




# EXPERT: transfer learning-enabled context-aware microbial community classification

Hui Chong <sup>†</sup>, Yuguo Zha <sup>†</sup>, Qingyang Yu<sup>†</sup>, Mingyue Cheng, Guangzhou Xiong, Nan Wang, Xinhe Huang, Shijuan Huang, Chuqing Sun, Sicheng Wu, Wei-Hua Chen, Luis Pedro Coelho and Kang Ning 

Corresponding author: Kang Ning, Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China. E-mail: ningkang@hust.edu.cn

<sup>†</sup>These authors contributed equally to this work.

## Abstract

Microbial community classification enables identification of putative type and source of the microbial community, thus facilitating a better understanding of how the taxonomic and functional structure were developed and maintained. However, previous classification models required a trade-off between speed and accuracy, and faced difficulties to be customized for a variety of contexts, especially less studied contexts. Here, we introduced EXPERT based on transfer learning that enabled the classification model to be adaptable in multiple contexts, with both high efficiency and accuracy. More importantly, we demonstrated that transfer learning can facilitate microbial community classification in diverse contexts, such as classification of microbial communities for multiple diseases with limited number of samples, as well as prediction of the changes in gut microbiome across successive stages of colorectal cancer. Broadly, EXPERT enables accurate and context-aware customized microbial community classification, and potentiates novel microbial knowledge discovery.

**Keywords:** microbial community classification, transfer learning, context-aware, disease classification, knowledge discovery

## Introduction

Numerous microbial community samples from diverse niches have been sequenced such as those from the 'Human Microbiome Project' [1, 2] and the 'Earth Microbiome Project' [3, 4]. Knowledge about microbial communities and their interactions with environment and human health has been thus expanded [5, 6], such as water pollution [7], land degradation [8], microbial dysbiosis linked to disease pathogenesis of colorectal cancer (CRC),

inflammatory bowel disease (IBD) and type 2 diabetes [9–11]. Such massive number of microbial community samples provides the opportunity to study the inconspicuous evolutionary and ecological patterns of microbial communities, especially habitat-specific patterns.

Microbial community classification has found its application in diverse contexts (e.g. classification among multiple categories including habitating niches, hosts or associated diseases). In a

**Hui Chong** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Yuguo Zha** is a PhD student at the Huazhong University of Science and Technology. His research interests are in microbiome associated data mining, including gene mining, species mining and pattern mining.

**Qingyang Yu** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Mingyue Cheng** is a PhD student at the Huazhong University of Science and Technology. His research interests are in microbiome associated disease analysis.

**Guangzhou Xiong** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Nan Wang** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Xinhe Huang** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Shijuan Huang** is interested in the application of deep learning to parse dynamic patterns of microbial communities and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Chuqing Sun** is interested in microbiome associated disease analysis and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Sicheng Wu** is interested in microbiome associated disease analysis and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Wei-Hua Chen** is interested in microbiome associated analysis methods development and worked at the College of Life Science and Technology, Huazhong University of Science and Technology.

**Luis Pedro Coelho** is interested in computational biology and worked at the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University.

**Kang Ning** is a professor at Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology.

Received: June 7, 2022. Revised: August 8, 2022. Accepted: August 15, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

typical context, there are several community samples from multiple classes, and the aim is to assign samples to the correct classes. And in all of these contexts, the classes are also referred to as the biomes. For example, a previous study reported thousands of community samples from the gut of patients with different stages of CRC (e.g. stages I, II, III and IV), and the purpose of classification in this context is to assign gut community samples to the correct stages of CRC [12]. Generally, the complexity of microbial community classification is positively correlated with the number of classes and negatively correlated with the number of community samples [13]. For a given context that involved multi-disease classes but limited number of samples, such as classification of 4026 samples from 28 case-control microbial studies spanning 10 diseases [14], low prediction accuracies are naturally expected for identification of disease-specific patterns, rendering microbiome-based classification unpractical.

Faced with these contexts, current methods for microbial community classification have limitations in dealing with such paramount of complex relationships and biome-specific patterns. While it becomes extremely difficult when there exists biomes in which there are only a few samples, a 'Big Data, Small Sample' problem [13]. Random forest model is suitable for classification among numerous samples, and it has been used in many applications, such as chronological age prediction [15] and fecal source identification [16, 17]. SourceTracker [18] and FEAST [19] are the two representative unsupervised learning methods for microbial community classification. These unsupervised learning methods are based on profile-based statistical models, either the Bayesian model used in the SourceTracker method, or the Expected-Maximization model used in the FEAST method. However, since unsupervised methods still do not consider the intricate but important patterns of a set of samples from similar biomes, their tolerance to noisy signals in samples is not high, hence potentially would lead to biased mismatches. In addition, both SourceTracker and FEAST require an insufferable tradeoff between running time and accuracy, especially when faced with the 'Big Data, Small Sample' context. For instance, performing microbial community classification among thousands of samples within hundreds of biomes may take hours for these methods [19]. Recently, ONN4MST was proposed to solve the irreconcilable contradiction between efficiency and accuracy [20]. ONN4MST is a supervised learning method based on ontology-aware neural network (ONN) model, which contains multiple output layers fitting with a general biome ontology. Notably, it was designed specifically to search microbial community samples against a general biome ontology. However, if there comes a new biome ontology with more detailed biomes involved (such as different stages of CRC), or simply with more biome relationships involved, then the general ONN model will be not applicable. Thus, ONN4MST's general model cannot be customized for specific contexts such as classification of microbial communities to specifically designated diseases or hosts. Moreover, training with a limited number of community samples from additional specialized biomes makes obtaining a robust classification model extremely challenging.

To address the above limitations, we proposed an exact and pervasive expert model for microbial community classification based on transfer learning (TL), namely EXPERT. EXPERT employs the ONN framework and gains advantage as regard to the trade-off between efficiency and accuracy. More importantly, EXPERT benefits from the TL technique that enabled the classification model to be adaptable in multiple contexts, especially those that classifies a few of community samples from a large number of

biomes or classes. Specifically, EXPERT utilizes TL technique to build transferred ONN model, which inherited partial parameters (i.e. weights) from general ONN model (e.g. the general ONN model of ONN4MST). Thus, EXPERT can utilize the knowledge of fundamental models (e.g. general ONN model) to aid in the learning of the transferred ONN models. In other words, EXPERT can deal with not only the general biome ontology, but also adapt to biome ontology such as a disease-related human gut biome ontology [21] consisting of terms of diseases associated with human gut microbial communities. Here, we first evaluated the efficiency and accuracy of EXPERT on newly deposited microbial community data in MGnify [22]. We then demonstrated its adaptivity in classification of community samples under diverse contexts, including: (1) different body sites, (2) different age of hosts, (3) different diseases and (4) different stages of CRC. The analyses of EXPERT in these contexts have shown its superior performance in microbiome sample classification in a broad-spectrum of contexts.

