstrate inconsistent alterations in abundance across different studies (Pittayanon et al. 2020). Consequently, by comparing the scores of these four contributions, users can comprehensively investigate the impact of specific microbes on the disease status of the host. These contributions reflect that modifying the abundance of certain species, either through increase or decrease, may result in a sample being more likely classified as diseased or healthy, respectively. For example, “Increase2Disease” represents that increasing the raw abundance of the species will result in a higher probability of the sample being classified as a diseased one. Thus, we can analyze a biomarker’s contribution by changing its raw abundance and checking the class (health or disease) probability predicted by the model. Based on the idea, we design a function utilizing the output probability of the GCN model to calculate the score for each kind of contribution. The higher the score, the more likely the biomarker has that specific contribution to the host’s disease status. Figure 2 illustrates the workflow of this function.

As shown in Fig. 2, the function initiates by recording the classification probability for each sample. Then, the program modifies the raw abundance of an identified biomarker in a sample and records the new classification probability under the modified abundance. Six modification rules are applied: “Max2Median,” “Max2Min,” “Middle2Max,” “Middle2Min,” “Min2Max,” and “Min2Median.” In these rules, “Middle” represents all in-between abundance in addition to the maximum and minimum abundance. Each rule represents the method of altering the raw abundance of the identified biomarker. For example, “Max2Median” indicates that the abundance of the selected biomarker is the maximum value among all species in the given sample, and will be adjusted from the maximum to the median value. With the original classification probabilities and those after applying various abundance modification rules, the function will compute the contribution value for each biomarker. The contribution value is calculated as the absolute difference between two probabilities: the probability of being healthy before and after modifying the abundance. This value reflects the biomarker’s contribution to disease status in the given sample. Lastly, the function calculates the contribution values of the identified biomarkers for all samples. Based on these values, the function determines the contribution score of one specific biomarker to the disease status of a host. The formula is $\sum_{i=1}^{n} x \cdot C_{i,x}$, where $n$ represents correctly predicted samples in the training data, $x$ is the specific sample, $i$ is the specific contribution type, $j$ is the identified marker, and $C_{i,j,x}$ is the contribution value.

### 3 Results

In this section, we describe GDmicro’s results on 10-fold cross-validation and cross-study experiments in 10 cohorts.
GDmicro

Figure 2. One example of using GDmicro to analyze a biomarker’s contribution to the host disease status. “s1, s2, s3:” identified biomarkers. “Rule:” the rule to change the raw abundance of the analyzed biomarker. For example, “Max2Median” means the raw abundance of the biomarker will change from the maximum value to the median value. “Contribution value:” the absolute value of the difference between the originally predicted probability of being healthy and the predicted probability of being healthy after changing the abundance.

with 5 diseases. In order to mimic real-world scenarios, we conducted these experiments using two different strategies for GDmicro. The first strategy, named “batch_run,” applies GDmicro to predict the host disease status for all the samples in the cohort altogether. In contrast, the second strategy, referred to as “single_run,” applies GDmicro to predict the host disease status for each individual separately, requiring multiple runs. Then, we analyzed the effect of different model architectures and parameters on the performance of GDmicro. Finally, we explored the ability of GDmicro to identify biomarkers for CRC and IBD cohorts and explore their influence on hosts’ disease status. We benchmarked four state-of-art tools, including SIAMCAT (Wirbel et al. 2021), MetAML (Pasolli et al. 2016), DeepMicro (Oh and Zhang 2020), and PopPhy-CNN (Reiman et al. 2020) with GDmicro in the 10-fold cross-validation and cross-study experiments. We take microbial compositional data as input for all of these tools. All tools are run with the recommended parameters or the best performance parameters reported in the corresponding paper.

3.1 Datasets and evaluation metrics

To benchmark with other tools (Pasolli et al. 2016, Oh and Zhang 2020, Rahman and Rangwala 2020, Reiman et al. 2020), we first used the same set of five cohorts covering five diseases: liver cirrhosis (Cirrhosis), CRC, IBD, obesity (Obesity), and type 2 diabetes (T2D). We then collected four CRC cohorts and one IBD cohort for the cross-study experiment. Finally, we explored one additional CRC cohort, which is never used in any training or validation process, to evaluate the performance of GDmicro and other tools in real-world scenarios. These cohorts contain 1702 samples in total and have composition profiles available, which are generated by Metaphlan2 (Truong et al. 2015) or mOTU2 (Milanese et al. 2019). We downloaded five CRC cohorts from Wirbel et al. (2019) and all other cohorts from curatedMetagenomicData package (Pasolli et al. 2017). The detailed information of all cohorts is summarized in Table 1.

