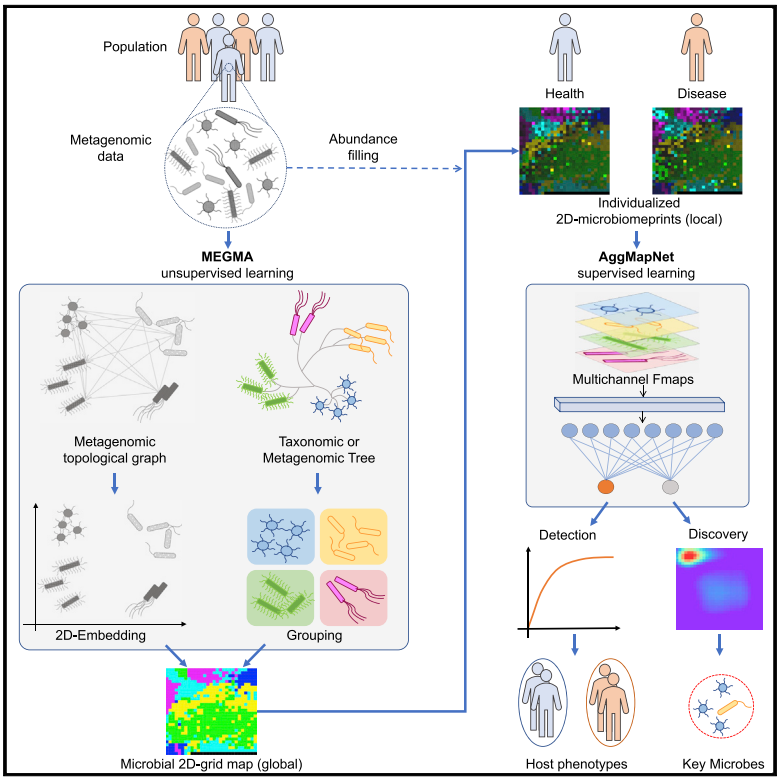
Article

Patterns

Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations

Graphical Abstract



Highlights

- Unsupervised MEGMA and supervised AggMapNet pipeline for host phenotype prediction
- Aggregating correlated microbes for signal amplification through manifold embedding
- Grouping-based multichannel operations boost the model performance significantly
- Identifying key microbial markers through global feature importance and saliency map

Authors

Wan Xiang Shen, Shu Ran Liang, Yu Yang Jiang, Yu Zong Chen

Correspondence

jiangyy@sz.tsinghua.edu.cn (Y.Y.J.), chenyuzong@sz.tsinghua.edu.cn (Y.Z.C.)

In brief

An interpretable DL-based metagenomic learning pipeline, MEGMA-AggMapNet-GFI, has been developed in this paper to facilitate host disease prediction and the discovery of consistent and replicable microbial biomarkers. In the pipeline, structured multichannel and signal-amplified metagenomic feature maps are constructed through unsupervised learning to enhance downstream supervised tasks of disease prediction and key biomarker recognition.

Shen et al., 2023, Patterns *4*, 100658
January 13, 2023
https://doi.org/10.1016/j.patter.2022.100658

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations

Wan Xiang Shen,[1,2] Shu Ran Liang,[1] Yu Yang Jiang,[1,*] and Yu Zong Chen[1,3,4,*]

[1]The State Key Laboratory of Chemical Oncogenomics, Key Laboratory of Chemical Biology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

[2]Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore, Singapore 117543, Singapore

[3]Shenzhen Bay Laboratory, Shenzhen 518000, China

[4]Lead contact

*Correspondence: jiangyy@sz.tsinghua.edu.cn (Y.Y.J.), chenyuzong@sz.tsinghua.edu.cn (Y.Z.C.)

https://doi.org/10.1016/j.patter.2022.100658

THE BIGGER PICTURE Machine learning methods have been practically employed for metagenomic phenotype prediction and biomarker discovery in clinical diagnostic investigations. Deep learning methods have also been explored as potential tools, but their practical applications are hindered by high dimensionality and low sample size. In this paper, we developed the unsupervised microbial embedding, grouping, and mapping algorithm (MEGMA) to enhance the downstream tasks of disease prediction and key biomarker recognition. Our study suggests that, through MEGMA unsupervised learning, structured multichannel and signal-amplified metagenomic feature maps can be constructed to enhance downstream supervised tasks of disease prediction and key biomarker recognition.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Metagenomic analysis has been explored for disease diagnosis and biomarker discovery. Low sample sizes, high dimensionality, and sparsity of metagenomic data challenge metagenomic investigations. Here, an unsupervised microbial embedding, grouping, and mapping algorithm (MEGMA) was developed to transform metagenomic data into individualized multichannel microbiome 2D representation by manifold learning and clustering of microbial profiles (e.g., composition, abundance, hierarchy, and taxonomy). These 2D representations enable enhanced disease prediction by established ConvNet-based AggMapNet models, outperforming the commonly used machine learning and deep learning models in metagenomic benchmark datasets. These 2D representations combined with AggMapNet explainable module robustly identified more reliable and replicable disease-prediction microbes (biomarkers). Employing the MEGMA-AggMapNet pipeline for biomarker identification from 5 disease datasets, 84% of the identified biomarkers have been described in over 74 distinct works as important for these diseases. Moreover, the method also discovered highly consistent sets of biomarkers in cross-cohort colorectal cancer (CRC) patients and microbial shifts in different CRC stages.

