Environmental Science and Ecotechnology 17 (2024) 100304

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Refining biome labeling for large-scale microbial community samples: Leveraging neural networks and transfer learning

Nan Wang [1], Teng Wang [1], 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

ARTICLE INFO

Article history:
Received 18 January 2023
Received in revised form
9 July 2023
Accepted 22 July 2023

Keywords:
Microbial community
Transfer learning
Sample classification
Environmental scientific research
Novel knowledge discovery

ABSTRACT

Microbiome research has generated an extensive amount of data, resulting in a wealth of publicly accessible samples. Accurate annotation of these samples is crucial for effectively utilizing microbiome data across scientific disciplines. However, a notable challenge arises from the lack of essential annotations, particularly regarding collection location and sample biome information, which significantly hinders environmental microbiome research. In this study, we introduce Meta-Sorter, a novel approach utilizing neural networks and transfer learning, to enhance biome labeling for thousands of microbiome samples in the MGnify database that have incomplete information. Our findings demonstrate that Meta-Sorter achieved a remarkable accuracy rate of 96.7% in classifying samples among the 16,507 lacking detailed biome annotations. Notably, Meta-Sorter provides precise classifications for representative environmental samples that were previously ambiguously labeled as "Marine" in MGnify, thereby elucidating their specific origins in benthic and water column environments. Moreover, Meta-Sorter effectively distinguishes samples derived from human-environment interactions, enabling clear differentiation between environmental and human-related studies. By improving the completeness of biome label information for numerous microbial community samples, our research facilitates more accurate knowledge discovery across diverse disciplines, with particular implications for environmental research.

1. Introduction

With the development of next-generation sequencing technology, the annual volume of sequencing data increased promptly. Utilizing Metagenomics Sequencing, 16S sequencing, and annotation technology, we could obtain the taxonomic structure and abundance of information about microbial communities [1]. The analyses of these microbial data play an important role in environmental protection [2], water pollution monitoring [3], disease diagnosis or prevention [4,5], and other aspects, especially for environmental scientific research. For example, investigations into indoor microbial communities have shed light on the impact of human–environment interactions could on the taxonomic structure patterns of households [6] and newly opened hospitals [7]. Multiple public databases, such as EBI-MGnify (or MGnify) and Qiita, have become repositories for numerous microbiome samples, including those collected from the environment [8-11]. These databases provide automated pipelines for the analysis and archiving of microbiome data, enabling the determination of taxonomic diversity and metabolic potential within these samples. The millions of microbiome samples deposited in these public databases could facilitate the comparison, clustering, and mining of the microbiome data.

Accurate labeling of microbiome samples is crucial for downstream analysis and interpretation of the data. To achieve precise labeling, it's essential for researchers to timely and carefully document the location, time, and other relevant meta-data about the collection site [12,13]. Portable and reliable instruments for real-time measurement of recording the geographic position are also feasible [14]. Moreover, public databases that specify strict data submission standards will filter out inaccurately annotated microbiome sample submissions [8-10]. The ideal microbiome database should prioritize comprehensive, clear, and precise annotations for

* Corresponding author.
E-mail address: ningkang@hust.edu.cn (K. Ning).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ese.2023.100304

all samples, particularly those collected from the environment. However, the current absence of a strict and unified submission standard, combined with the complex sources of samples during the initial database construction, has resulted in many sample annotations being "rough sketches" with imprecise or non-detailed source labels. For instance, a considerable proportion of samples that should have been attributed to diverse microbial categories have instead been broadly classified as "Mixed biome" without further elaboration. These rough sketches primarily consist of three types of improperly annotated samples: un-annotated samples (samples annotated as "Mixed biome"), under-annotated samples (samples with coarse annotations that could be refined), and mis-annotated samples (samples with incorrect annotations). Since the establishment of the database, the total sample size of MGnify has significantly increased, leading to a rising proportion of inaccurately annotated samples, as exemplified by un-annotated samples (Fig. 1). This trend poses serious challenges, which are listed as follows.

Firstly, inaccurate annotations could lead to a substantial proportion of microbiome samples being wasted, particularly those labeled as "Mixed biome", resulting in the exclusion of valuable unannotated samples. This issue is pronounced in environmental research focused on coral tissue, hindering the study of microbial community dynamics and marine coral colony protection (MGYS00003856).

Secondly, inaccurate annotations could result in the misinterpretation of microbiome samples or even research failures. The lack of strict meta-data standards for data submission has led to massive under-annotated samples in databases. For instance, freshwater samples coarsely labeled as "root: Environmental: Aquatic" are probably mistakenly used as marine samples during data mining for marine ecosystem studies. Given the limited availability of robust data-cleaning methods, such misclassifications can yield erroneous conclusions and ultimately compromise the validity of the study.

Thirdly, inaccurate annotations can cause a cascading accumulation of errors when included in secondary databases that link data from different primary databases, such as GM-repo [15]. This leads to an increase in inaccurately annotated samples in secondary databases, making the true origin of the samples unverifiable. With the ever-increasing proportion of samples accumulated in the current microbiome databases (Fig. 1), the aforementioned issues pose a significant obstacle for microbiome research in environmental scientific research. Therefore, a highly intelligent and automatic method that could disentangle and refine the biome information for microbiome samples is desirable.

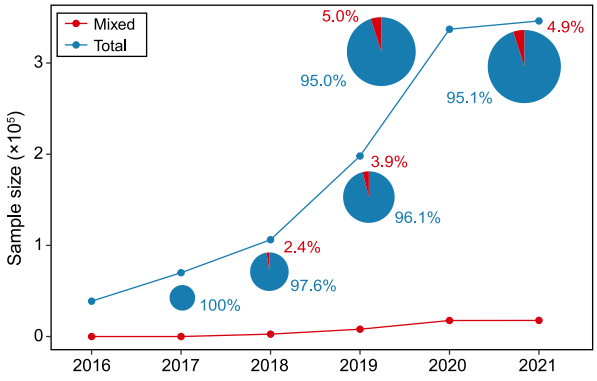

Fig. 1. The increasing number of all samples and those annotated as "Mixed biome" in MGnify. The statistic of the annual amount of all samples and samples annotated as "Mixed biome" in MGnify from 2016 to 2021. The lines show the annual amount of all samples and samples annotated as "Mixed biome". The pie charts show the annual proportion of the samples annotated as "Mixed biome" in all samples.