Like many previous studies (Pasolli et al. 2016, Lo and Marculescu 2019, Nguyen and Zucker 2019, Zhu et al. 2019), we selected the AUC as the evaluation metric in this work, which summarizes the true positive and false positive rates and has a robust evaluation for the unequal ratio of each outcome. For the 10-fold cross-validation experiment, each experiment is repeated 10 times. Then, the margin of error for the mean of all experiments is calculated with a 95% confidence interval, and it is defined as:

$$ME = t_{1-z/2,n-1} \times \frac{s_r}{\sqrt{n}},$$  (5)

where $1 - \alpha$ is the significance level, $n$ is the sample size, $t_{1-z/2,n-1}$ refers to critical value of t-distribution with degrees of freedom $n - 1$ for an z/2 area of for the upper tail, and $s_r$ is the sample standard deviation.

3.2 Ten-fold cross-validation experiment

To evaluate the performance of GDmicro, we applied all tools to classify disease status for samples from five cohorts (upper part in Table 1). We conducted 10-fold cross-validation using StratifiedKFold ($k = 10$) function in the sklearn package. The mean AUC and the margin or errors of all tools were recorded in Table 2. GDmicro with the “batch_run” strategy achieved the best performance in all tested cohorts. In particular, GDmicro-batch achieved 0.923 and 0.936 AUC in the CRC-FR and IBD-DK cohorts, which are 6.4% and 6.3% improvements over the second-ranked tool, respectively. In addition, while GDmicro-single does not match the performance of GDmicro-batch, it still outperforms the other tested methods. We further investigated the impact of the number of test samples on the performance, revealing that the model’s performance improves as the number of test samples increases (Supplementary Section S2.1 and Supplementary Fig. S1). Although some tools have competitive results, they have larger fluctuations than GDmicro. For example, while SIAMCAT achieved 0.949 AUC in the Cirrhosis cohort, its mean AUC in the T2D cohort was only 0.691. Similarly, DeepMicro had 0.940 AUC in the Cirrhosis cohort, but its mean AUC in the CRC-FR cohort was only 0.803, which was much lower than the top two tools. The results also reveal that the performance of all methods was not satisfactory in the Obesity and T2D cohorts. A possible reason is that the
Additional validation experiment

CRC-IN Colorectal cancer India 58 28 30 Gupta et al. (2019)

Table 2. The mean AUC (10-fold cross-validation) of five tools on five popular disease cohorts. a

<table>
<thead>
<tr>
<th>Cohorts</th>
<th>GDMicro-batch</th>
<th>GDMicro-single</th>
<th>SIAMCAT</th>
<th>MetAML</th>
<th>DeepMicro</th>
<th>PopPhy-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cirrhosis</td>
<td>0.968 (0.0004)</td>
<td>0.956 (0.001)</td>
<td>0.949 (0.0008)</td>
<td>0.945 (0.029)</td>
<td>0.94 (0.032)</td>
<td>0.94 (0.041)</td>
</tr>
<tr>
<td>CRC-FR</td>
<td>0.923 (0.003)</td>
<td>0.881 (0.005)</td>
<td>0.857 (0.003)</td>
<td>0.809 (0.053)</td>
<td>0.803 (0.077)</td>
<td>0.796 (0.075)</td>
</tr>
<tr>
<td>IBD-DK</td>
<td>0.856 (0.004)</td>
<td>0.882 (0.005)</td>
<td>0.867 (0.005)</td>
<td>0.873 (0.044)</td>
<td>0.863 (0.081)</td>
<td>0.781 (0.094)</td>
</tr>
<tr>
<td>Obesity</td>
<td>0.741 (0.003)</td>
<td>0.697 (0.003)</td>
<td>0.628 (0.005)</td>
<td>0.655 (0.071)</td>
<td>0.659 (0.068)</td>
<td>0.676 (0.074)</td>
</tr>
<tr>
<td>T2D</td>
<td>0.816 (0.001)</td>
<td>0.782 (0.001)</td>
<td>0.691 (0.002)</td>
<td>0.744 (0.048)</td>
<td>0.763 (0.064)</td>
<td>0.753 (0.055)</td>
</tr>
</tbody>
</table>