INTRODUCTION

Metagenomic analysis has been explored for non-invasive diagnosis and biomarker discovery.[1,2] Machine learning (ML) and deep learning (DL) methods facilitate metagenomics-based disease prediction and the discovery of consistent, replicable, and cross-cohort microbial biomarkers.[3–9] However, metagenomic data of individual clinical investigations are typical of low sample sizes (dozens-to-hundreds of samples),[3,4,10] high dimensionality (hundreds-to-thousands of microbes),[3,4,10] sparsity (sparsely distributed across taxonomic hierarchies), and high variations (biological and environmental).[11] These problems confound statistical inference and learning outcomes to random chances and false discoveries[12] and mask the identification of genuine

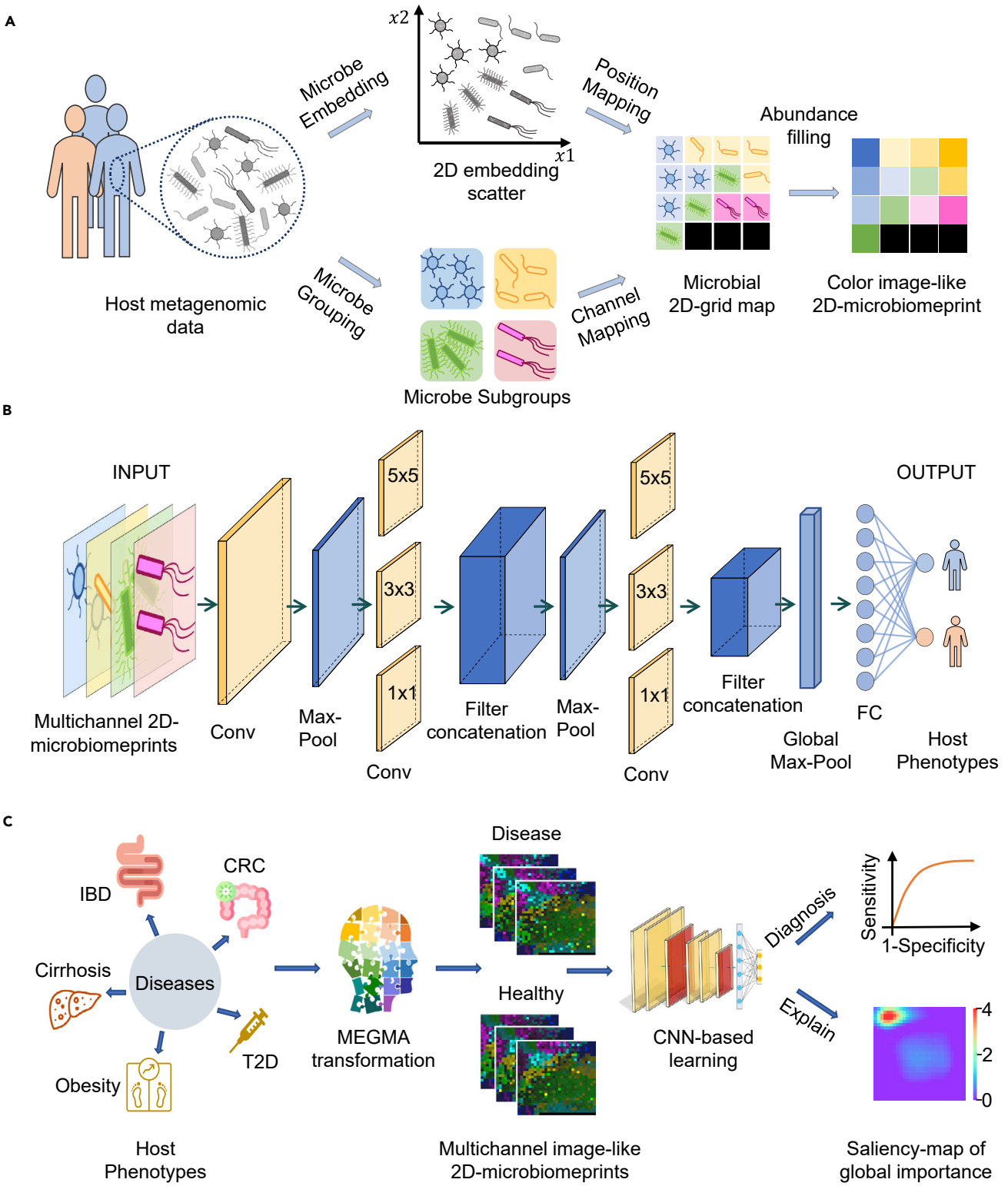


Figure 1. Metagenomic MEGMA pipeline and ConvNet-based AggMapNet for host disease prediction and replicable biomarker identification

(A) The unsupervised MEGMA to embed and group the host microbes, then map them on to a regular 2D grid map and finally transform the metagenomic data into the color image-like 2D microbiomeprint.

(B) The supervised ConvNet-based AggMapNet architecture for host phenotype prediction with the multichannel 2D microbiomeprints as inputs.

(legend continued on next page)

biomarkers.[12,13] DL outcomes are difficult to interpret, particularly in microbiome-wide association studies.[9,14] Instead of the end-to-end DL methods, ML methods with feature selection strategy have been practically used for metagenomic investigations of low sample sizes.[3,4,15] For example, the "Meta-Singer" is to rank the microbial features based on the aggregation of identified features from multiple ML models,[9] while the novel "predomics" tool employs the genetic algorithm to find the best number of features for simple condition models, leading to better accuracy and interpretability than the previous state-of-the-art (SOTA) ML models using fewer features.[15] However, microbiome data are complex, and ML methods with fewer selected features may be limited in the representation capability of the models and in learning complex patterns from the data.[14] New interpretable DL methods are needed for enhanced learning and interpretability of metagenomic data to complement existing ML and DL methods.

The widely used metagenomic ML methods include the least absolute shrinkage and selection operator (LASSO), ensemble tree-based random forest (RF), and support vector machines (SVMs) in combination with various feature selection techniques.[3–5] They learn unordered 1D vectors of taxa or microbe abundances. These tabular 1D vectors of high dimensionality and sparsity are not the most appropriate data form for efficient DL. The disease prediction capability of DL may be improved by converting metagenomic data into phylogenetically ordered representations based on taxa hierarchical trees.[6] Hence, appropriate metagenomic representation is important for enhanced learning. Two convolutional neural networks (ConvNets) Ph-CNN[7] and PopPhy-CNN[8,9] have been developed with the abundances of the taxonomically ordered microbes and 2D matrix of the embedded phylogenetic tree as input data, respectively.