In this study, we designed Meta-Sorter, a tool to disentangle the biome labels for samples with inaccurate biome annotations, based on a neural network and transfer learning. A neural network model was first constructed based on 94,874 samples introduced into MGnify before January 2020 (existing samples) with detailed biome annotations and showed high robustness and accuracy, with the average area under the receiver operating characteristic (AUROC) curve of 0.896. This model was then applied to all existing samples annotated as "Mixed biome" for their detailed biome prediction, with results showing that 95.41% of samples were consistent with their meta-data. Secondly, due to the reduced accuracy of the neural network model on samples introduced into MGnify after January 2020 (newly introduced samples), we designed a transfer neural network model based on transfer learning, with the classification accuracy (average AUROC) again boosting to 0.989, and 97.62% of newly introduced samples annotated as "Mixed biome" correctly predicted. Combining the results on existing and newly introduced samples, we found that out of 16,507 samples with no detailed biome annotations, 96.65% could be correctly classified, largely solving the missing biome labeling problem. Finally, we assessed the practical application performance of Meta-Sorter on several environmental concrete cases, such as differentiating the actual sources of samples only labeled as "Marine" in MGnify into benthic and water columns and classifying samples from studies that involved human–environment interactions into environment or human. Collectively, we have designed Meta-Sorter as a highly intelligent and automatic method that could disentangle and refine the biome information for microbial samples. Meta-Sorter is thus a useful tool for better classification and knowledge discovery from millions of microbiome samples.

2. Material and methods

2.1. Datasets

We examined 118,592 samples from 1447 studies introduced into MGnify before January 2020 to *ab initio* training the neural network model, 25 studies annotated as "Mixed biome" before January 2020, which included 7941 samples. 34,209 samples newly introduced into MGnify after January 2020, which belonged to 32 studies, were applied for implementing transfer learning to the neural network model to generate the transfer neural network model, and we also introduced ten studies annotated as "Mixed biome" after January 2020 which included 10,862 samples.

2.2. Process of model construction and transfer learning

Meta-Sorter is based on the ontology-aware neural network model combined with transfer learning.

The inputting files of Meta-Sorter: the biome ontology, which is a hierarchical structure that represents the taxonomic hierarchy of the samples, for instance, the biome "root: Host-associated: Human: Digestive system: Large intestine: Fecal" is represented in the ontology as layer-1 represents "root", layer-2 represents "Host-associated", …, layer-6 represents "Fecal"); the taxonomic structures for microbial community samples with detailed biome information, which are generally obtained by 16S sequencing or Metagenomics Sequencing and annotated by software, such as QIIME [12] and KRAKEN [16].

Data processing process. First, import the microbial data (the biome ontology and microbial taxonomic structures), establish a mapping relationship between the taxonomic profiles and the

phylogenetic tree, construct a regular abundance matrix, and then standardize the abundance matrix and convert it into a relative abundance matrix. Then, the relative abundance matrix is standardized by Z-score to become a standard abundance matrix.

The modeling process of ontology-aware neural network based on the standard abundance matrix. The neural network (NN) consists of four modules: (1) the "base" module is to obtain the features of the input standard abundance matrix on the low level, (2) the "inter" module, which contains three Dense NN layers, is to obtain the features of different hierarchy layers, (3) the "integ" module, which contains a concatenation NN layer and a Dense NN layer, is to integrate the features of different hierarchy layers, and (4) the "ouput" module, which contains a Dense NN layer, is to estimate the contribution of each source according to the integrated representations of different hierarchy layers. During forward propagation, the representation of each lower layer is integrated into the corresponding higher layer using multiple "integ" modules, which establish the initialization parameters between neural network layers. Backward propagation involves optimizing the parameters of the entire model. This is achieved through the utilization of gradient descent coupled with the backpropagation algorithm, enabling the solution of the model's parameters.

Transfer learning process. The existing model could be divided into bottom- and top-level nodes. The bottom-level nodes have the potential to be applied in emerging datasets, while the top-level nodes can only be applied to the existing datasets. The transfer learning process involves three steps for effective knowledge transfer. Firstly, the lower-level nodes are locked to ensure their exclusion from the transfer process, then the new community structure is encoded and introduced, accompanied by modifications to the structure and weights between the higher-level nodes. This process, referred to as "Transfer", aims to facilitate the incorporation of microbial data from the new community. The forward and backward algorithms are then applied iteratively to update the parameters of higher-level nodes until convergence is achieved. Subsequently, the optimized higher-level nodes are deployed to analyze new datasets, referred to as the "Fast Adaptation" process. In the final stage, the bottom-level nodes are unlocked, and the parameters of these nodes are updated iteratively using the new microbial data and the training of forward and backward algorithms. This phase is referred to as "Fine-tuning" (See Fig. S1 for details).

2.3. Performance measures

For the area under the receiver operating characteristic (AUROC) curve, the area under the precision−recall (AUPR) curve, and Maximum F-score (F-max) computation, we set the threshold from 0 to 1 with a step size of 0.01. The result of the logical operation is 1 if the contribution of the node is greater than the threshold, else 0. We calculated True Positive, True Negative, False Positive, and False Negative for calculating True Positive Rate, False Positive Rate, Precision, Recall, and F1-measure at every threshold, and then we obtained the AUC curve and PR curve. By calculating the area under the AUC curve as AUROC and the area under the PR curve as AUPR of each node, F-max stands for the maximal F1-measure. Each node represents an ecological classification of a community.

2.4. Assessments of Meta-Sorter

We accessed the neural network model by applying five-fold cross-validation to the 118,592 samples collected from 134 biomes. The neural network model with the best performance was used for Meta-Sorter.

We accessed the transfer neural network model and the independent neural network model by applying five-fold cross-validation (80% samples as the source to construct the independent neural network model and to implement transfer learning to the neural network model for generating the transfer neural network model, the rest 20% samples were used to assess the performance of the models) to the 34,209 samples from 32 studies, the transfer neural network model with the best performance was used for Meta-Sorter.

3. Results and discussion

3.1. The workflow of Meta-Sorter

Meta-Sorter has a neural network model constructed based on 118,592 microbial samples from 134 biomes and their biome ontology. Notice that in this study, the 118,592 samples with detailed biome information used here were those deposited into the MGnify database before January 2020 (existing samples) (Fig. 2a and Table S1). Moreover, to adapt the neural network model to the newly introduced samples, part of which were probably from new biomes, we introduced transfer learning (Fig. S1) to Meta-Sorter. 34,209 newly introduced microbial samples from 35 biomes (including eight new biomes) (Table S1) and their biome ontology were applied in the transfer learning process to generate the transfer neural network model. During the transfer learning process, the parameters and structures of an existing neural network model could be updated, and the resulting transfer neural network model was suitable for newly introduced samples. Notably, the 34,209 newly introduced samples used here for building the transfer neural network model were those deposited into the MGnify database after January 2020 (newly introduced samples). With the neural network model and the transfer neural network model, Meta-Sorter could decode the samples' biome labels, which were annotated as "Mixed biome", into detailed biome labels (Fig. 2b). Besides, Meta-sorter could refine the biome labels to obtain more valuable information for reference (Fig. 2c) and correct the mis-annotated samples' labels to avoid cascading accumulation (Fig. 2d), as well as other applications, such as classifying the actual sources of ancient DNA.