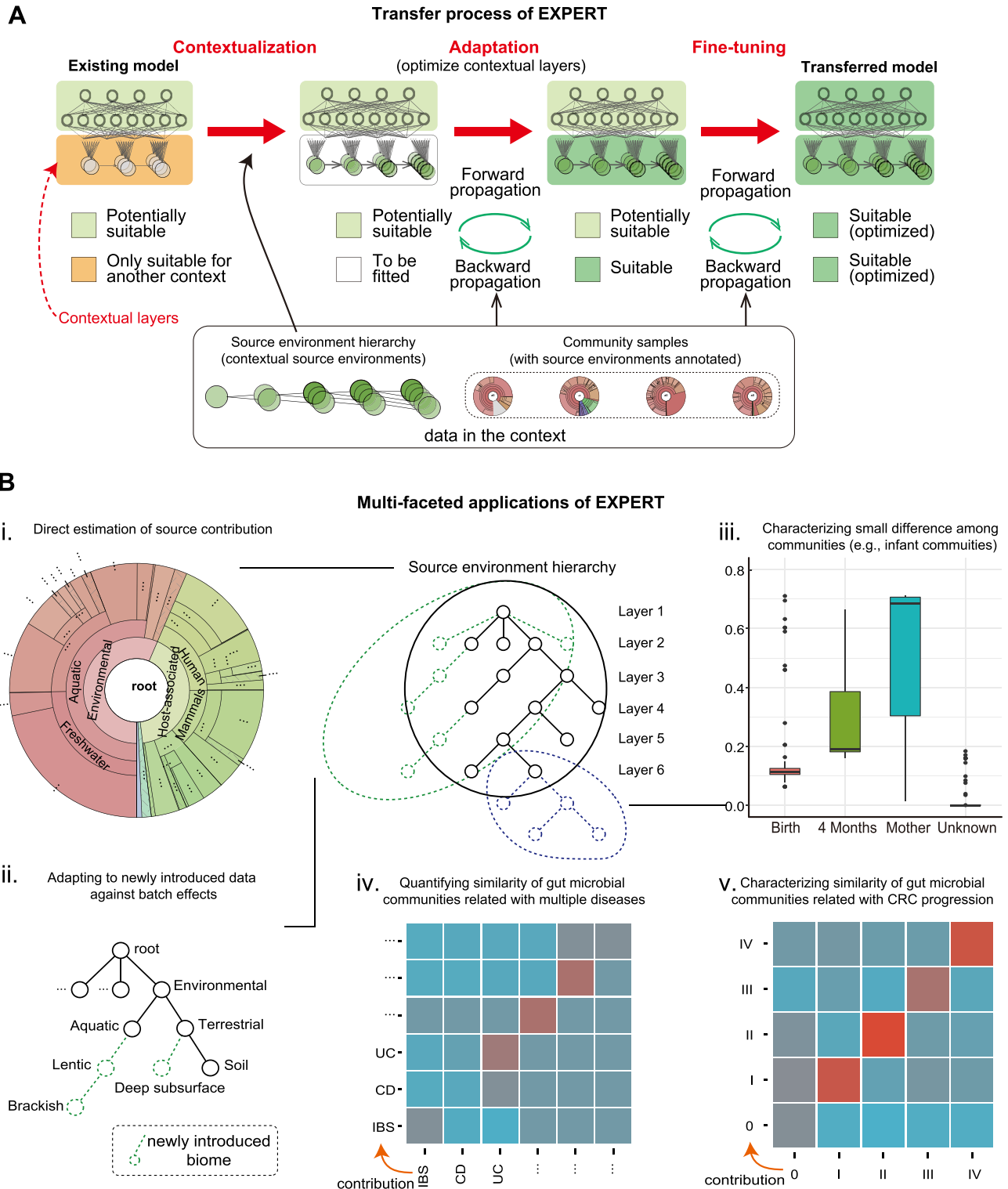
## Results

### Rationale, adaptive modeling and multi-faceted applications of EXPERT

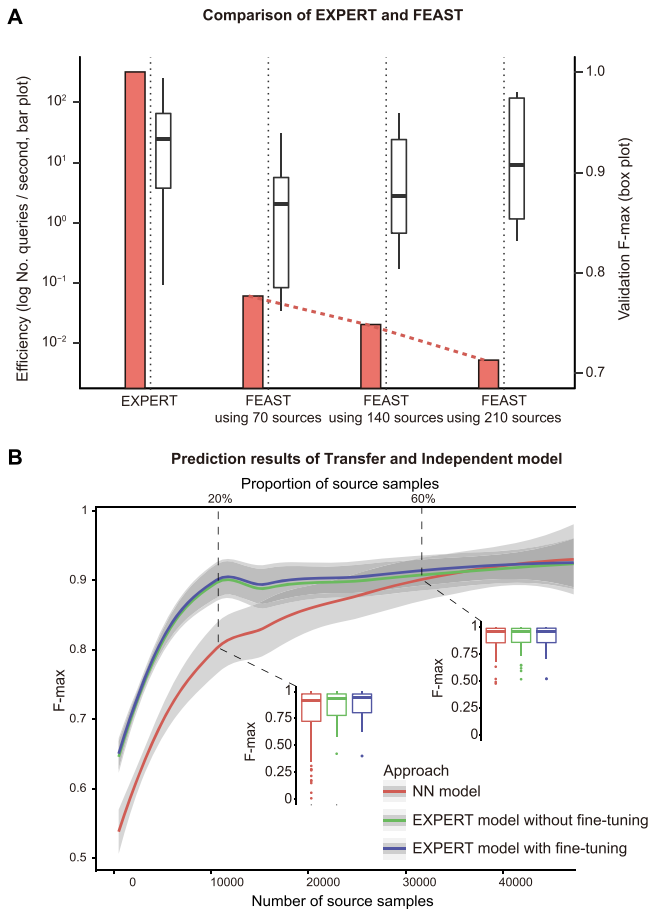
In this study, we proposed an exact and pervasive expert model for microbial community classification based on TL, namely EXPERT, which is a context-aware method for microbial community classification that employs both neural network (specifically ONN) and TL technique. First, EXPERT enabled adaptation to new biome ontology by employing a structurally adaptive fundamental ONN model (Supplementary Figure S1). The fundamental ONN model consists of two parts: fixed fully connected layers for extracting general representation of input microbial community data and contextual layers for extracting representations that are specific to biome ontology layers (Methods, Supplementary Figure S2). The contextual layers of the fundamental ONN model could be reinitialized to fit with different biome ontology. The input is a relative abundance matrix of all the samples. Each row represents a sample and each column represents a taxon. The outputs are the source contributions of biomes belonging to each ontology layer. Second, TL technique enables EXPERT to have the context-aware ability for microbial community classification. Specifically, EXPERT utilizes TL technique to build transferred ONN model, which inherited partial parameters (i.e. weights) from general ONN model (e.g. the general ONN model of ONN4MST). Thus, EXPERT can utilize the knowledge of fundamental models (e.g. general ONN model) to aid in the learning of the transferred ONN models.

The knowledge transfer process of EXPERT is illustrated in Figure 1A. EXPERT adopted the rationale of TL technique [23], allowing context-aware microbial community classification through three steps namely, transfer, adaptation and fine-tuning. In the transfer step, EXPERT adapts the fundamental model to the biome ontology under a given context. In the adaptation and fine-tuning steps, EXPERT optimizes the parameters (weights) of the transferred ONN model (Methods, Supplementary Note 1). The contextualized model (i.e. the transferred ONN model) can serve as a broad-spectrum of classification applications (Figure 1B).

In this study, three fundamental models were introduced for knowledge transfer (Supplementary Tables S1–S5): the general model (GM, cross-validated on 118 592 community samples from 131 representative biomes in MGnify), the human model (HM, cross-validated on 52 537 community samples from 27 human-associated biomes in MGnify) and the disease model (DM,



**Figure 1.** Illustration of EXPERT’s knowledge transfer process. (A) EXPERT can adapt the knowledge of a fundamental model to a classification context through three steps: transfer (reuse parameters of a fundamental model and reinitialize contextual layers according to the context, red dotted arrows), adaptation (quickly optimize only the contextual layers using iterative forward-backward propagation, green circular arrows) and fine-tuning (further optimize the entire model using the iterative forward-backward propagation). The fundamental model is a pre-trained EXPERT model to be adapted, with several NN layers relatively independent to contexts and a series of contextual NN layers highly specified to a context. Different background colors of the model indicate the suitability of different modules to the context. (B) The contextualized model can serve a broad-spectrum of applications (based on research purposes). To avoid the impact of batch effect between datasets, we normalized the reads’ count belonging to each taxon by the sequencing depth of the sample. We chose the relative abundances at phylum, order and genus ranks in all the experiments. We also applied Z-score standardization when training the model and used the mean and standard deviation of training data to normalize the testing samples. Data preprocessing detail is explained in [Supplementary Notes 3 and 4](#). Abbreviations: NN, neural network.



**Figure 2.** Efficiency, accuracy and adaptivity of EXPERT. **(A)** Comparison of transfer (GM) EXPERT model with FEAST on efficiency (number of queries/sinks per second, left Y-axis) and accuracy (based on cross-validation, right Y-axis). For FEAST, the sources were randomly selected 70, 140 and 210 samples (10, 20 and 30 samples per biome, respectively). EXPERT's performance was measured by contextualizing the GM. **(B)** The performance (validation F-max, Y-axis) of three models along with different proportions of sources used (X-axis). The NN model was trained solely based on contextual data. The results were obtained by using cross-validation and different proportions (1–10% by a step size of 1% and 10–90% by a step size of 10%) of source samples. Loess regression was applied to these points using the number of source samples and F-max.

cross-validated on 13 642 fecal community samples from patients of 19 diseases and healthy controls).