Moreover, to alleviate the over-fitting issue when conducting disease predictions by DL, a promising algorithm, Met2Img,[16,17] has been introduced to exploit taxonomic (the so-called "Fill-up") and manifold embeddings (MEs), such as t-distributed stochastic neighbor embedding (t-SNE) to transform abundance data into "synthetic images," which enables the efficient exploration of ConvNets for disease classification based on the image created.[17] In the "synthetic images," the abundances are binned to generate the color space based on a given color map type. Testing results have indicated that the integration of phylogenetic information alongside abundance data improves classification performances.[16,17] While the Met2Img tool can generate color "synthetic images," the corresponding color space simply duplicates the abundance information, but it lacks the local-coherence characteristics on the transformed images and has overlap issues of the feature points (FPs) when using MEs, such as t-SNE.[16] Nevertheless, highly efficient ConvNet models rely on the recognition of 2D local-coherence and multichannel characters of natural color image data.[18] Natural images are highly structured and low-noise data, where character-distinguishing features of the data are concentrated in local regions of images. The restructuring of metagenomic data into spatially correlated 2D color image-like data is still needed for efficient DL with ConvNets.

In this work, we developed an unsupervised metagenomic microbial embedding, grouping, and mapping algorithm (MEGMA) to transform tabular metagenomic data into spatially correlated color image-like 2D representations named 2D microbiomeprints (3D tensor data in the form of width, height, and channel). Each channel contains a group of microbes, marked with a different color (Figure 1A). The MEs and position mappings were used to enhance the local connectivity and local coherence of image-like 2D representations, while the grouping operations were used to generate the multichannel (i.e., the number of the colors) characteristics of the 2D representations. Therefore, the final MEGMA 2D microbiomeprints are structured multichannel 3D feature maps (Fmaps) for enhanced performances in the subsequent learning tasks. For example, ConvNet-based AggMapNet (Figure 1B) DL models can be trained with MEGMA 2D microbiomeprints as inputs to learn the metagenomic data for disease prediction and biomarker discovery (Figure 1C).

Nineteen publicly available low sample size metagenomic datasets were used in this study, including a Disease-Set1[8,9] and a Disease-Set2[16,17] that are related to five diseases (cirrhosis, obesity, type 2 diabetes [T2D], inflammatory bowel disease [IBD], and colorectal cancer [CRC]), and two sets of recently published CRC gut metagenomic CRC-Nation[3] and CRC-Stage[4] datasets (Table 1). All these datasets used in our study were directly obtained from the processed data and results in previous studies. These datasets cover a different number of microbial species, and the metagenomic taxonomic profiles are generated by different pipelines, such as the MetaPhlAn2,[19] the mOTU2,[20] and the SILVA Living Tree Project (LTP).[21] We evaluated whether the ME methods are better than the random uniform embedding (RUE) method in generating 2D microbiomeprints. Moreover, the performances of ConvNet-based AggMapNet models trained on multichannel 2D microbiomeprints were compared with those trained on single-channel grayscale 2D microbiomeprints to determine if the former is more superior than the latter. These enable AggMapNet models to outperform the commonly used ML and DL models in the metagenomic benchmark datasets of Disease-Set1 (Table S1). We also compared MEGMA with the existing method Met2Img[16,17] to assess which image-like 2D representation generation algorithm performs better for disease prediction (Table S2).

We further show that the AggMapNet explainable module in the analysis of the 2D microbiomeprints led to the identification of the important microbes (IMs), consistent with literature-reported biomarkers and biological mechanisms. A saliency map[22] is used to reflect the degree of importance of a feature (i.e., a microbe) in the input 2D microbiomeprints. On the identification of the biomarkers for the Disease-Set1, 84 of the 100 identified IMs, which include the top 20 species for each disease of cirrhosis, IBD,

(C) The host phenotype prediction scenarios from metagenomic data and biomarker discovery based on the MEGMA personal multichannel 2D microbiomeprints using ConvNet-based AggMapNet. These were tested for host phenotype predictions and important disease-specific microbial biomarkers identification (explanation saliency map of important features) on five disease datasets cirrhosis, obesity, type 2 diabetes (T2D), inflammatory bowel disease (IBD), and colorectal cancer (CRC).

Table 1. Summary of the human gut metagenomic datasets in this study

Data Group	Reference	Dataset	No. of cases	No. of controls	No. of species
Disease-Set1	PopPhy-CNN and Meta-Singer, Reiman et al.[8,9]	cirrhosis	114	118	542
		IBD	25	85	443
		obesity	164	89	465
		T2D	223	217	606
		CRC	48, 39	47	507
Disease-Set2	Met2Img, Nguyen et al.[16,17]	cirrhosis	118	114	542
		IBD	25	85	443
		obesity	164	89	465
		T2D	170	174	572
		CRC	48	73	503
Cross-nation sets of CRC-Nation	Wirbel et al.[3]	AUS	46	63	849
		CHN	74	54	849
		DEU	60	60	849
		FRA	53	61	849
		USA	52	52	849
Disease-stage sets of CRC-Stage	Yachida et al.[4]	MP	40	127	7,278
		S0	27	127	7,278
		SI/II	69	127	7,278
		SIII/IV	54	127	7,278

T2D, obesity, and CRC, are consistent with the disease-relevance reports in the 74 distinct literature reports. On identifying of the consistent and replicable microbial signatures across cohorts of five nations in the early detection of CRC,[3] the global feature importance (GFI) of AggMapNet is better than the commonly used marker-identifying methods, such as generalized fold change (FC),[3] two-sided Wilcoxon rank-sum test (WRST) p value with Benjamini-Hochberg false-discovery rate (FDR) correction (q value),[3] LASSO model coefficient, and RF feature importance (FI). The AggMapNet GFI-based saliency map can also detect the microbial shifts in different stages of CRC. In conclusion, we show that the use of 2D microbiomeprints as metagenomic representations can significantly enhance downstream tasks of DL-based disease prediction and discovery of key signatures, and an interpretable DL-based metagenomic learning MEGMA-AggMapNet-GFI pipeline (released in aggmap: https://pypi.org/project/aggmap/1.1.7/) with good performance has been developed for disease prediction and biomarker discovery.