3.2. The neural network model decoded the biome information for un-annotated samples accurately

The neural network model worked well on different layers of the biome ontology. To disentangle the samples annotated as "Mixed biome", we need a prediction model that covers as many biomes as possible. Here, we chose 118,592 samples (Table S1), which included 134 biomes and were deposited into the MGnify database before January 2020 (existing samples) to generate a neural network model. The neural network model's benchmark revealed that the average AUROC, AUPR, and F-max are 0.89, 0.76, and 0.73, respectively. Though the average AUROC on each layer of the biome ontology decreased slightly as the layer increased, the prediction accuracy on each layer exceeded 0.99, indicating the robustness of the neural network model (Fig. 3a)**.** Therefore, the neural network model worked well on classifying the existing samples annotated in detail and covered a comprehensive set of biomes, making it feasible to decode the biome labels for samples without detailed biome information.

Meta-sorter based on the neural network model decoded the samples' biome labels annotated as "Mixed biome" into detailed biome labels. We assessed the performance of Meta-Sorter on predicting the detailed source biome for samples annotated as "Mixed biome". We examined 7941 existing samples from 25 studies annotated as "Mixed biome" and utilized the neural

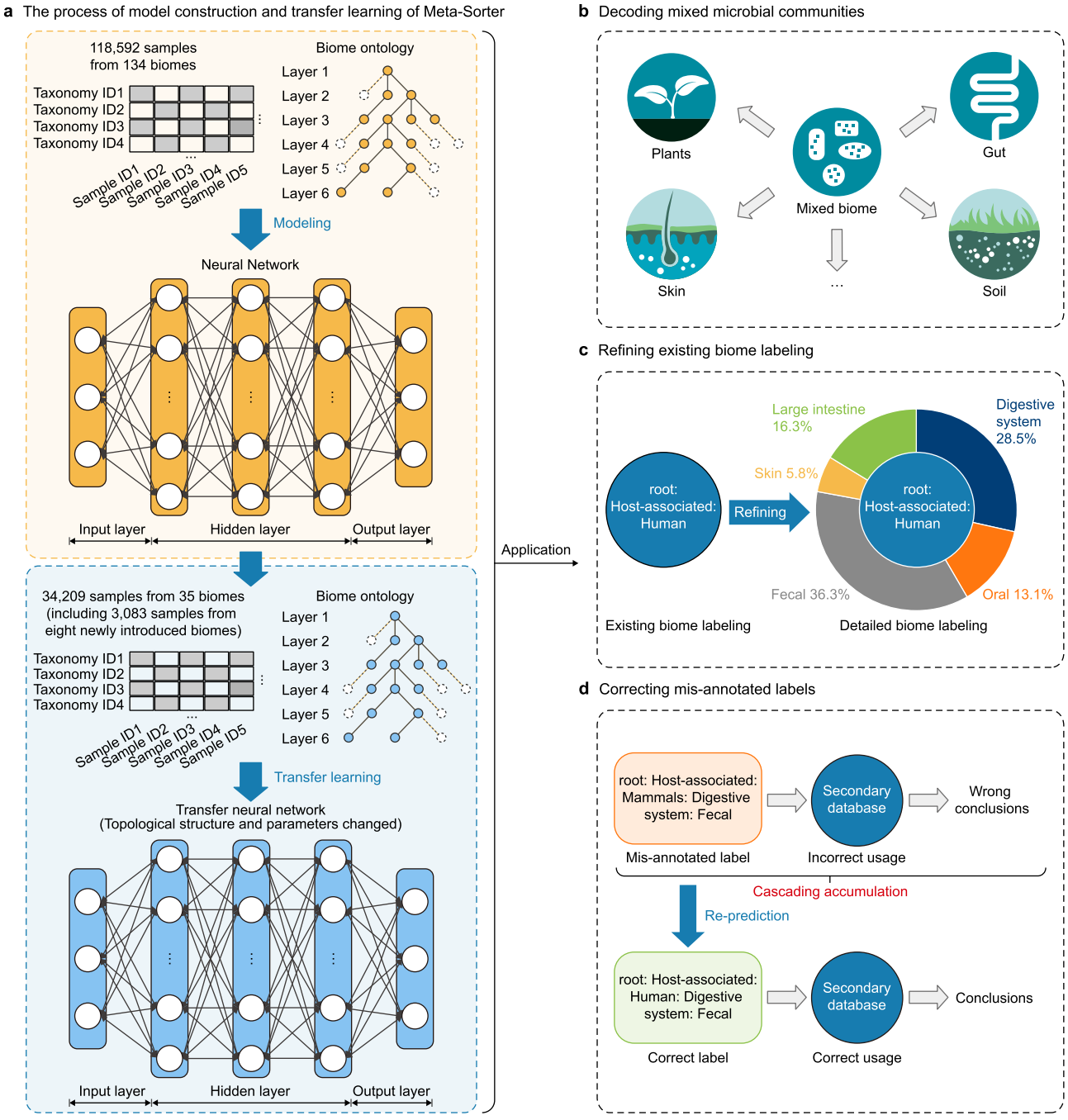

Fig. 2. The rationale and applications of Meta-Sorter. **a**, The process of model construction and transfer learning of Meta-Sorter. Two input files, biome ontology and samples' taxonomic structures with detailed biome information, are required in model construction and transfer learning. The yellow box shows that the neural network model was constructed based on 118,592 existing samples with detailed information on 134 biomes and their biome ontology. The blue box shows that the transfer neural network model was constructed using 34,209 newly introduced samples from 35 biomes (including 3083 samples from eight newly introduced biomes) and transfer learning to the existing neural network model. **b**–**d**, The applications of Meta-Sorter. Meta-Sorter decoded the samples' biome labels annotated as "Mixed biome" into detailed biome labels (**b**). Meta-sorter refined the biome labels in more detail to obtain more valuable information for reference (**c**). Meta-Sorter corrected the mis-annotated samples' labels to avoid cascading accumulation (**d**).

network model to predict their actual sources (manually curated documented source biome labels as reported in the literature) (Tables S2 and S3). By comparing with the actual sources of these samples, we found that a high proportion of biomes predicted by the neural network model were consistent with the actual sources, up to 95.41% of the prediction results of the chosen samples were in line with the original reference information (Fig. 3b). Moreover, since the neural network model covered comprehensive biomes, it can be applied to disentangle the biome information for samples of a wide range.

Case studies on decoded samples from "Mixed biome". We then focused our examination and in-depth analysis on several representative environmental sets of samples. McCall et al. [17] collected microbial samples from different rooms in four human

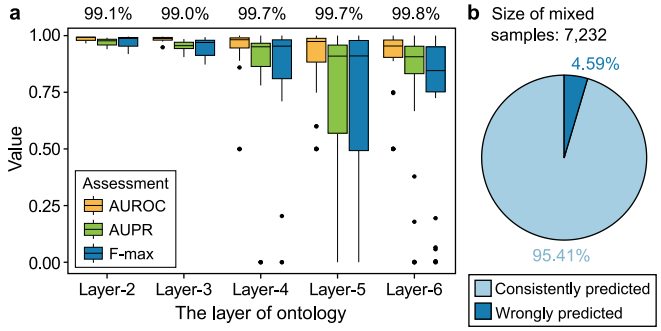

Fig. 3. The neural network model benchmark and the decoding accuracy of the "Mixed biome" labels. **a**, 118,592 existing samples were randomly divided into training subsets (80%, 94,874 samples) and testing subsets (20%, 23,718 samples). The source biome annotation for samples of the testing subset was predicted by the neural network model. The boxplots represent AUROC, AUPR, and F-max of the neural network model for source biome annotation, categorized by different layers of the biome ontology. The percentages above the boxplots represent the prediction accuracy for each layer of the biome ontology. **b**, 7941 existing samples annotated as "Mixed biomes" were predicted by the neural network model of Meta-Sorter, of which 7232 samples had reference information in the original literature. Results were based on manually comparing the predicted labels with reference information in the original literature, marked as consistent predicted if the predicted labels were consistent with reference information, and marked as wrongly predicted if not consistent.

environments with different urbanization levels and recorded meta-data to explore the role of microbiome structure in urbanization and migration. Despite the high quality and research value, uploaded samples were classified as "Mixed biome" in MGnify, and the users needed to access more channels to obtain the samples' meta-data for re-analysis, causing inconvenient access to these samples or even study failure.