### Efficiency, accuracy and adaptivity of EXPERT

In this part, we assessed EXPERT using 52 537 community samples from 27 human-associated biomes from MGnify [22] (Supplementary Tables S1, S3, Supplementary Figure S3). Benchmark tests demonstrated EXPERT's superior efficiency, accuracy and adaptivity for microbial community classification (Figure 2). Specifically, EXPERT outperformed SourceTracker and FEAST in terms of efficiency and accuracy, and outperformed ONN4MST in terms of accuracy and adaptability.

We compared EXPERT's performance with that of FEAST, which is comparably accurate but faster than the other most used SourceTracker [19]. Here 47 283 (90% of 52 537) community samples were used, in combination with the GM as the fundamental model, for training the transferred GM model, while 5254 (10% of 52 537 communities) samples were used for testing both methods. Since FEAST can only deal with a few hundred to thousand

community samples, we randomly selected small sets of community samples in seven independent biomes to run FEAST (Figure 2A, Methods). We found that along with the increase of samples used in FEAST ( $n=70, 140$  and  $210$ ), its efficiency decreases dramatically (0.06, 0.02 and 0.005 queries/second, Figure 2A and Supplementary Table S6) with an increased accuracy (F-max = 0.847, 0.884 and 0.911). However, EXPERT could balance the tradeoff between accuracy and efficiency, which could simultaneously reach higher accuracy and efficiency (F-max = 0.923, over 200 queries/second, Figure 2A).

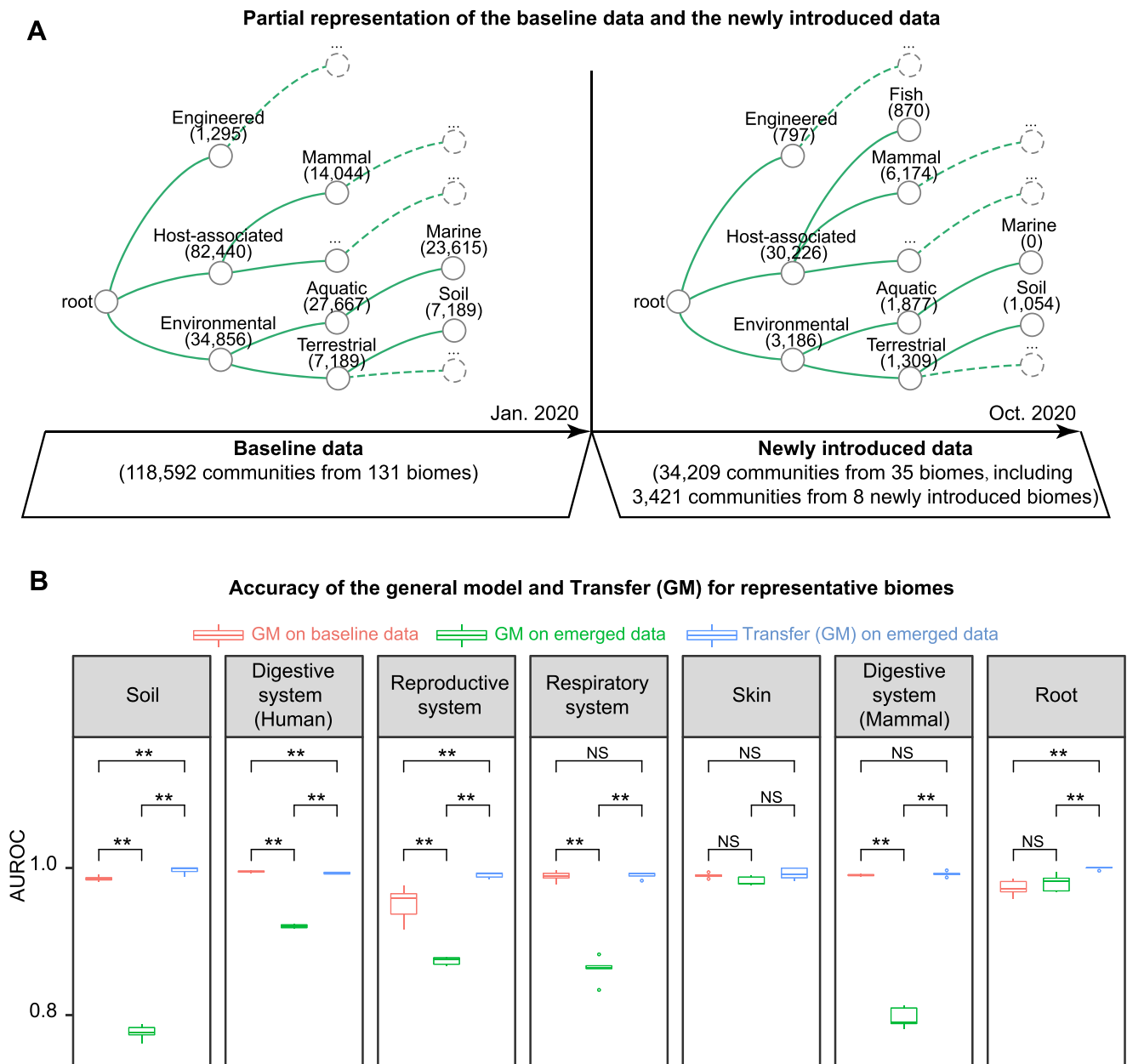
We also compared EXPERT's accuracy with ONN4MST, by using different gradients (from 1% to 90%) of 47 283 samples for training, while using the same 5254 samples for testing (Figure 2B). Notably, we have reimplemented the ONN4MST model for evaluation, since the original ONN4MST could not be directly applied in this context. Results showed that the EXPERT model outperforms ONN4MST on accuracy: although the accuracy of both methods steadily increased with the increased proportion of samples used for training, the EXPERT model only required 10% of training samples to achieve a validation F-max of 0.814, whereas ONN4MST required three times as many training samples to reach a similar validation F-max of 0.813 (Supplementary Table S7). Therefore, benefited from TL technique, EXPERT models were able to fit a given context based on less training samples compared to ONN4MST. Notably, as the fine-tuning optimization clearly improved the accuracy (Figure 2B), the knowledge transfer with fine-tuning was considered the default setting in the following sections.

### EXPERT classifies newly introduced microbial community samples in less studied contexts

In this context, we aim to validate EXPERT's adaptability to newly introduced microbial community samples, such as those obtained through new sequencing and analytical technologies or from rarely studied environments. To test EXPERT's capability in such context, in addition to the 118 592 community samples accessed as on January 2020 from MGnify (referred to as baseline data, Supplementary Tables S1, S2 and Figure 3A), we selected 34 209 community samples from MGnify between January 2020 and October 2020 (referred to as newly introduced data, Figure 3A and Supplementary Tables S1, S8, Supplementary Figure S4). Among the newly introduced data, 30 788 community samples were originated from 27 biomes included in the baseline data, whereas 3421 community samples were originated from eight newly introduced biomes (Supplementary Figure S4).

We first tested the applicability of the GM and EXPERT on the 30 788 communities that originated from 27 biomes included in the baseline data. The GM performed on baseline data showed an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.982, while the GM performed on newly introduced data showed a much lower AUROC of 0.884 (Supplementary Note 2). The reason behind this might be the data heterogeneity between the two datasets (Supplementary Figure S5). However, these potential effects might be reduced by using EXPERT to transfer the GM to the newly introduced data, obtaining an improved AUROC of 0.993 (Figure 3B).

We then tested the applicability of EXPERT on the 3421 communities that originated from eight newly introduced biomes. We found that even though these newly introduced biomes were not included in the baseline data, the EXPERT could transfer the GM to the newly introduced biomes with AUROC of 0.988. As demonstrated by these results, EXPERT has the potential for



extending fundamental models (e.g. GM, HM and DM) into previously unexplored contexts such as those understudied biomes.