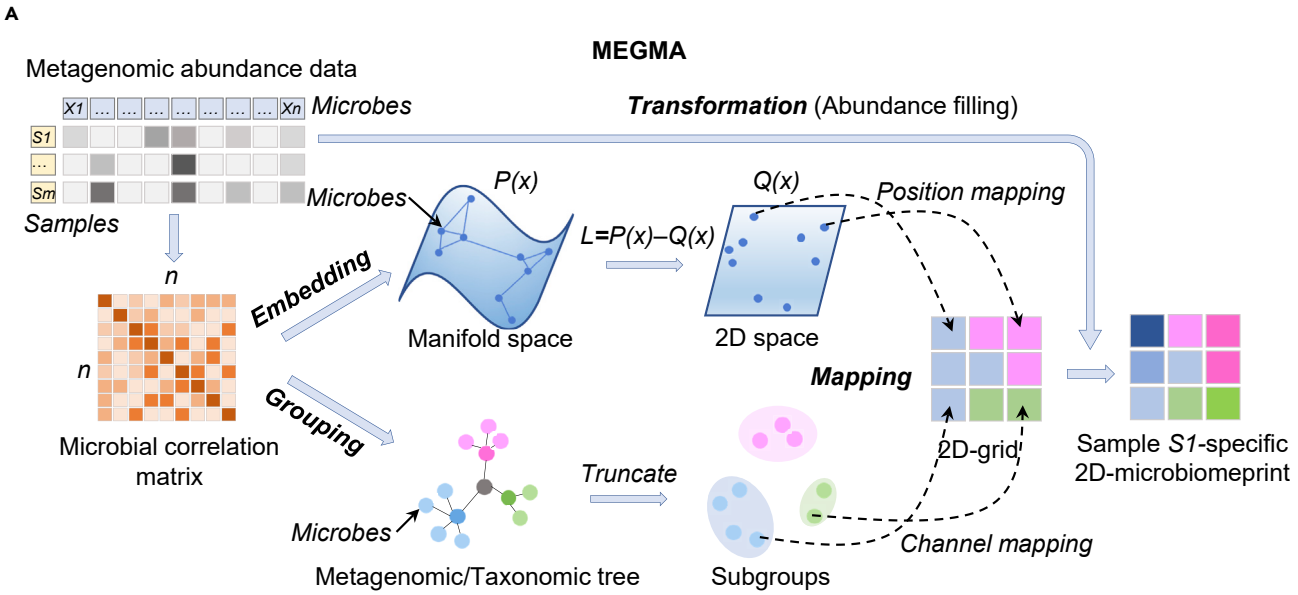
RESULTS

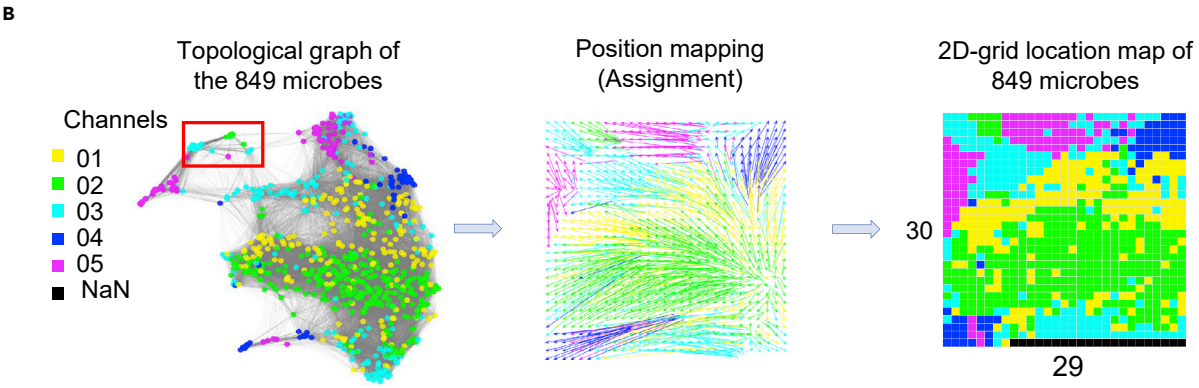
MEGMA approach for restructuring metagenomic data into multichannel microbiome 2D representations

MEGMA was developed to transform high-dimensional and sparse metagenomic data from the tabulated 1D vector forms into color image-like multichannel 2D microbiomeprints. Each 2D microbiomeprint represents a microbial abundance 2D imprint of individual samples. Natural color images are highly structured and low-noise data with two important characteristics, namely local coherence and multiple channels (e.g., RGB channels). The ME and metagenomic/taxonomic grouping (MG/TG) of microbes were particularly designed to construct the local coherence and multichannel (i.e., the number of the colors or groups) characters of the 2D microbiomeprints, respectively.

As shown in Figure 2A, the input is the sparse, high-dimensional metagenomic tabular data with logarithm transformation, then the pairwise correlation matrix of the microbes is calculated. The ME and MG were based on the correlation matrix (except for the TG). In the grouping stage, the microbes were grouped into several subgroups by truncating the metagenomic hierarchical clustering tree (Figure S1A) or taxonomic phylogenetic tree. In the embedding stage, each microbe was assigned to its x and y coordinates in the 2D space by various manifold learning algorithms (Figure S1B). In the mapping stage, position mapping is aimed at the assignment of each microbe to one optimized position in the 2D grid map, while channel mapping is to split the microbes into different channels based on the subgroups (Figure S1C). Finally, the microbial abundance vector of one sample was then transformed into a color image-like multichannel 2D microbiomeprint based on the microbial locations in the regular 2D grid (Figure S1D).