We used Meta-Sorter based on the neural network model to disentangle the detailed biome for samples in this built-environment study. By comparing the prediction results (Table S4) with their actual biomes [17,18], we found that despite the presence of non-negligible differences, the categories of the sample sources were substantially consistent (Fig. 4a, b), which showed the prediction results by Meta-Sorter were rational and

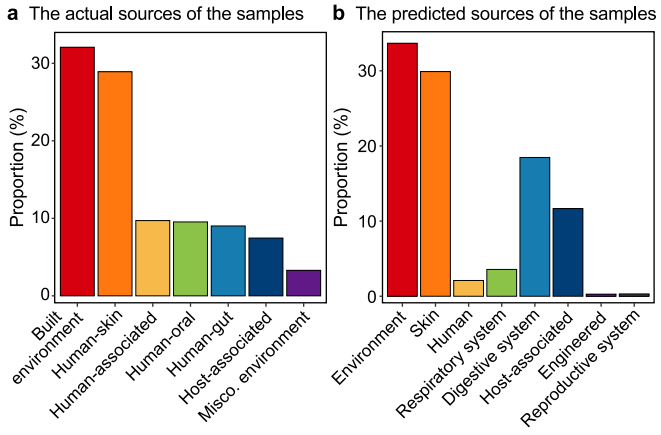

Fig. 4. Comparison of predicted sources and actual sources for samples in the case study on microbial communities in a built environment for decoding samples from "Mixed biome". **a**, The composition of actual sources and their proportion (%): build environment (32.08), human-skin (28.91), human-associated (9.71), human-oral (9.54), human-gut (9.02), host-associated (7.46), and misc. environment (3.29). **b**, The composition of predicted sources and their proportion (%): environmental (33.67), skin (29.9), digestive system (18.47), host-associated (11.67), respiratory system (3.57), human (2.10), reproductive system (0.31), and engineered (0.28).

included interesting details listed below.

Firstly, the original descriptions lacked refinement and accuracy, while Meta-Sorter provided more detailed and concrete annotations. The samples predicted as "Root: Environmental: Aquatic" accounted for the highest proportion of those predicted as "Root: Environmental", and the original description of these samples was "indoor genome", which belonged to "root: Environmental". Recognizing this issue, the researchers in this study used a Bayesian approach called SourceTracker [19] to estimate the source environments for each group of samples, with results showing that water source, such as domestic water, was the source of a large proportion of the microorganisms, which was highly consistent with Meta-sorter's results and further validated our model.

Secondly, Meta-Sorter predicted more samples from "Root: Host-associated: Human: Digestive system" and fewer samples from "Root: Host-associated: Human: Respiratory system", the most notable variation in prediction results. This finding was reasonable and reliable since oral is a component of the respiratory and digestive systems of humans, and the samples collected from these sites have a high degree of consistency.

More importantly, Meta-Sorter could assign samples to biome labels intelligently and automatically, with the biome labels predicted by Meta-Sorter not included in the manually pre-defined set of biome labels used by SourceTracker. For example, Meta-Sorter identified a fraction of samples sourced from the "root: Host-associated: Human: Reproductive system", which was not included in the original sample description but was reasonable based on the literature indicating some samples were collected from bathroom floors and walls using sterile swabs [18]. This demonstrates Meta-Sorter's potential to identify human interference in environmental studies, enhancing the reliability of results.

3.3. *Transfer learning enabled the decoding of the biome information for newly introduced un-annotated samples*

The limitation of applying the neural network model on newly introduced microbial samples. Though the neural network could perform exceptionally well on disentangling the biome information for un-annotated samples, it should be admitted that all these un-annotated samples were already in the database (existing samples), and their detailed biome information was already known by the neural network model. As the microbial samples in the database accumulated, it would be intriguing if Meta-Sorter could be used to disentangle the biome information for these newly introduced samples. However, the neural network model was only aware of the existing biome information, which was insufficient for such a purpose.

We observed that numerous microbial samples were newly introduced into MGnify after January 2020 (newly introduced samples) (Fig. 1), we then examined if the neural network model of Meta-Sorter could be utilized to predict the actual sources for these samples. Here, we have collected 32 newly introduced studies, which included 34,209 samples annotated with detailed information and 10,862 samples from ten studies annotated as "Mixed biome". Meta-Sorter based on the neural network model has been applied to the 34,209 samples annotated with detailed information and compared with the independent neural network model, which was constructed solely on these 34,209 samples. The average AUROC of the neural network model is 0.872 while the independent neural network model is 0.989 ($P_{Wilcox} = 2.22 \times 10^{-16}$; Fig. 5a), a pivotal reason for which was that the distribution between existing samples and newly introduced samples had significant differences (Fig. 5b). This phenomenon restricted the applicability of existing neural network model on the newly introduced samples.

However, as more and more samples annotated as "Mixed

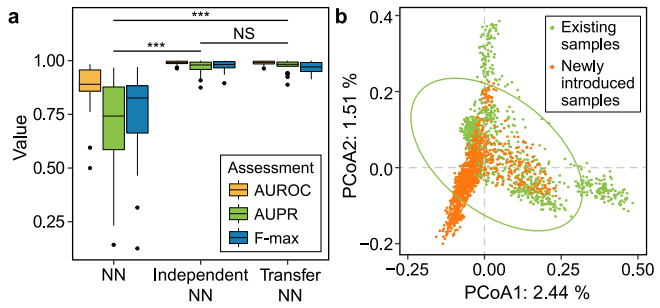

Fig. 5. The benchmark of different models and the heterogeneity of existing samples and newly introduced samples. **a**, Comparison of different models in source biome annotation for newly introduced samples. 34,209 newly introduced samples with detailed biome information were randomly divided into training subset (80%, 27,355 samples) and testing subset (20%, 6854 samples), the samples of testing subset were predicted by NN model, independent NN model, and transfer NN model, respectively. The boxplots represent the AUROC, AUPR, and F-max of the three neural-network models used for source biome annotation. NS, not significant; ***$P < 0.005$; Mann-Whitney-Wilcoxon test. **b**, The different distribution of existing samples and newly introduced samples. The confidence level is 95%. NN, neural network.

biome" would be introduced into public databases like MGnify (Fig. 1), it's desirable to find out a solution for exceeding the effect of the heterogeneity of datasets and disentangling the biome information of those samples in "Mixed biome". To solve this problem, we introduced transfer learning to update the existing neural network model and generated a transfer neural network model which could be utilized for disentangling the biome information for these newly introduced microbial samples, which were annotated as "Mixed biome" (Fig. 2a and Fig. S1). The average AUROC of the transfer neural network model was 0.989, outperformed the neural network model ($P_{Wilcox} = 2.22\times10^{-16}$), and was as good as the independent neural network model ($P_{Wilcox} = 0.7$), with the consistent results of AUPR and F-max (Fig. 5a). This result indicated that transfer learning could efficiently exceed the limitation for the application of neural network model caused by heterogeneity between the existing samples and the newly introduced samples.