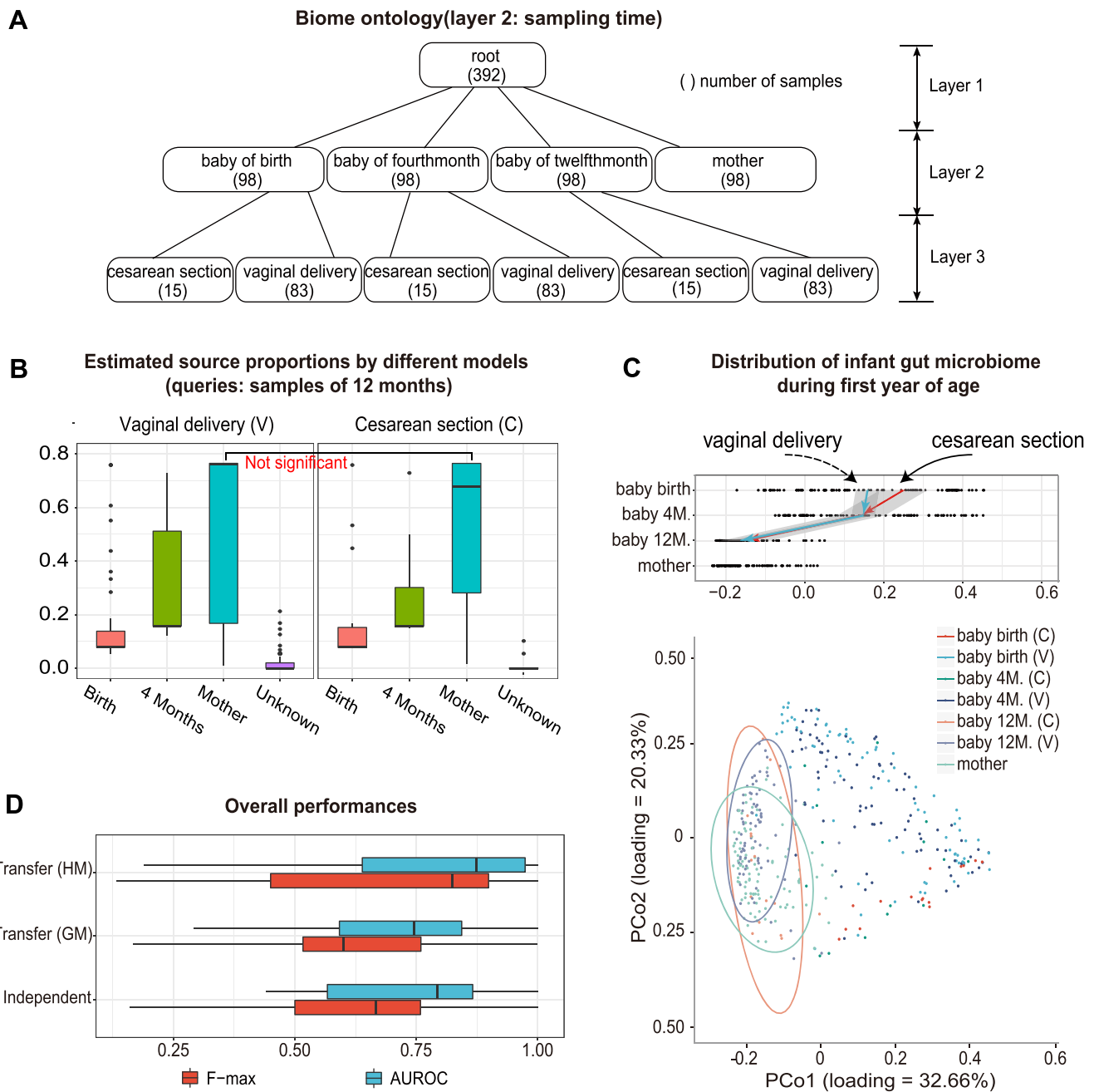
### EXPERT dissects the infant gut microbial communities according to sampling time and delivery mode

We next used EXPERT to characterize tiny successive changes among infant gut microbial communities during the first year of life, through quantifying the microbial sources. We used longitudinal data from Backhed *et al.* [24], including fecal samples from 98 infants and their mothers, delivered by vaginal delivery or cesarean section (Figure 4A and Supplementary Tables S1 and

S9), in this study. In this context, the classification probability of a prediction model was considered as a proxy of source contribution. We considered samples from earlier time points and samples from mothers as the potential microbial sources to train the model. We then used the model to predict the microbial sources for infant samples at 12 months of age. Here we have assessed multiple prediction models: EXPERT transferred models based on different fundamental models, and the independent model built by using samples in this context alone.

Based on the biome ontology that divided samples by sampling time first followed by delivery mode (Figure 4A), we noticed that for infant gut microbial communities at 12 months of age, the





**Figure 4.** EXPERT's performance in characterizing gut microbial community development over time for infants. **(A)** The hierarchy representing source environments, corresponding to infant samples collected from the ENA database. Environments in the second and third layers were grouped by sampling time and delivery modes. For this part of the study, sources include the gut microbiome of the mother, infant at birth and 4 months, queries include the gut microbiome of the infant at 12 months. **(B)** Estimated contributions by transfer (HM) model, separated by two delivery modes. **(C)** Distribution of infant gut microbial communities during their first year, using PCoA and distance metric of Jensen Shannon divergence. The dotted line refers to samples delivered vaginally, and the full line refers to samples delivered via cesarean section. The baby of 4 months is abbreviated to baby 4M, the baby of 12 months is abbreviated to baby 12M. The letters 'C' and 'V' stand for cesarean section and vaginal delivery, respectively. Top panel: samples from the infant's gut are plotted according to their source and collection date on the Y-axis, and position on the X-axis is plotted according to their first principal coordinate in the PCoA. **(D)** The overall performance of models generated based on different fundamental models, in which the Independent model was solely based on the samples used in this context; transfer (GM) and transfer (HM) refer to models built based on the GM and HM with fine-tuning, respectively.

maternal contribution is dominant (Figure 4B). Moreover, there is no significant difference in the maternal contribution between cesarean-born and vaginal-born infants (Wilcoxon test,  $P=0.929$ , Figure 4B), consistent with principal coordinate analysis (PCoA) using distance metric either in weighted-UniFrac [25] or Jensen Shannon divergence [26] (Figure 4C and Supplementary Figure S6). We concluded that the infant gut at 12 months is largely

adapted to exposed environments, resulting in an insignificant difference between samples collected from hosts of different delivery modes, consistent with previous studies [27, 28].

We then assessed the utility of different fundamental models in this context (Supplementary Figures S7 and S8). We found that the HM can facilitate microbial community classification in this context with significantly better performance compared with the

GM [Transfer (HM): AUROC = 0.773, Transfer (GM): AUROC = 0.720, Wilcoxon test,  $P = 0.072$ ], suggesting the use of the HM in this application. Therefore, we suggest that when using EXPERT, it is necessary to choose a proper fundamental model according to the specific context (Figure 4D).

### EXPERT reveals disease-specific patterns within gut microbial communities

The pattern of gut microbial communities could be disease-specific, reflecting the distinct inflammation patterns across diseases. In this context, we aimed to demonstrate EXPERT's utility in characterizing human gut microbial communities associated with different types of diseases. Using EXPERT, we can measure patterns across multiple diseases. Specifically, we assembled a large gut microbial community dataset, including 13 642 community samples representing 19 diseases (Figure 5B) and healthy controls, collected from 101 studies and 27 countries (Figure 5A and Supplementary Tables S1 and S4). There are profound differences among the number of samples for different diseases, with only 268 samples for liver cirrhosis, 298 samples for IBD, while 1145 samples for colitis ulcerative (Supplementary Table S4). For EXPERT, we have used the HM as the fundamental model for disease classification, and we have implemented repetitive cross-validation (90% for training and 10% for validation, five repeats) for assessment. Results revealed that, except for Crohn's disease, the pattern is specific to each of the other 18 diseases (Supplementary Figure S9), consistent with a previous study of disease-specific patterns within the human gut microbial communities [29].

We further validated the disease-specific patterns by utilizing an independent model constructed entirely from the same training and validation samples. We found that both Independent model and Transfer (HM) model could distinguish diseases with high AUROC (over 0.8 for most diseases, Figure 5C, D), and confirmed that the gut microbial communities may be used to discriminate between these diseases. This demonstrated the utility of EXPERT in large-scale microbial community classification analysis, particularly when comparing a wide variety of microbial communities from multiple environments.