Figure 2B presents the topological exploration of the 2D microbiomeprint generation. Firstly, the ME algorithm generates the topological graph based on the abundance correlation distance of the 849 microbes in the CRC-Nation set.[3] In the graph, each node is one microbial species, the correlation distance is represented by the weights of the edges. The nodes can be divided into several groups to generate the multichannel character by a metagenomic hierarchical clustering tree or a taxonomic tree. The optimized topological graph ensures that the correlated microbes can be aggregated as neighbors in the 2D space, such as the subgraph in the red box of Figure 2B (Figures S2 and S3). Secondly, the position mapping ensures that the embedded microbes in the optimized graph can be assigned to the regular 2D grid while preserving the adjacent relationships as much as possible. Finally, the 2D grid location map





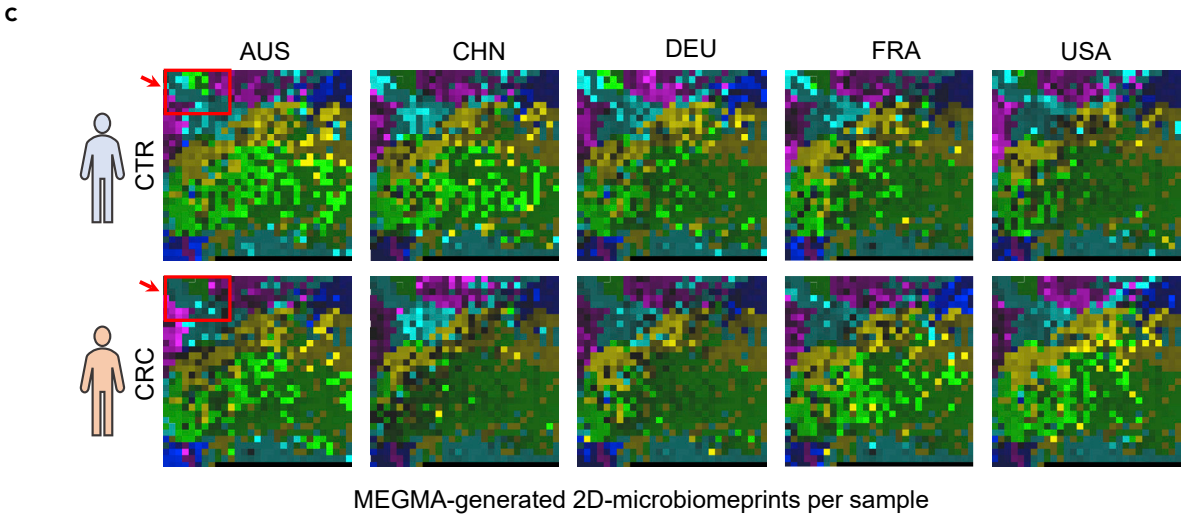


Figure 2. Generation of host-specific image-like multichannel 2D microbiomeprints by MEGMA

(A) The flowchart of MEGMA. The input is the metagenomic tabular data with logarithm transformed, then the pairwise correlation matrix of the microbes is calculated, followed by the following procedures: in the embedding stage, each microbe will get the x and y coordinates in the 2D space by manifold learning

(legend continued on next page)

of the 849 microbes is generated, each pixel in the 2D grid map represents one species of the microbe and each color is one channel (subgroup).

The image-like 2D microbiomeprints were further generated by the abundance filling of the 2D grid map (i.e., the transformation stage). Since each sample (a patient or a healthy subject) has an abundance vector, MEGMA can transform individual-specific 2D microbiomeprint. The example 2D microbiomeprints for each class (CRC or healthy control, CTR) in each of the five nations are shown in Figure 2C. In the CRC group, the upper left corner of the 2D microbiomeprints is quite distinct from the same area in the CTR group, the microbes in this area are mostly of higher abundance in CRC patients than in healthy controls. The microbes in this area are in the subgraph of Figure S2, most of which are the known microbes associated with CRC. In conclusion, through the transformation of the MEGMA, high-dimensional, noisy, and sparse metagenomic data can be transformed into highly structured data of color 2D microbiomeprints, where the key microbes will be gathered as a hot zone to highlight the abundance signal. In addition, unrelated microbes are distributed in other areas, helping to reduce noise. Therefore, this transformation can potentially enhance the performance of downstream tasks, such as disease prediction and key signature identification.

Manifold-guided metagenomic embedding substantially improves host disease-prediction performance over random embedding

Five ME algorithms (multi-dimensional scaling [MDS], isometric mapping [ISOMAP], locally linear embedding [LLE], t-SNE, and Uniform Manifold Approximation and Projection [UMAP]) and one RUE method are compared in the 2D microbiomeprint generation to verify whether ME is advantageous. Four binary tasks disease datasets of Disease-Set1 (cirrhosis, IBD, T2D, obesity) were used to evaluate the performance of disease prediction based on the 2D microbiomeprints generated by different embedding methods. All ME-derived 2D microbiomeprints are more structured and spatially correlated than RUE-derived 2D microbiomeprints (Figure 3A). ConvNet-based AggMapNet models with the same hyperparameters trained on these 2D microbiomeprints show varying performance. Based on 10-fold cross-validation and the lower kernel size of the first layer in AggMapNet, the disease-prediction performance The AUC of all ME-based models consistently outperformed those of the RUE-based models by substantial margins (Figure 3B). Paired t tests were used for evaluating the significance of differences based on 100 paired metrics from RUE and each ME method. All ME methods, except for MDS or UMAP, are significantly better than random method RUE in the AUC metric for the two relatively balanced cirrhosis and T2D datasets, respectively. For the unbalanced IBD and obesity datasets, UMAP ($p < 0.001$) and t-SNE ($p < 0.05$) are significantly better than the random method RUE in F1 score (Figure S4). These results strongly suggest that manifold-guided feature embedding substantially enhances ConvNet-based metagenomic disease-prediction performance. This arises from the superior ability of CNN in learning the local patterns (local connectivity) of images,[23] the convolutional layers take advantage of the local spatial coherence of the 2D microbiomeprint.