Transfer learning enhanced the adaptability of the neural network model to the newly introduced biome. In addition to the newly introduced samples, which only include the existing biomes, there are a considerable number of newly introduced biomes, despite the obvious fact that the neural network model was not suitable for classifying these samples from new biomes. The first and foremost reason is that remodeling in each newly introduced biome is unrealistic due to the enormous computing resources and time required. There's another reason if the amount of data newly added to the database in a certain year is rather small or when the researchers aim to excavate specialized data (e.g., marine microbial data mining), the independent neural network model cannot be adequately trained due to the limited size of the dataset. In these contexts, a transfer learning scheme is suitable since less time and computing resources are required to generate a reliable transfer neural network model. Therefore, we applied transfer learning to adapt the neural network model to 34,209 newly introduced samples from 35 biomes (including 3083 samples from eight newly introduced biomes) to generate the transfer neural network model (Fig. 2a, 6a, and Fig. S1).

We assessed the AUROC and AUPR of the transfer neural network model predicting the newly introduced biomes and found that the transfer neural network model worked well on those biomes (Fig. 6b and c). Plenty of cases support the superiority of the transfer neural network model. For example, we noticed a new biome out of the 35 biomes, which were annotated as "root: Host-associated: Birds: Digestive system: Ceca". However, only "root: Host-associated" was introduced in the neural network model, which was obviously not suitable for those newly introduced biomes, while by implementing transfer learning to the neural network model with samples in the new biome, the transfer neural network model had high prediction accuracy (AUROC = 0.999 and AUPR = 0.868) on the biome. In short, transfer learning enhanced the adaptability of the neural network model to the newly introduced biomes, and the robustness of the model has been improved.

Prediction of the detailed source biome for samples annotated as "Mixed biome" from newly introduced studies. Apart from samples with detailed biome labels, there were a proportion of newly introduced samples without biome information and annotated as "Mixed biome". Due to differences among existing and newly introduced samples, the neural network model may not be appropriate for decoding their biome labels. Therefore, we assessed the accuracy of Meta-Sorter based on the transfer neural network model on predicting the detailed source biome for newly introduced samples annotated as "Mixed biome". We chose 10,862 newly introduced samples from ten studies annotated as "Mixed biome" (Table S2), and utilized the transfer neural network model to disentangle the detailed biomes for these samples, then compared them with their actual sources from the original literature (Table S5). We found that 97.62% of these samples were assigned to detailed biomes, consistent with their actual sources (Fig. 7a), indicating that the transfer neural network model could effectively disentangle the biome information for those samples without detailed annotations.

Combined the results on existing and newly introduced samples, we found that Meta-Sorter based on the neural network model and transfer neural network model could largely solve the missing biome labeling problem: 16,507 samples out of 18,803 total samples had reference information in the original literature, of which 96.65% (15,954 samples) could be consistently predicted by Meta-Sorter (Fig. 7b). This demonstrated the effectiveness of Meta-Sorter.

3.4. Meta-Sorter refined the biome annotation for under-annotated and mis-annotated samples

In addition to decoding samples labeled as "Mixed biome", Meta-Sorter can also classify data too high in the classification hierarchy to be refined. In an environmental study, "anchialine metagenome raw sequence reads" (MGYS00005510), researchers collected benthic and water column samples from nine anchialine habitats in the Hawaiian Archipelago and identified environmental factors driving microbial diversity [20]. Despite the arduous sample-collecting procedure and genomes information extremely exploitable for microbiomes and metagenomics researchers, all samples were labeled as "Root: Environmental: Aquatic", and even the sample descriptions lacked precise site information. We utilized Meta-Sorter based on transfer learning to refine the label of 250 samples (Table S6) in this study, with results showing as follows:

Firstly, Meta-Sorter successfully deciphered the sample information for more refined classification labels. 85.6% of the predicted results from Meta-Sorter reached the classification of layer-5 or layer-6 (Fig. 8a), while the original classification label for all samples was "root: Environmental: Aquatic", which may provide added offer additional information to other researchers when mining data at finer classification levels.

Secondly, Meta-Sorter's rational results indicated extra clues for this study. The sampling environment is anchialine open pools or ponds and caves, which are part of the marine/anchialine ecosystem, whereas Meta-Sorter predicted that 90% of the samples would be classified as "root: Environmental: Aquatic: Marine". The original work has divided the samples into water columns and