### EXPERT characterizes gut microbial communities during cancer progression

Gut microbial communities undergo compositional changes as cancer progresses, and this can be observed in the human gut microbiota, which has been shown to influence the progression of CRC [12]. In this context, we demonstrate EXPERT's utility in characterizing gut microbial communities during the progression of CRC. We assessed the applicability of EXPERT by leveraging the DM as the fundamental model for cancer stage classification (Figure 6A). We considered 635 samples from five stages in the progression of CRC: 0 (Healthy control) I, II, III and IV according to the study of Zeller *et al.* [12] (Figure 6B and Supplementary Tables S1 and S10). Preliminary analysis using weighted-UniFrac [25] could not show the compositional shifts of the human gut within such progression (Figure 6C and Supplementary Figure S10). However, by repetitive cross-validation (90% for training and 10% for validation, five repeats), we found that, based on the gut microbial community, EXPERT can clearly predict the CRC progression stage (Figure 6D). We also assessed the EXPERT's capability in monitoring the progression of CRC, by comparing the performance of different models: the Transfer (DM) model, the Transfer (HM) model and the independent model (solely based on the CRC samples). Results showed that the Transfer (DM) model

achieved a better performance (AUROC = 0.977, Figure 6E) among these three models, highlighting the EXPERT's utility on classifying the different stages of CRC progression using gut microbial communities, which is superior than most of the contemporary methods. These results indicated the association between gut microbiota and CRC progression and EXPERT could track the progression of CRC based on gut microbiota [12, 14].

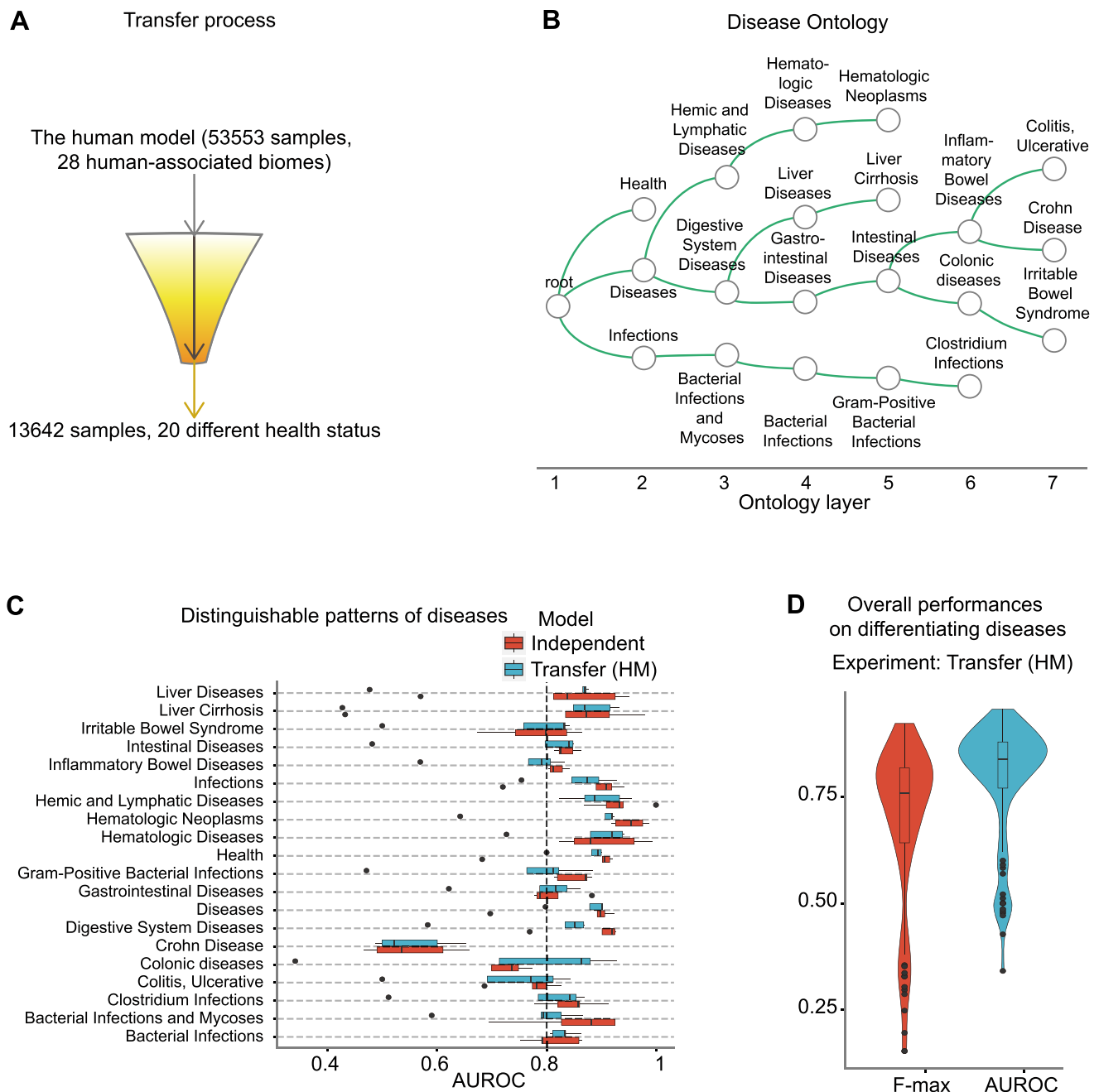
## Discussion

Broadly, EXPERT adopted neural network and TL techniques to profoundly expand the applicability of microbial community classification, enabling discovery of unique microbiological knowledge in diverse contexts. EXPERT presented a high adaptability to diverse contexts, such as classification among multiple categories including habitating niches, hosts or associated diseases, for in-depth knowledge discovery.

Our analytical results have confirmed that EXPERT has enabled microbial community classification with high speed and fidelity. We should emphasize that the method could be used on a regular laptop (e.g. a Linux laptop with dual cores CPU, 8GB memory and 128GB disk). The evaluation showed that EXPERT reached a high efficiency (over 200 queries/second, Figure 2A), which is several orders of magnitude faster than the FEAST method (under 0.1 query/second). In this study, three fundamental models were introduced for knowledge transfer, EXPERT could adapt the fundamental models to newly introduced data, and thus could be utilized in a broad-spectrum of microbial community classification contexts, especially less studied contexts. Furthermore, EXPERT benefits from the TL technique that enabled the classification model to be adaptable in multiple contexts, especially those that classifies a few of community samples from a large number of biomes or classes, such as predicting the CRC progression stage.

We have demonstrated EXPERT's utility in context-aware microbial community classification in several applications. First, EXPERT can characterize the tiny compositional difference associated with environmental changes. By adapting EXPERT to microbial communities of infant gut across delivery modes, we found that due to environmental exposure during the first year, cesarean-born infants have a largely restored gut microbial community compared with infants born vaginally, consistent with the results of other published analyses [27, 28]. Second, we demonstrated the utility of EXPERT beyond traditional methods by incorporating a dataset of multi-disease gut microbial communities. By using EXPERT on the dataset, we discovered that the human gut microbial community exhibits disease-specific patterns, which is consistent with previous cross-disease research [29]. Third, we showed EXPERT's utility in characterizing the gut microbiota for patients at various stages of CRC. By using communities from five stages of CRC progression, we found that hosts sampled at the same stage shared similar gut microbial communities, enlightening us to realize that the compositional changes within gut microbial communities could reflect the progression of CRC, supported by Zou *et al.* [14].

Context-aware is becoming an important direction in microbiome data mining field. Our study shows that TL enables context-aware microbial community classification in a broad-spectrum of applications, such as classification of microbial communities for multiple diseases with limited number of samples, as well as prediction of the changes in gut microbiome across successive stages of CRC. Context-aware has also been used to other microbiome data mining fields, such as classifying



**Figure 5.** EXPERT reveals disease-specific patterns within gut microbial communities. **(A)** Illustration of knowledge transfer utilized for disease pattern analysis. The knowledge transfer between classification contexts was illustrated using different colors (white for human-associated biomes, yellow for gut microbiota-associated disease status). In this analysis, the knowledge from the HM was contextualized (transferred) to the dataset containing 13 642 samples and 19 diseases as well as healthy control. **(B)** The hierarchical organization of 19 diseases and healthy control. The hierarchy was constructed by referring disease names to Medical Subject Headings and Human Disease Ontology. The hierarchy includes 20 different health statuses (19 different diseases and infections, plus healthy control) distributed in seven different layers (X-axis). **(C)** The performance of the EXPERT models on the gut microbial community associated with each disease or healthy control, evaluated based on the source contribution and biome-specific evaluation (Methods). The dashed line indicates an AUROC of 0.800. **(D)** The overall performances of the transfer (HM) model. Settings of quantification and assessment were the same as Figure 5C.

of metagenomic reads [30] and parsing gut microbial community dynamic [31].