It is noted that, although DL models of different ME algorithms performed differently on different datasets, the manifold-based ME algorithms (e.g., UMAP) exhibited distinguished ability. For example, on the four datasets with lower feature numbers, DL models of ISOMAP and LLE algorithms performed slightly better than those of UMAP and t-SNE algorithms (Figure 3B). But on the CRC-Nation[3] datasets (Table 1) with a higher number of features (Figure S5), the models of UMAP and t-SNE algorithms outperformed those of ISOMAP and LLE algorithms. Under the study-to-study transfer (STST) (model trained on the data of one nation and tested on the data of rest nations) tests, the performances of the DL models of the UMAP algorithm significantly outperformed those of the RUE algorithm on all nations, and they significantly outperformed those of the ISOMAP and LLE algorithms on some of the nations, such as Germany, France, Australia, or the US (Figure S5). The levels of these performance variations may be partly influenced by the hyperparameters of these ME algorithms. Nonetheless, UMAP, as a SOTA dimensionality reduction algorithm, is able to preserve both local and global data structures during embedding by a trade-off hyperparameter.[24,25] Preserving the local or global structures of the data may critically affect DL performances in the 2D representations[18,26] because the ConvNet is shown to strongly rely on local texture cues.[27] Moreover, UMAP is much faster in embedding large numbers of microbes and has been shown to be able to restore the original images from MINST data with randomly arranged pixels in our recent study.[18] Therefore, UMAP was chosen as the default embedding algorithm in MEGMA.

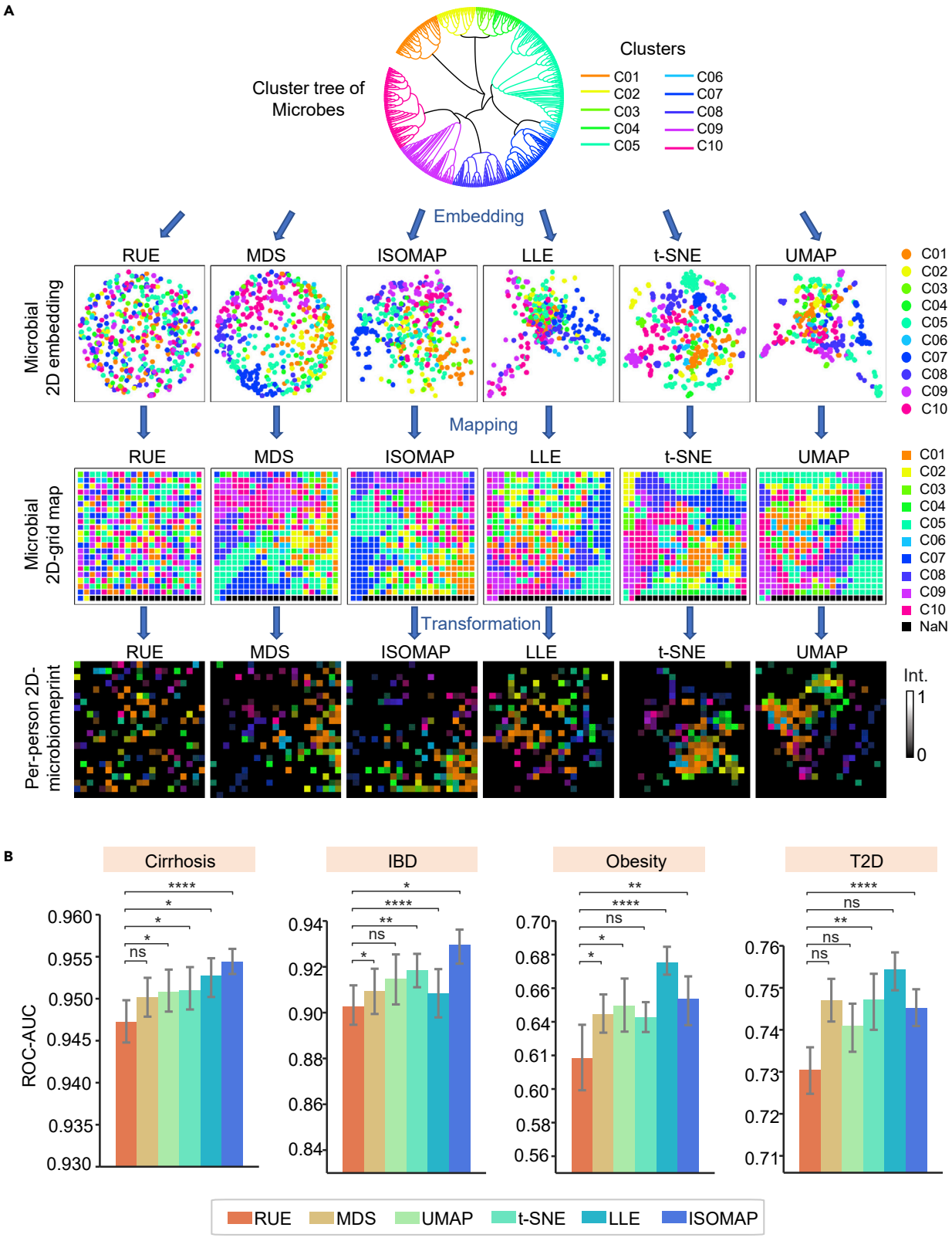
The grouping-based multichannel 2D microbiomeprints further enhance the host disease-prediction performance remarkably

To explore whether grouping-based multichannel 2D microbiomeprints can improve the disease-prediction performance

algorithms. In the grouping stage, the host microbes are grouped into several subgroups by truncating the metagenomic hierarchical clustering tree or taxonomic phylogenetic tree. In the mapping stage, the position mapping is response for the assignment of each microbe to one optimized position in the 2D grid map, while the channel mapping is to split the microbes into different channels based on the subgroups. Finally, in the MEGMA transformation, the species abundance vector of one sample is then transformed into the host-specific multichannel image-like 2D microbiomeprint based on the species locations in the regular 2D grid. Each pixel in the 2D microbiomeprint represents one species (or operational taxonomic units) of microbe, and each color (group) is one channel, the brightness of the color represents the magnitude of the abundance value.

(B) The 2D embedded topological graph by the UMAP-mediated MEGMA, the microbial position mapping process, and the final 2D grid map of the 849 microbial species in CRC-Nation data (i.e., each pixel is one microbial species). The number of the cluster channel is 5, the size of the regular 2D grid map is (30, 29), and the black pixels are used to amend the grid. For the topological graph, only edges with weights greater than 0.2 are shown to simplify the complete graph, the zoomed subgraph (red box) is in Figure S2.