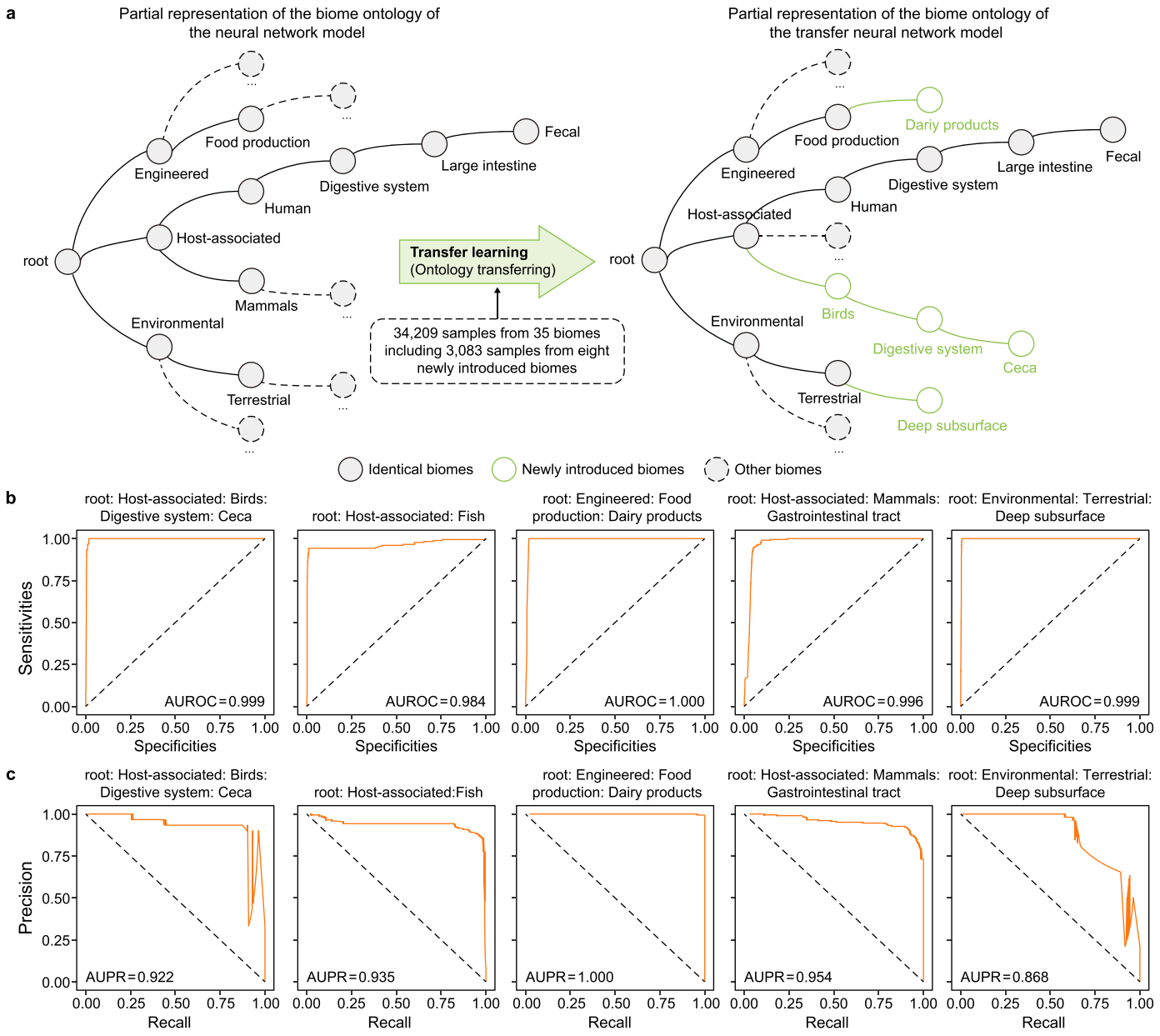

Fig. 6. The robust adaptation of Meta-Sorter to the newly introduced samples. **a**, The left panel shows the partial representation of the biome ontology of existing samples used to construct the neural network model, and the right panel shows the partial representation of the biome ontology of the newly introduced samples used for transfer learning to generate the transfer neural network model. The existing samples included 118,592 samples from 134 biomes, and the newly introduced samples included 34,209 samples from 35 biomes (including eight newly introduced biomes, e.g., bird-related biomes). **b**–**c**, The AUROC (**b**) and AUPR (**c**) assessments of Meta-Sorter on representative newly introduced biomes.

benthic categories, and Meta-Sorter predicted that 26% of the samples correspond to the "water column" and 25% correspond to the "benthic" categories (Fig. 8a). Collectively, the aforementioned results demonstrate the validity of our predictions.

Interestingly, Meta-Sorter provided insights into the similarities between water columns and benthic communities. A significant proportion of predictions (39%) belong to "root: Environmental: Aquatic: Marine: Intertidal zone", which denotes the area above water level at low tide and underwater at high tide. This high variability in environment and biome is attributed to the neritic, deep zones, and seabed biomes [21,22]. In the NMDS analysis, the researchers found that both water columns and benthic communities had higher numbers of shared communities and limited variation in structure at the same sampling site (Fig. 8b). This was most likely due to the substantial proportion of samples from the "intertidal zone" in both taxa. Further analysis by excluding or categorizing this sample part may yield additional insights.

Finally, Meta-Sorter's predictions may indicate biome migration or labeling errors due to human factors. 1% of the predictions were identified as "host-associated" and 9% as non-marine environments, such as "Freshwater Lake" and "Non-marine Saline and Alkaline", which can be partially attributed to the sampling sites. Furthermore, the biomes predicted as the label "host-associated" and other environments may have some tendency to have inhabited both ecological niches, indicating community migration between diverse environments and hosts [23]. In addition, there is a non-negligible possibility that this inconsistency is due to human factors such as contamination of the samples resulting in labeling

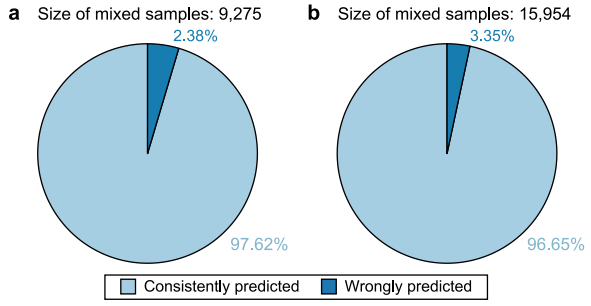

Fig. 7. The accuracy of Meta-Sorter on decoding samples annotated as "Mixed biome". **a**, 10,862 newly introduced samples annotated as "Mixed biomes" were predicted by the transfer neural network model of Meta-Sorter, of which 9275 samples had reference information in the original literature. We manually compared the predicted labels with reference information in the original literature and marked them as consistent predicted if the predicted labels were consistent with reference information, while marked as wrongly predicted if not consistent. **b**, The general prediction accuracy of Meta-Sorter on the samples in "Mixed biome", including those from both existing samples and newly introduced samples. Meta-Sorter predicted 18,803 samples annotated as "Mixed biome" based on the neural network model and the transfer neural network model, of which 16,507 samples had reference information in the original literature.

errors, highlighting Meta-sorter's potential to resolve the mis-annotation problems.

It has become an increasingly difficult problem for data mining from publicly available microbiome samples in environmental scientific research, largely due to three kinds of inaccurately annotated samples (un-annotated, under-annotated, and mis-annotated samples), which has led to the inability for biomarker discovery or even cascading accumulation of errors. In this study, we have designed Meta-Sorter, a neural network and transfer-learning-enabled AI method for improving the biome labeling of thousands of microbial community samples with inaccurate biome information. By establishing a comprehensive neural network model and introducing a transfer neural network model, the problems caused by inaccurately annotated samples were largely solved. Results have shown that out of 16,507 samples with no detailed biome annotations, 15,954 (96.65%) could be consistently predicted, largely solving the biome labeling problem for inaccurately annotated samples. Interestingly, Meta-Sorter was able to assign samples to biome labels in a highly intelligent and automatic manner, providing insights beyond the original literature. In addition to environmental scientific research, Meta-Sorter has a broad spectrum of applications, such as indicating the potential of microbial community migration, disentangling the hidden information of ancient microbial community samples, and detecting sample contamination.

Meta-Sorter could be further improved when more microbial community samples and their meta-data are accumulated. With the accumulation of samples from more diverse biomes, Meta-Sorter's model could include more biome information for more accurate biome labeling. Additionally, for under- and mis-annotated samples, mining Meta-Sorter's results might reveal intricate but important connections among samples. Finally, the idea of Meta-Sorter using both neural networks and transfer learning for sample annotation might be expanded to other domains of biological data mining in a wide range of contexts, such as gene mining.