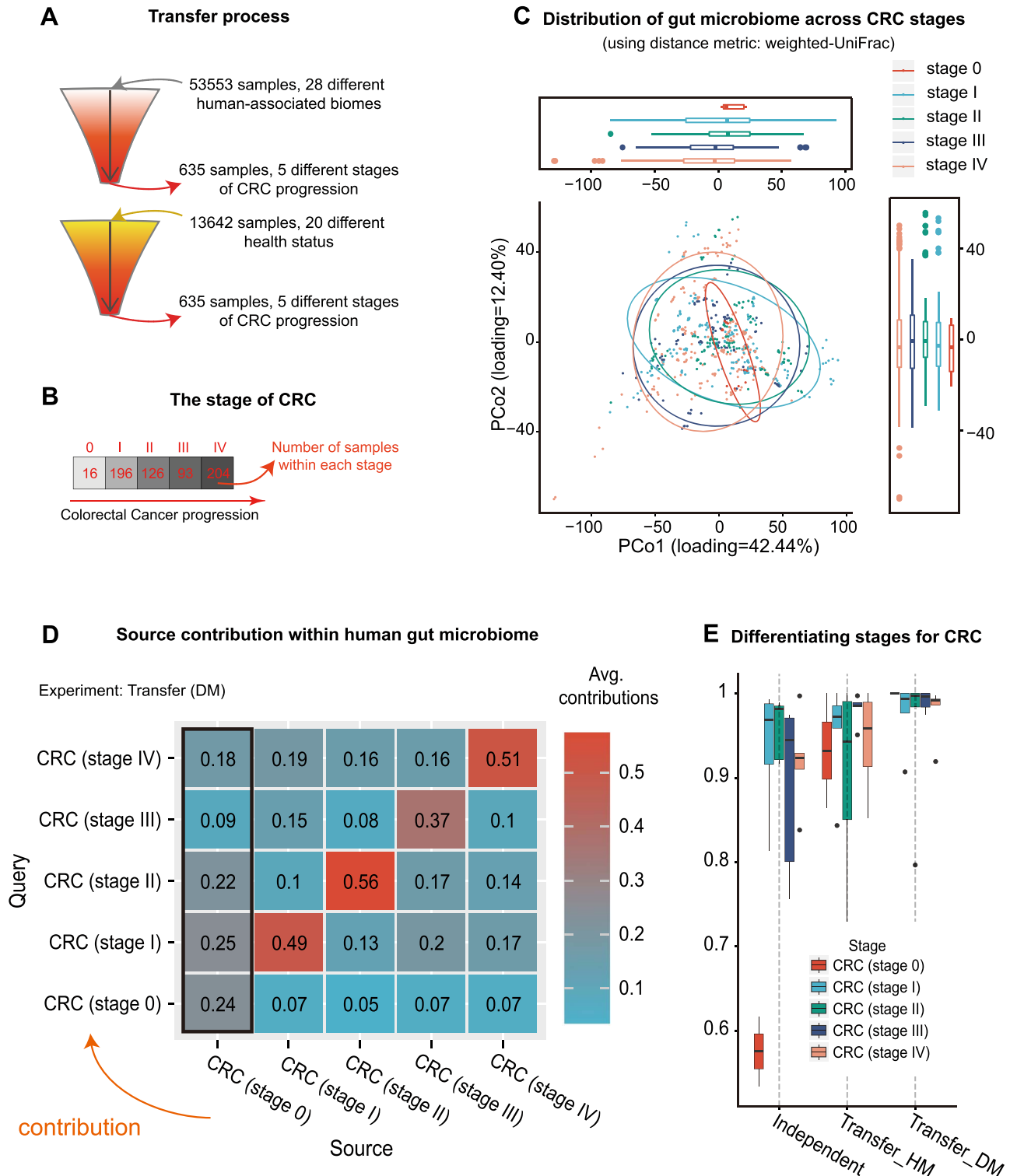
Several issues need to be looked into further in the future: We noted that in certain contexts (e.g. characterizing gut microbial communities during cancer progression), the accuracy could be improved if the fundamental model was properly selected by referring to the standard ontology [21, 32]. EXPERT should provide a collection of fundamental models to enable effective adaptation in diverse contexts, and provide an approach for intelligently

selecting appropriate fundamental models for a given context. Additionally, the application of EXPERT on the newly introduced data has indicated its robustness.

## Conclusions

In conclusion, EXPERT enabled accurate and rapid microbial community classification, as well as biologically informed novel microbial knowledge discovery, by utilizing a TL approach. We





**Figure 6.** EXPERT characterizes compositional shifts within host gut microbiota during CRC progression. **(A)** Illustration of knowledge transfer utilized for characterizing the compositional shifts. The knowledge from the HM learned from 52 537 human-associated communities, as well as the DM learned from 13 642 human gut communities associated with 19 diseases and healthy control, were transferred to characterize the CRC-related compositional shifts. **(B)** The five stages of CRC progression, and the number of samples for each stage. Stage 0: healthy control. **(C)** The distribution of gut microbiomes, visualized by PCoA (using distance metric of weighted-UniFrac). The source samples were randomly selected 90% out of the entire dataset. The query samples were the remaining 10% samples. This process of random selection and quantification was repeated five times. There is no sample overlap between source samples and query samples. **(D)** The average contribution of different stages of CRC. The source samples were randomly selected 90% out of the entire dataset. The query samples were the remaining 10% samples. This process of random selection and quantification was repeated five times. **(E)** The stage-specific performances (AUROC) of EXPERT on different CRC stages (see Methods for details of stage-specific evaluation).

have demonstrated the applicability of TL in the discovery of microbiome knowledge using this method, particularly when dealing with newly introduced data or context-dependent settings. We believed that EXPERT could facilitate high-fidelity microbial community classification in a broad-spectrum of applications.

## Methods

### Microbial community classification with EXPERT

EXPERT is a supervised learning (i.e. neural network and TL) method for microbial community classification. Specifically, the goal is to train a classification model  $Y = f(X, W)$ , where  $X$  is the species relative abundance vector of the input microbial community sample,  $Y$  is the biomes (e.g. CRC stages) to predict and  $W$  is the parameters (e.g. weights and ontology) of the model. The input  $X$  is a relative abundance matrix of all the samples. For the relative abundance matrix, each row represents a sample and each column represents a taxon. The output  $Y$  is the contribution (probability) of biomes belonging to each ontology layer.

### The EXPERT framework

The EXPERT framework adopts NN approach and TL technique (Supplementary Figure S1). Considering a query sample  $q$  represented by its community structure, as well as its potential sources represented by a hierarchy  $O$ , to quantify contributions  $\hat{y}_q$  from the sources to  $q$ , we employed an adaptive and multi-task NN to learn a mapping  $M$  from a series of source samples  $s \in D_S$  to their biome sources,  $y_s = (y_s^1, \dots, y_s^l)$  (where  $y_s^i$  is biome source for source sample in the second layer of the biome hierarchy), and then apply  $M$  on  $q$  to determine the contributions for the query community:

$$\hat{y}_q = \left( \hat{y}_q^i \right)_{0 < i \leq l_O} = M(q).$$

### Fast inference via forward propagation

We adopt the rationale of multi-task learning. EXPERT integrates the representation of each lower layer (which is calculated by its ‘inter’ modules  $M_{\text{inter}}$ ) into its higher layer, by employing several ‘integ’ modules  $M_{\text{integ}}$ . Therefore, together with ‘output’ module  $M_{\text{output}}$ , the representation of the contributions is given by

$$M(q) = \left( M_{\text{output}}^i \left( R_{\text{integ}}^i \right) \right)_{0 < i \leq l_O}$$

where

$$R_{\text{integ}}^i(q) = \begin{cases} M_{\text{integ}}^i \left( M_{\text{inter}}^i \left( M_{\text{base}}(q) \right), 0 \right), & \text{if } i = 1 \\ M_{\text{integ}}^i \left( M_{\text{inter}}^i \left( M_{\text{base}}(q) \right), R_{\text{integ}}^{i-1} \right), & \text{otherwise.} \end{cases}$$

The NN structures of these modules are described in the following section ‘TL model’.