(C) The MEGMA-generated individualized multichannel 2D microbiomeprint (transformed from the 2D grid map) in (B) for a CRC patient and a CTR individual in each of the five nations of AUS, Australia; CHN, China: DEU, Germany: FRA, France; USA, United States of America. The red box indicates the important area identified by the naked eye to distinguish the CRC and CTR samples.



(legend on next page)

more than grayscale (single-channel) 2D microbiomeprint, we chose UMAP as the embedding method and investigated the two grouping algorithms (MG and TG). The MG algorithm clusters microbes by phenotype distances (i.e., similarity in microbial abundance patterns), while the TG algorithm clusters microbes by genotype distances (i.e., distances on a phylogenetic tree). These two algorithms were used to group microbes on different dimensions and scales. With TG on the IBD dataset, truncating the phylogenetic tree at levels of control, kingdom, phylum, class, order, family, and genus led to 1, 3, 10, 18, 23, 49, and 68 channels of 2D microbiomeprints, respectively (Figure 4A). For comparison of the two grouping algorithms, the number of clusters in MG was also set at 1, 3, 10, 18, 23, 49, and 68, respectively (Figure 4B). The control (the channel parameter c = 1) corresponds to cases of no groupings and all microbes were embedded into one channel of 2D microbiomeprints, leading to the grayscale image-like Fmap.

We found that AggMapNet models trained on both MG-based and TG-based multichannel 2D microbiomeprints showed remarkable AUC performances than those trained on single-channel 2D microbiomeprints (i.e., the controls) (Figures 4C, 4D, and S6), suggesting that these groupings are highly effective for enhanced metagenomic data learning. To a certain extent, model performances were enhanced by increasing the number of grouping clusters, because of more fine-grained data separations, and more specific feature learning. The TG boosted the model performance more hierarchically with the number of channels than MG (Figures 4C and S7). However, the metagenomic clusters are more balanced in size than the taxonomic clusters (Figures 4A and 4B). At a lower number of clusters (e.g., c = 3, 10, 18, …), MG can boost model performance more than TG (Figures 4C and 4D). These results showed that, irrespective of which of the two grouping algorithms was used, the finer the grouping scale, the better the model performance can be. MG seems to be more robust than (and thus can practically replace) TG for enhanced model performances.

AggMapNet with 2D microbiomeprints as input significantly outperforms commonly used machine learning and deep learning models in disease prediction

AggMapNet with 2D microbiomeprints as input (under UMAP embedding and MG) was evaluated on all five datasets in Disease-Set1. Firstly, AggMapNet models were constructed on the five datasets (cirrhosis, T2D, obesity, IBD, and CRC), and compared with four existing ML (RF, SVM, LASSO, MLPNN)[9] and one existing DL (PopPhy-CNN)[8] model of each dataset. Under the same performance metrics (mean AUC, MCC, Precision, Recall, and F1 score) and the same stratified 10-fold cross-validation (repeated 10 times using different random seeds, with 100 data points for each dataset) as in the reported studies,[9] AggMapNet outperformed all ML and DL models on all five datasets by larger margins, with an exception of only one metric in one dataset. The single exception is the AUC metric on the unbalanced obesity dataset, in which case AggMapNet scored slightly lower AUC than that of RF (0.642 versus 0.648) but the F1 score of AggMapNet is significantly better than that of RF (0.636 versus 0.556, $p \leq 0.0001$) (Table S1; Figure S9). Since the performance metrics are roughly normally distributed (Figure S8), a paired t test was used to evaluate whether the performance is significantly different between AggMapNet and RF (Figure S9). The results showed that AggMapNet outperformed RF significantly on IBD ($p \leq 0.001$), T2D ($p \leq 0.01$), and obesity ($p \leq 0.001$) datasets by F1-score and the MCC metric. Under the AUC metric, AggMapNet outperformed RF significantly ($p \leq 0.001$) on the datasets of cirrhosis and IBD. Overall, except for the multi-class dataset of CRC, AggMapNet models significantly outperformed the RF models on the remaining four datasets.