a The values in parenthesis refer to the margin of errors. Bold: the best performance of each cohort. Underlined: the performance of the second-ranked tool of each cohort. Notably, GDMicro-single has the second-best performance in the table, where underlining is used to highlight the best non-GDMicro approach. “GDMicro-batch:” run GDMicro once for all individuals in the validation sets. “GDMicro-single:” run GDMicro for one individual each time, requiring multiple runs.

These two cohorts. GDMicro, MetAML, and DeepMicro achieve comparable performance to the 10-fold cross-validation experiment in the CRC-FR cohort, while the AUC of SIAMCAT and PopPhy decreases more than 3%. However, the AUC of all tools decreases in the IBD-DK cohort. This result indicates that some models can still maintain good robustness in the cross-study experiments when sufficient training samples are available (e.g. 421 training samples for CRC-FR). However, when the training sample size is small, and some samples are from the same individual (e.g. 94 training samples from 50 individuals for IBD-DK), the classification of cross-study samples can be a challenge for all tested tools. Then, we will discuss the overall performance of these tested tools in all seven test cohorts.

3.3 Cross-study experiment

Test data in real-world applications usually come from different studies or domains from the training data. Thus, it is imperative to evaluate the robustness of the model across different studies or domains from the training data. Therefore, it is important to test the model’s performance on data from different countries. Because CRC and IBD are more associated with human gut microbes based on previous literature (Yu and Fang 2015, Rodriguez et al. 2020) and our experiments, we conducted cross-study experiments for these diseases.

We collected additional human gut microbiome data for CRC and IBD. The additional four CRC cohorts were all sampled from different individuals, while some samples in the additional IBD cohort were sampled from the same individual at different ages. In total, we have five CRC cohorts and two IBD cohorts. All these cohorts are from different countries. Then, we used the leave-one-study-out (LOSO) strategy to evaluate the performance of different tools. Specifically, we remove all samples from one study in the training data and retrain the model. Then, the model is tested on the removed samples. This setup mimics the real scenario where the query sample is from a different research group. The AUC of all tools in seven test cohorts was shown in Fig. 3.

Considering that both CRC-FR and IBD-DK cohorts were used in the 10-fold cross-validation experiment (Table 1), we first analyzed the performance change of different tools in these two cohorts. GDMicro, MetAML, and DeepMicro achieve comparable performance to the 10-fold cross-validation experiment in the CRC-FR cohort, while the AUC of SIAMCAT and PopPhy decreases more than 3%. However, the AUC of all tools decreases in the IBD-DK cohort. This result indicates that some models can still maintain good robustness in the cross-study experiments when sufficient training samples are available (e.g. 421 training samples for CRC-FR). However, when the training sample size is small, and some samples are from the same individual (e.g. 94 training samples from 50 individuals for IBD-DK), the classification of cross-study samples can be a challenge for all tested tools. Then, we will discuss the overall performance of these tested tools in all seven test cohorts.

As shown in Fig. 3A and B, GDMicro with the “batch_run” strategy achieved the best and most stable performance in the LOSO experiments. Compared to the second-best result, GDMicro-batch improved the AUC from 0.837 to 0.915 and 0.861 to 0.907 in CRC-FR and CRC-AT cohorts, respectively. MetAML and SIAMCAT achieved comparable performance in the CRC-US cohorts. However, their performance was not stable in the remaining cohorts. Although DeepMicro and PopPhy-CNN achieved satisfactory performance in the previous experiments, their performance was poor in the IBD-DK cohort. Thus, we further draw a ROC curve of each tool in the IBD-UK cohort in Fig. 3C, which shows that GDMicro has more reliable classification results than other tools.
However, all the tools do not perform well in the IBD-DK cohort, which contains data from the fewest number of individuals out of the seven cohorts. This result indicates that the performance of the learning-based methods may fluctuate with the change of the training cohort size. Nevertheless, GDmicro achieves more than 5% improvement in AUC.