### Robust optimization via backward propagation and TL

We adopt the rationale of TL. Considering  $M_{\text{base}}$  of a fundamental model as a static mapping, the parameters of the rest modules  $\hat{w}$

could be solved using gradient descent as well as backpropagation algorithm:

$$\hat{w} = \operatorname{argmin}_{\hat{w}} \sum_{i=0}^{l_O} \left( \alpha \left( B_O^i \right) \sum_{s \in S} \beta_s^i L \left( \hat{y}_s^i \left( \hat{w} \right), y_s^i \right) \right),$$

where

$$\beta_s^i = \begin{cases} 1, & \text{if } y_s^i \text{ exists} \\ 0, & \text{otherwise} \end{cases},$$

$$L \left( \hat{y}_s^i, y_s^i \right) = \sum_{b \in O^i} \left( \text{CrossEntropy} \left( \hat{y}_s^i(b), y_s^i(b) \right) \right).$$

Here  $\alpha \left( B_O^i \right) = \frac{B_O^i}{B_O}$  stands for the assigned loss weight for  $i$ th layer (i.e.  $i - 1$ th task in the multiple task).  $\beta_s^i$  stands for the sample weight assigned for a sample  $s$  on  $(i - 1)$ th task during learning, enabling the learning from partially labeled data.  $B_O^i$  stands for the number of biomes contained in the  $i$ th layer of the biome hierarchy  $O$ .  $O^i$  stands for the  $i$ th layer of the biome hierarchy  $O$ .  $b \in O^i$  is a biome in the biome hierarchy  $i$ th layer of the biome hierarchy  $O$ .

Then, optimizing the parameters of the entire model (including  $\tilde{M}_{\text{base}}$ ), the parameters of the entire model  $w$  can be solved by using gradient descent as well as backpropagation algorithm:

$$w = \operatorname{argmin}_{\hat{w}} \sum_{i=0}^{l_O} \left( \alpha \left( B_O^i \right) \sum_{s \in S} \beta_s^i L \left( \hat{y}_s^i \left( \hat{w} \right), y_s^i \right) \right)$$

For independent optimization (optimization based on completely random initialization), EXPERT directly optimizes the entire model. See Supplementary Note 1 for a detailed description for optimization.

### TL model

NN approach has limited capability when there is a series of newly introduced source environments, as researchers need to modify the NN model at the code level and retune its hyperparameters. We developed EXPERT’s NN model that changes internal NN structure according to source environments in different contexts, namely the adaptive NN model (Supplementary Figure S1). The EXPERT framework initializes the model according to the hierarchy representing source environments. In the model, there are four conceptual modules.

To extract low-level representations for input data, the model employs the ‘base’ module with two Dense NN layers. The NN layers have fixed structures of 1024 and 512 neurons, and use ReLU activation and He initializer with Uniform distribution.

To extract representations that are specific to different hierarchy layers, the model employs the ‘inter’ module with three adaptive Dense NN layers. Denoting  $n$  as the number of source environments in each hierarchy layer, the three NN layers have adaptive structures of  $8n$ ,  $4n$  and  $2n$  neurons, respectively. The three NN layers use ReLU activation and He initializer with Uniform distribution.

To integrate representation of different hierarchy layers, the model employs the ‘integ’ module with a Concatenation NN layer and an adaptive Dense NN layer. Denoting the number of source environments in each hierarchy layer as  $n$ , the NN layer has adaptive structures of  $1.5n$  neurons, and uses Tanh activation and Xavier initializer with Uniform distribution.

To estimate according to the integrated representations of different hierarchy layers, the model employs the ‘output’ module with an adaptive Dense NN layer. Denoting the number of source

environments in each hierarchy layer as  $n$ , the NN layer has adaptive structures of  $n$  neurons, and uses Sigmoid activation and Xavier initializer with Uniform distribution.

## Datasets

We used six datasets to assess the utility of EXPERT (Supplementary Table S1). The hierarchy is essentially a refined subset of an ontology (e.g. Environmental Ontology [32] or the Human Disease Ontology [21]) or self-defined according to the context of classification. Refer to Supplementary Notes 3, 4 and Supplementary Table S11 for the unified data processing pipeline used in the study.

For systematic assessment of our GM, the dataset was obtained from MGnify [22], which consists of 118 592 communities collected from 131 biomes. Among them, 52 537 samples originated from human biomes, 14 045 samples originated from mammal biomes, 7189 samples originated from terrestrial biomes and 27 667 samples originated from aquatic biomes. These samples were analyzed by MGnify before January 2020 (Supplementary Table S2). The source environment hierarchy is constructed by referring to the hierarchical biome classification from MGnify and the ecosystem classification paths from the GOLD database [33] (Supplementary Table S12).

For systematic assessment of our HM, the dataset was a part of the first dataset, in which 52 537 communities from 27 human biomes were selected (Supplementary Table S3). The source environment hierarchy is constructed by referring to the hierarchical biome classification from MGnify and the ecosystem classification paths from GOLD.

We also used the newly introduced data in 2020 from MGnify, which consists of 34 209 communities collected from 35 biomes. Throughout the dataset, 3421 samples belonging to eight biomes were newly added by MGnify after January 2020 (Supplementary Table S8). The source environment hierarchy is constructed by referring to the hierarchical biome classification from MGnify and the ecosystem classification paths from GOLD.

For the succession of infant gut microbiome, the dataset was obtained from MGnify, consisting of 392 fecal samples collected from 98 infants and their biological mothers. Among them, 85 infants were born by vaginal delivery and 13 infants were born by cesarean section. The infant samples were collected at three time points including birth, 4 months and 12 months. The maternal samples were collected during the first week after delivery (Supplementary Table S9).

For disease modeling, the dataset was obtained from GMrepo [34], including 13 642 communities collected from feces of hosts diagnosed with 19 diseases as well as healthy controls, Supplementary Table S4). The source environment hierarchy is constructed by referring to NCBI MeSH [35] and Human Disease Ontology.

For cancer monitoring, the dataset was obtained from GMrepo, which consists of 16, 93, 126, 196 and 204 communities, respectively, collected at CRC stages 0, I, II, III and IV, 635 in total (Supplementary Table S10). The source environment hierarchy is constructed by referring to the five stages of CRC.

## Performance measures

To assess the performance of EXPERT models and other methods, we used these measures:

$$TP_b(t) = \sum_s I(\hat{y}_s(b) > t \wedge b \in y_s),$$

$$TN_b(t) = \sum_s I(\hat{y}_s(b) < t \wedge b \notin y_s),$$

$$FP_b(t) = \sum_s I(\hat{y}_s(b) > t \wedge b \notin y_s),$$

$$FN_b(t) = \sum_s I(\hat{y}_s(b) < t \wedge b \in y_s),$$

$$TPR_b(t) = \frac{TP_b(t)}{TP_b(t) + FN_b(t)},$$

$$FPR_b(t) = \frac{FP_b(t)}{FP_b(t) + TN_b(t)},$$

$$Recall_b(t) = \frac{TP_b(t)}{TP_b(t) + FN_b(t)},$$

$$Precision_b(t) = \frac{TP_b(t)}{TP_b(t) + FP_b(t)},$$

where  $TP$  is true positive,  $TN$  is true negative,  $FP$  is false positive,  $FN$  is false negative,  $y_s(b)$  is the quantified contribution from a biome source  $b$  for a microbial community sample  $s$ , threshold  $t \in [0, 1]$  with a step size of 0.01,  $y_s$  is a set of actual biomes for a sample  $s$  and  $I$  is a logical operation function; the value of  $I$  is 1 when the result of logical operation is TRUE, else 0.