4. Conclusions

Our proposed Meta-Sorter has made advancements in improving the completeness of biome label information for tens of

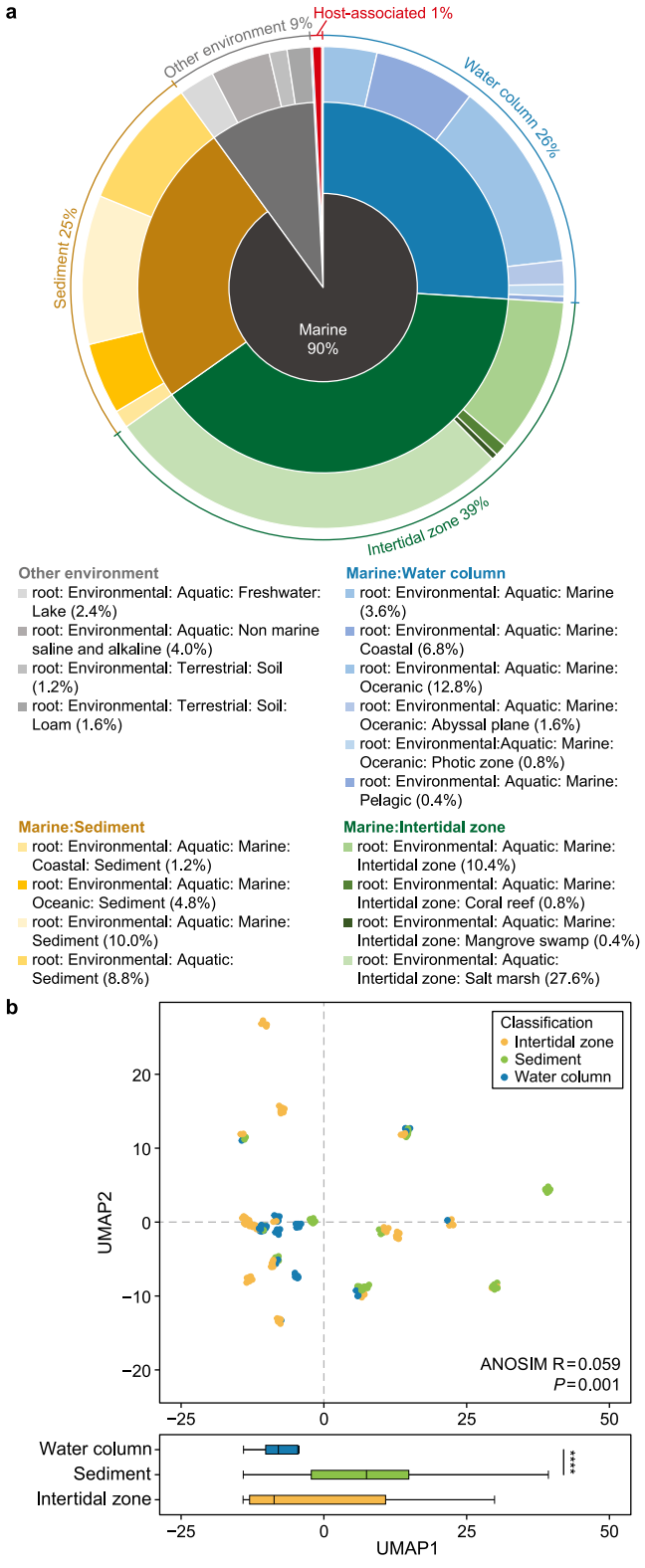

Fig. 8. Meta-Sorter refined the source labels of representative samples from a case study on marine samples. **a**, The detailed biome contribution predicted by Meta-Sorter. The results were manually divided into three sources: water column, intertidal zone, and sediment. More detailed sources and proportions were as well as shown under the three sources. **b**, The distribution of the samples with the predicted three-category source labels (water column, sediment, and intertidal zone) as the samples' actual sources (ANOSIM, R = 0.059, P = 0.001). The box plot showed the distribution of samples on the x-axis, demonstrating substantial differences between the water column and sediment sample. **** $P_{Wilcoxon} < 0.0001$.

thousands of microbial community samples. This has solved the problem of the cascading accumulation of errors, facilitating a wide range of applications, including sample classification, source tracking, and novel knowledge discovery from millions of microbiome samples. Although this AI approach effectively improves microbial community labeling, we acknowledged that hundreds of samples still require improvement. To ensure the continuous optimization of accurate annotation for microbiome samples, it is essential to employ additional effective strategies. These strategies may include establishing standardized protocols for sample collection and labeling, implementing positioning methods, such as Global Positioning System, for quality control of geographic location labeling, incorporating supplementary meta-data information, and developing strict and uniform database data submission standards. In conclusion, our AI-assisted approach has improved the accuracy of microbial community labeling, resulting in a significant improvement in discovering microbiome knowledge in multiple disciplines, particularly in environmental scientific research.

Credit authorship contribution statement

Nan Wang: Methodology, Software, Validation, Data Curation, Writing - Original Draft, Visualization. **Teng Wang:** Software, Validation, Writing - Original Draft, Visualization. **Kang Ning:** Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Data availability

The datasets analyzed during the current study are publicly available in EBI-MGnify (https://www.ebi.ac.uk/metagenomics). The accession numbers are included in Table S1.

Codes and models availability

The pre-trained models and source codes of scripts for training, querying, and transfer learning are publicly available at https://github.com/HUST-NingKang-Lab/Meta-Sorter. The program "EXPERT" is available at https://github.com/HUST-NingKang-Lab/EXPERT.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the 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). We thank Hui Chong for the technical assistance in the transfer learning process. Numerical computations were performed at the Hefei Advanced Computing Center.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2023.100304.