In line with the results from the 10-fold cross-validation experiment, GDmicro with the “single_run” strategy has decreased performance compared to GDmicro-batch. When there are multiple samples, the edges between them or the increased number of edges between training and test samples can help the semi-supervised learning of GCN, which might explain the performance difference between the two scenarios. Despite the lower AUC, GDmicro-single outperforms all other tools on the two IBD cohorts. In addition, its average AUC achieved a 4% improvement compared to the second-best tool across all tested cohorts (Supplementary Table S1). Similar to the previous experiment, we also investigated the influence of test sample size on the cross-study experiment (Supplementary Section S2.1). The result shows that GDmicro exhibited improved performance with an increase in the number of test samples (Supplementary Fig. S2). Especially, the performance has a rapid convergence as the number of test samples increases in some tested cohorts.

From Fig. 3A and B, we also observed that machine-learning-based tools (MetaML and SIAMCAT) tend to have better performance than deep-learning-based tools (DeepMicro and PopPhy-CNN) in most LOSO experiments. This result indicated that deep-learning-based tools are more likely to overfit the training set because of the complexity of the model. However, by using a deep adaptation network to learn the transferable latent features, GDmicro maintains good generalization capabilities in the cross-study host disease status classification task.

3.4 Ablation study and parameter analysis

In this experiment, we study how different architectures and parameters influence the performance of GDmicro using ablation study and parameter analysis. Specifically, we analyzed the influence of the adaptation loss function, GCN model, and hyper-parameter $k$ in the $k$NN graph on the performance of GDmicro. To be more consistent with the usage of real-world data, we analyzed datasets of the cross-study experiment. The analysis results show that the current architecture outperforms all other tested architectures, demonstrating the efficiency of domain adaptation and the GCN model (Supplementary Fig. S3A). In addition, we also found that the performance of GDmicro is not very sensitive to $k$ (Supplementary Fig. S3B). By default, we use $k=5$ to construct the $k$NN graph. Additional details regarding this experiment can be found in Supplementary Section S2.2.

3.5 GDmicro identifies important biomarkers and analyzes their contribution to CRC and IBD

In this experiment, we applied GDmicro to identify important biomarkers and explore their contribution to hosts’ disease status from cohorts used in the cross-study experiment. It should be noted that the goal of this experiment is to identify potential biomarkers, and it is a common practice to conduct biomarker discovery using a batch of samples (Segata et al. 2011, Liu et al. 2022). Thus, we run GDmicro with the “batch_run” mode in this experiment. As mentioned in Section 2, GDmicro could identify important biomarkers using a performance-based method. According to the method, the species with better classification ability will have higher AUC. Therefore, we can evaluate the classification ability of all species in the given data and record their AUC. Then, we selected the top 10 species with the highest AUC values for further analysis. The identified biomarkers and their relative abundance distribution are shown in Supplementary Fig. S4. We observed that many identified species of CRC cohorts were enriched in CRC samples, and their relative abundance was also higher. Unlike CRC, many identified species in the IBD cohort were enriched in healthy samples, which was in line with previous observations (Ma et al. 2021, Olbjørn et al. 2022). Given that abundance-based statistical analysis is a prevalent method for discovering microbial biomarkers, we examined the differences between biomarkers identified by GDmicro and those detected using the Wilcoxon test, a popular statistics-based approach. The details about using the Wilcoxon test to discover biomarkers are described in Supplementary Section S2.3. In the following paragraphs, we...
first analyzed the biomarkers identified by GDmicro and the Wilcoxon test in terms of robustness, discriminatory capabilities, and cross-study consistency. Then, we investigated the contribution of these cross-study biomarkers to the host disease status.

To assess the robustness of identified biomarkers, we further calculated the *P*-value of identified features in each study (Supplementary Fig. S5) and defined two kinds of features. The first feature is called “robust feature” that exhibits a consistent positive or negative *P*-value of <0.05 across all studies. The second type of feature, called “inconsistent feature,” refers to features that exhibit both positive and negative *P*-values across different datasets. In real-world applications, we expect that effective biomarker discovery methods can uncover more robust features while reducing the number of inconsistent features. As Fig. 4A shows, GDmicro identified more robust features and produced fewer inconsistent features than the statistics-based method.