Then, two evaluation metrics (F-max and AUROC) were introduced. F-max stands for the maximal F1-measure and was calculated with the following formula. AUROC stands for the area under the ROC and was calculated using the composite trapezoidal rule:

$$Fmax_b(t) = \max_t \frac{2Precision_b(t)Recall_b(t)}{Precision_b(t) + Recall_b(t)}$$

Finally, we treated the average performance across all biomes as the performance of the entire model. Notably, in the section 'Efficiency, accuracy and adaptivity of EXPERT', we only considered biomes with the number of samples >100 to compute the average performance for the GM, the independent model, Transfer (GM) model, and Transfer (GM0) model.

## Evaluating fundamental models

We assessed each model of the fundamental models through cross-validation, and selected the best model among all trained models as the final model.

We assessed the GM by applying 8-fold cross-validation to the 125 823 microbial community data collected from 132 biomes, and selected the best model among the eight trained models as the GM to be transferred.

We assessed the HM by applying repetitive cross-validation (90% as sources to train a model, the rest 10% as queries to test its performance, repeated for five times) to the 52 537 microbial community data collected from 25 biomes, and selected the best model among the five trained models as the GM to be transferred.

The assessment of the DM is the same as the assessment of the HM, but using another dataset consists of 13 462 gut microbial communities associated with 19 diseases.

## Experiment design

We compared EXPERT's performance with FEAST and the NN approach using the human-associated dataset (Supplementary Tables S1, S3). We measured the running time using the Linux command 'time' and considered the real-time usage for comparison. The efficiency was then calculated using the running time we measured. Refer to Supplementary Note 5 for detailed comparison procedure for each experiment.

We demonstrated EXPERT's utility in three contexts. In these contexts, we used standard hyperparameters for training the model (Supplementary Note 1). Detailed descriptions are provided in Supplementary Note 6.

## Statistical analysis

Statistical analyses of the contributions have been performed utilizing the Wilcoxon test, at the significance level of 0.05. For all the tests, when the *P*-value associated is lower than the significance level, one should reject the null hypothesis  $H_0$ , and accept the alternative hypothesis  $H_a$ .

## Visualization of data distribution

Throughout the paper, the box-plot elements are center-line, median; box limits, upper and lower quartiles; whiskers,  $1.5 \times$  interquartile range; points and outliers. The Violin plot is also used for data distribution analysis, mainly for comparison. The PCoA is also used for data distribution analysis, with ellipses representing a confidential interval of 0.95. The principal coordination is obtained through applying beta diversity measurement (Scikit-bio version 0.5.6, Supplementary Table S5) on the abundance of all taxa in seven ranks, namely Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus and Species. The source code of the PCoA analysis is hosted on GitHub at <https://github.com/AdeBC/UniPCoA>.

### Key Points

- We developed the context-aware method EXPERT, which employs TL technique to facilitate microbial community classification in diverse contexts.
- EXPERT could balance the tradeoff between accuracy and efficiency, which could simultaneously reach higher accuracy and efficiency.
- EXPERT enables context-aware customized microbial community classification, and potentiates novel microbial knowledge discovery.

## Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

## Authors' contributions

H.C. and K.N. designed the study, conceived of and proposed the idea, designed and developed the framework. H.C., Y.Z., Q.Y., M.C., G.X., N.W., S.H. and X.H. performed the experiments and analyzed the data. H.C., Y.Z., Q.Y., M.C., G.X. and N.W. visualized the data. C.S. and S.W. provided valuable data. H.C., Y.Z., Q.Y., M.C., W.C., L.P.C. and K.N. contributed to editing and proofreading the manuscript. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Funding

National Natural Science Foundation of China (grants 32071465, 31871334 and 31671374) and the China Ministry of Science and

Technology's National Key R&D Program (grant no. 2018YFC0910502).

## Data availability

The collected samples from the MGnify and GMrepo databases were annotated with their associated biomes/phenotypes in Supplementary Tables S2–S4, S8–S10. All the processed data are uploaded and hosted at <https://github.com/HUST-NingKang-Lab/EXPERT-use-cases>.

## Code availability

All source codes have been uploaded to the website at <https://github.com/HUST-NingKang-Lab/EXPERT>. The package of EXPERT is available on PyPi (see <https://pypi.org/project/expert-mst/>). Detailed software and models used in this study are provided in Supplementary Table S5.

## References

1. Turnbaugh PJ, Ley RE, Hamady M, et al. The Human Microbiome Project. *Nature* 2007;**449**:804–10.
2. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 2014;**16**:276–89.
3. Thompson LR, Sanders JG, McDonald D, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63.
4. Gilbert JA, Jansson JK, Knight R. The earth microbiome project: successes and aspirations. *BMC Biol* 2014;**12**:69.
5. Dominguez-Bello MG, De Jesus-Laboy KM, Shen N, et al. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat Med* 2016;**22**:250–3.
6. Thomas S, Izard J, Walsh E, et al. The host microbiome regulates and maintains human health: a primer and perspective for non-microbiologists. *Cancer Res* 2017;**77**:1783–812.
7. Zhang L, Ji L, Liu X, et al. Linkage and driving mechanisms of antibiotic resistome in surface and ground water: their responses to land use and seasonal variation. *Water Res* 2022;**215**:118279.
8. Coban O, De Deyn Gerlinde B, van der Ploeg M. Soil microbiota as game-changers in restoration of degraded lands. *Science* 2022;**375**:abe0725.
9. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 2019;**25**:679–89.
10. Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–62.
11. Reitmeier S, Kiessling S, Clavel T, et al. Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. *Cell Host Microbe* 2020;**28**:258–72.
12. Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 2014;**10**:766.
13. Zha Y, Chong H, Ning K. Microbiome sample comparison and search: from pair-wise calculations to model-based matching. *Front Microbiol* 2021;**12**:642439.

14. Zou S, Fang L, Lee M-H. Dysbiosis of gut microbiota in promoting the development of colorectal cancer. *Gastroenterol Rep* 2018;**6**: 1–12.
15. Huang S, Haiminen N, Carrieri A-P, et al. Human skin, oral, and gut microbiomes predict chronological age. *mSystems* 2020;**5**:e00630-19.
16. Roguet A, Eren AM, Newton RJ, et al. Fecal source identification using random forest. *Microbiome* 2018;**6**:185.
17. Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, random forests, in a bacterial source tracking study. *Water Res* 2010;**44**:4067–76.
18. Knights D, Kuczynski J, Charlson ES, et al. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 2011;**8**:761–3.
19. Shenhav L, Thompson M, Joseph TA, et al. FEAST: fast expectation-maximization for microbial source tracking. *Nat Methods* 2019;**16**:627–32.
20. Zha Y, Chong H, Qiu H, et al. Ontology-aware deep learning enables ultrafast and interpretable source tracking among sub-million microbial community samples from hundreds of niches. *Genome Med* 2022;**14**:43.
21. Schriml LM, Mitraka E, Munro J, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 2019;**47**:D955–62.
22. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* 2020;**48**: D570–8.
23. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl-Data Eng* 2010;**22**:1345–59.
24. Bäckhed F, Roswall J, Peng Y, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 2015;**17**:690–703.
25. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;**71**:8228–35.
26. Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans Inform Theory* 1991;**37**:145–51.
27. Stokholm J, Thorsen J, Blaser Martin J, et al. Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. *Sci Transl Med* 2020;**12**:eaax9929.
28. Roswall J, Olsson LM, Kovatcheva-Datchary P, et al. Developmental trajectory of the healthy human gut microbiota during the first 5 years of life. *Cell Host Microbe* 2021;**29**:765–776.e3.
29. Duvallet C, Gibbons SM, Gurry T, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017;**8**:1784.
30. Müller A, Hundt C, Hildebrandt A, et al. MetaCache: context-aware classification of metagenomic reads using minhashing. *Bioinformatics* 2017;**33**:3740–8.
31. Martino C, Shenhav L, Marotz CA, et al. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat Biotechnol* 2021;**39**:165–8.
32. Buttigieg PL, Pafilis E, Lewis SE, et al. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperability. *J Biomed Semantics* 2016;**7**:57.
33. Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res* 2021;**49**:D723–33.
34. Wu S, Sun C, Li Y, et al. GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res* 2020;**48**:D545–53.
35. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2014;**42**:D7–17.