References

[1] A. Escobar-Zepeda, A. Vera-Ponce de León, A. Sanchez-Flores, The road to metagenomics: from microbiology to DNA sequencing technologies and Bioinformatics, Front. Genet. (2015) 6.
[2] M. Bahram, F. Hildebrand, S.K. Forslund, J.L. Anderson, N.A. Soudzilovskaia, P.M. Bodegom, J. Bengtsson-Palme, S. Anslan, L.P. Coelho, H. Harend, J. Huerta-Cepas, M.H. Medema, M.R. Maltz, S. Mundra, P.A. Olsson, M. Pent, S. Pölme, S. Sunagawa, M. Ryberg, L. Tedersoo, P. Bork, Structure and function of the global topsoil microbiome, Nature 560 (7717) (2018) 233–237.
[3] Z. Wang, M. Han, E. Li, X. Liu, H. Wei, C. Yang, S. Lu, K. Ning, Distribution of antibiotic resistance genes in an agriculturally disturbed lake in China: their links with microbial communities, antibiotics, and water quality, J. Hazard Mater. 393 (2020) 122426.
[4] C.A. Gaulke, T.J. Sharpton, The influence of ethnicity and geography on human gut microbiome composition, Nat. Med. 24 (10) (2018) 1495–1496.
[5] J. Halfvarson, C.J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W.A. Walters, L.M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E.E. McClure, M.F. Dunklebarger, R. Knight, J.K. Jansson, Dynamics of the human gut microbiome in inflammatory bowel disease, Nat. Microbiol. 2 (5) (2017) 17004.
[6] S.T. Kelley, J.A. Gilbert, Studying the microbiology of the indoor environment, Genome Biol. 14 (2) (2013) 202.
[7] S. Lax, N. Sangwan, D. Smith, P. Larsen, K.M. Handley, M. Richardson, K. Guyton, M. Krezalek, B.D. Shogan, J. Defazio, I. Flemming, B. Shakhsheer, S. Weber, E. Landon, S. Garcia-Houchins, J. Siegel, J. Alverdy, R. Knight, B. Stephens, J.A. Gilbert, Bacterial colonization and succession in a newly opened hospital, Sci. Transl. Med. 9 (391) (2017) eaah6500.
[8] A. Gonzalez, J.A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A.D. Swafford, S.B. Orchanian, J.G. Sanders, J. Shorenstein, H. Holste, S. Petrus, A. Robbins-Pianka, C.J. Brislawn, M. Wang, J.R. Rideout, E. Bolyen, M. Dillon, J.G. Caporaso, P.C. Dorrestein, R. Knight Qiita, rapid, web-enabled microbiome meta-analysis, Nat. Methods 15 (10) (2018) 796–798.
[9] K.P. Keegan, E.M. Glass, F. Meyer, MG-RAST, a metagenomics service for analysis of microbial community structure and function, Methods Mol. Biol. (2016) 207–233.
[10] A.L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M.R. Crusoe, V. Kale, S.C. Potter, L.J. Richardson, E. Sakharova, M. Scheremetjew, A. Korobeynikov, A. Shlemov, O. Kunyavskaya, A. Lapidus, R.D. Finn, MGnify: the microbiome analysis resource in 2020, Nucleic Acids Res. 48 (D1) (2020) D570–D578.
[11] E.W. Sayers, J. Beck, E.E. Bolton, D. Bourexis, J.R. Brister, K. Canese, D.C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T.L. Madden, N. O'Leary, L. Phan, S.H. Rangwala, V.A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the national center for biotechnology information, Nucleic Acids Res. 49 (D1) (2021) D10–D17.
[12] E. Bolyen, J.R. Rideout, M.R. Dillon, N.A. Bokulich, C.C. Abnet, G.A. Al-Ghalith, H. Alexander, E.J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J.E. Bisanz, K. Bittinger, A. Brejnrod, C.J. Brislawn, C.T. Brown, B.J. Callahan, A.M. Caraballo-Rodríguez, J. Chase, E.K. Cope, R. Da Silva, C. Diener, P.C. Dorrestein, G.M. Douglas, D.M. Durall, C. Duvallet, C.F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J.M. Gauglitz, S.M. Gibbons, D.L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G.A. Huttley, S. Janssen, A.K. Jarmusch, L. Jiang, B.D. Kaehler, K.B. Kang, C.R. Keefe, P. Keim, S.T. Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M.G.I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B.D. Martin, D. McDonald, L.J. McIver, A.V. Melnik, J.L. Metcalf, S.C. Morgan, J.T. Morton, A.T. Naimey, J.A. Navas-Molina, L.F. Nothias, S.B. Orchanian, T. Pearson, S.L. Peoples, D. Petras, M.L. Preuss, E. Pruesse, L.B. Rasmussen, A. Rivers, M.S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S.J. Song, J.R. Spear, A.D. Swafford, L.R. Thompson, P.J. Torres, P. Trinh, A. Tripathi, P.J. Turnbaugh, S. Ul-Hasan, J.J.J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K.C. Weber, C.H.D. Williamson, A.D. Willis, Z.Z. Xu, J.R. Zaneveld, Y. Zhang, Q. Zhu, R. Knight, J. G. Caporaso Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, Nat. Biotechnol. 37 (8) (2019) 852–857.
[13] J.A. Gilbert, J.K. Jansson, R. Knight, The Earth Microbiome project: successes and aspirations, BMC Biol. 12 (1) (2014) 69.
[14] J.T. Holah, 16 - microbiological environmental sampling, records and record interpretation, in: H.L.M. Lelieveld, J.T. Holah, D. Napper (Eds.), Hygiene in Food Processing, second ed., Woodhead Publishing, 2014, pp. 539–576.
[15] S. Wu, C. Sun, Y. Li, T. Wang, L. Jia, S. Lai, Y. Yang, P. Luo, D. Dai, Y.-Q. Yang, Q. Luo, N.L. Gao, K. Ning, L.-j. He, X.-M. Zhao, W.-H. Chen, GMrepo: a database of curated and consistently annotated human gut metagenomes, Nucleic Acids Res. 48 (D1) (2020) D545–D553.
[16] J. Lu, S.L. Salzberg, Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2, Microbiome 8 (1) (2020) 124.
[17] L.-I. McCall, C. Callewaert, Q. Zhu, S.J. Song, A. Bouslimani, J.J. Minich, M. Ernst, J.F. Ruiz-Calderon, H. Cavallin, H.S. Pereira, A. Novoselac, J. Hernandez, R. Rios, O.H. Branch, M.J. Blaser, L.C. Paulino, P.C. Dorrestein, R. Knight, M.G. Dominguez-Bello, Home chemical and microbial transitions across

urbanization, Nat. Microbiol. 5 (1) (2020) 108–115.
[18] J.F. Ruiz-Calderon, H. Cavallin, S.J. Song, A. Novoselac, L.R. Pericchi, J.N. Hernandez, R. Rios, O.H. Branch, H. Pereira, L.C. Paulino, M.J. Blaser, R. Knight, M.G. Dominguez-Bello, Walls talk: microbial biogeography of homes spanning urbanization, Sci. Adv. 2 (2) (2016) e1501061.
[19] D. Knights, J. Kuczynski, E.S. Charlson, J. Zaneveld, M.C. Mozer, R.G. Collman, F.D. Bushman, R. Knight, S.T. Kelley, Bayesian community-wide culture-independent microbial source tracking, Nat. Methods 8 (9) (2011) 761–763.
[20] MGnify. Anchialine metagenome raw sequence reads. https://doi.org/10.15468/4s5r6q.
[21] L.A. Levin, D.F. Boesch, A. Covich, C. Dahm, C. Erséus, K.C. Ewel, R.T. Kneib, A. Moldenke, M.A. Palmer, P. Snelgrove, D. Strayer, J.M. Weslawski, The function of marine critical transition zones and the importance of sediment biodiversity, Ecosystems 4 (5) (2001) 430–451.
[22] B.A. Menge, Top-down and bottom-up community regulation in marine rocky intertidal habitats, J. Exp. Mar. Biol. Ecol. 250 (1) (2000) 257–289.
[23] J.A. Fuhrman, J.A. Cram, D.M. Needham, Marine microbial community dynamics and their ecological interpretation, Nat. Rev. Microbiol. 13 (3) (2015) 133–146.