To evaluate the discriminatory capabilities of the identified biomarkers, we conducted the LOSO experiment using only the top 10 species selected by GDmicro and the Wilcoxon test, respectively. For comparison, we randomly selected 10 features from each study and reran the LOSO experiment, repeating this five times to avoid data bias. As shown in Fig. 4B, GDmicro consistently achieved higher AUC values in most tested datasets than the Wilcoxon test and outperformed the set with randomly selected species in all datasets. This result shows that the top-ranked species identified by GDmicro have better discriminatory capabilities, and an additional experiment further confirmed this superiority when using the top 50 selected features (Supplementary Section S2.3 and Supplementary Fig. S6).

Previous studies have reported that biomarkers with cross-study consistency (aka cross-study biomarkers) are crucial for disease identification and are valuable for clinical diagnosis (Wirbel et al. 2019, Zhang et al. 2022). Thus, we compared the top 10 biomarkers identified by GDmicro and the Wilcoxon test across studies. Figure 4C shows that GDmicro identified six cross-study biomarkers for CRC and four for IBD, while the Wilcoxon test detected only two for CRC. Many cross-study biomarkers identified by GDmicro are highly associated with CRC and IBD (Zeller et al. 2014, Zhang et al. 2017, Kwong et al. 2018). For example, Gemella morbillorum and Parvimonas micra are reported as biomarkers for non-invasive diagnosis of CRC (Yao et al. 2021), while Eubacterium rectale, a beneficial bacteria, decreases in

Figure 4. (A). The number of robust and inconsistent features identified by GDmicro and the statistics-based method (Wilcoxon test). (B). AUC comparison across studies in the LOSO experiment using the top 10 features identified by three different methods. The AUC of “10 randomly selected features” is the average AUC of five-time repeated experiments. (C). Venn diagrams for the identified top 10 biomarkers in different cohorts by GDmicro and the statistics-based method. (D). The boxplot of contribution scores of cross-study biomarkers in CRC and IBD datasets. The Y-axis represents the contribution score values of the biomarker in five CRC and two IBD cohorts.
In this work, we demonstrate that GDmicro outperforms the state-of-art tools in classifying host disease status. To reduce the domain discrepancy between data, we use a deep adaptation network to learn the transferable latent features from compositional abundance data. Then, we build an inter-host microbiome similarity graph and apply a GCN model to integrate structural and abundance features and utilize information from unlabeled and limited labeled samples. As a result, GDmicro can achieve more accurate and robust host disease status classification. Furthermore, GDmicro demonstrates superior accuracy in identifying disease-related species compared to the statistics-based method, and it elucidates their contribution to the host’s disease status. This insight offers valuable information for biomarker discovery.

Although GDmicro has improved host disease status classification, it has three major limitations. First, when the graph contains few test samples or only a single test sample, the GCN model cannot fully leverage its advantage to effectively utilize the features from the test samples. Consequently, this limitation can lead to a potential decrease in performance. Second, GDmicro is designed to classify a single disease at a time. If a sample exhibits multiple diseases, GDmicro is unable to simultaneously identify all of these diseases in a single classification. Finally, it is worth mentioning that GDmicro does not possess a specific function optimized for the disease classification of time-series samples, which is a highly relevant task within this field of research.

Thus, we have two goals to optimize in future work. First, current efforts are aimed at classifying whether a host has a particular disease or not. However, in practical applications, input samples may suffer from multiple diseases. Therefore, we intend to combine the graphs of different diseases for GCN to explore whether GDmicro can classify multiple diseases at the same time. Second, we will consider extending GDmicro to accommodate temporal longitudinal microbiome data. To support such an extension, methods like Long Term Memory Networks will be integrated into our current architecture for feature extraction and classification.

### Supplementary data

Supplementary data are available at Bioinformatics online.

### Conflict of interest

None declared.

### Funding

This work was supported by City University of Hong Kong [Project 9678241, 7005866, 7005453]; and Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

### Data availability

The source code of GDmicro and all datasets used in this work are freely available at https://github.com/liaoherui/GDmicro.

### References


Chen J, Chia N, Kalari KR et al. Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. Sci Rep 2016;6: 28484.