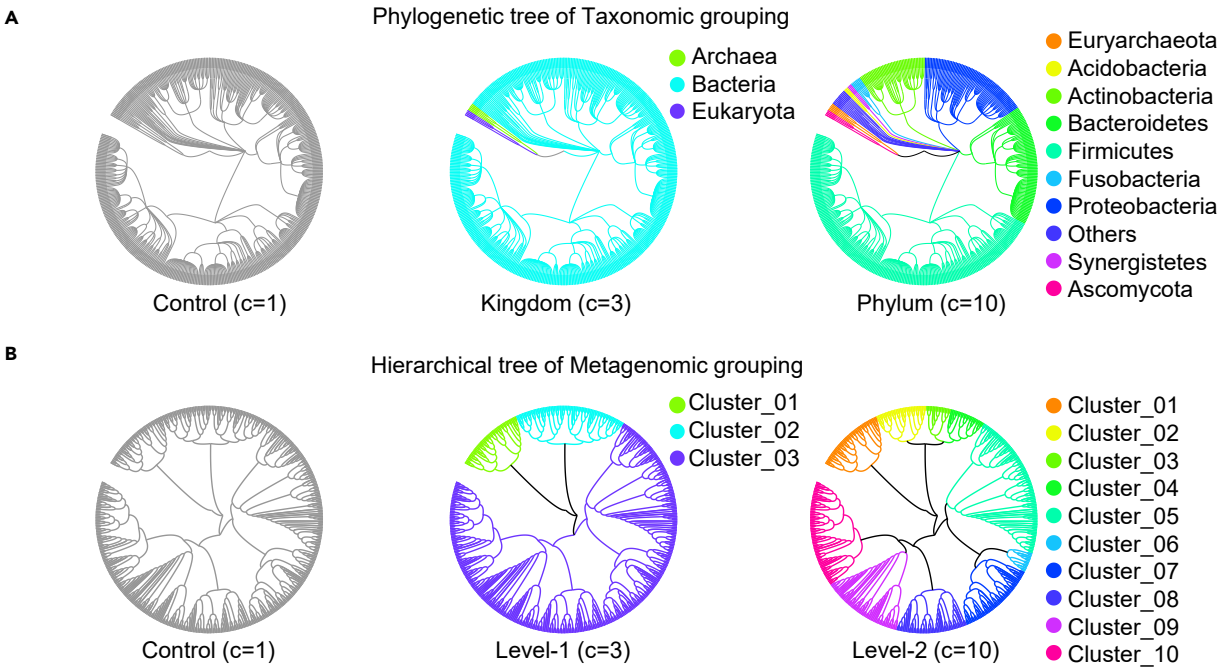
Secondly, AggMapNet models were developed on the cross-nation sets of CRC-Nation in comparison with the commonly used ML models by the STST validations as in the reported studies,[3] i.e., a model trained on the data of each of the five nations (France, Australia, Germany, China, and the US), and tested on the data of the rest of the nations, a total of 20 tests. LASSO was used because it achieved SOTA performances of the STST test in the previous study.[3] MEGMA-generated 2D microbiomeprints of individual classes (CRC and healthy control, CTR) across the nations are presented in Figure 2C. AggMapNet outperformed LASSO and RF in 15 and 14 of 20 STST tests of CRC-Nation by substantial margins, respectively (Figure 5C). In the 6 tests of lower LASSO performance (AUC < 0.7), AggMapNet markedly improved the performance over LASSO in all 6 tests (Australia-to-US 0.59 to 0.78, France-to-US 0.64 to 0.77, Australia-to-France 0.62 to 0.75, Australia-to-US 0.59 to 0.78, Australia-to-Germany 0.65 to 0.89, China-to-US 0.67 to 0.76). AggMapNet also markedly improved 3 tests of lower RF performance (AUC < 0.7) (Australia-to-US 0.57 to 0.78, Australia-to-China 0.65 to 0.75, Germany-to-Australia 0.59 to 0.86).

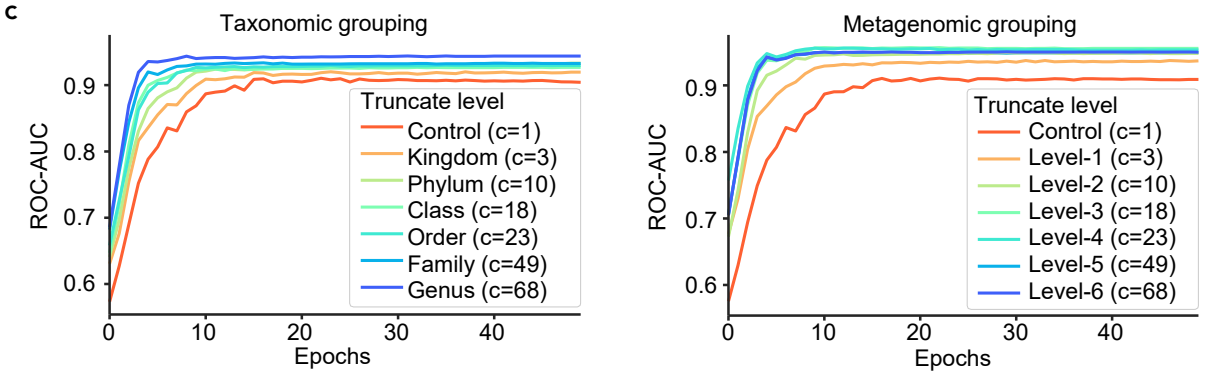
Thirdly, AggMapNet models were generated for the disease-stage sets of CRC stage in comparison with the SOTA ML models. Yachida et al.[4] have constructed RF and LASSO classifiers to identify diagnostic markers to distinguish samples from patients with S0 (n = 27) and SIII/IV (n = 54) CRCs from samples of healthy controls (n = 127). RF scored SOTA performance in classifying CRC stages S0 and SIII/IV at 0.73 and 0.83 AUC.[4]

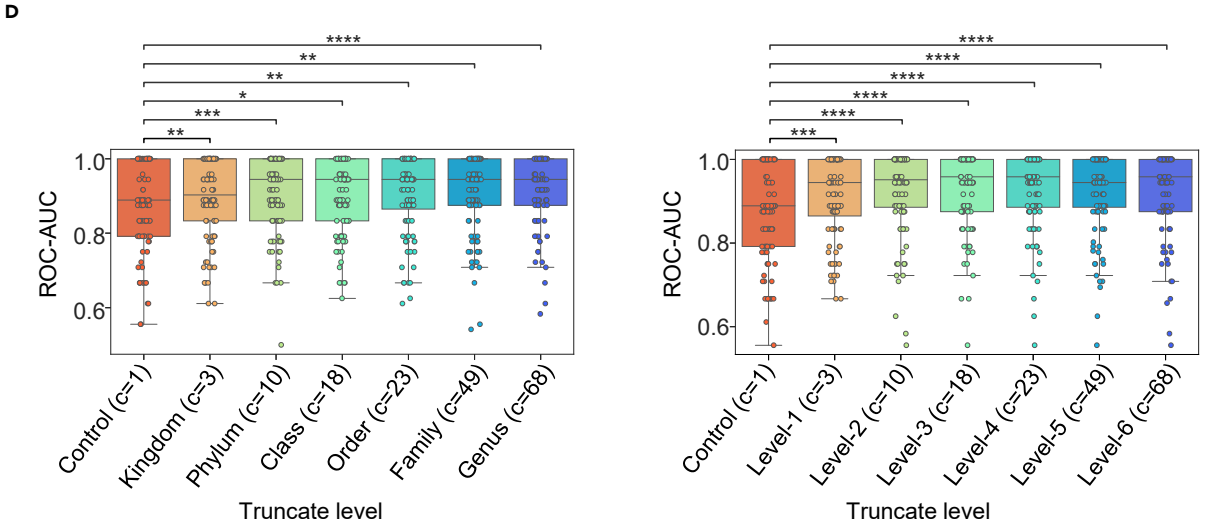
Figure 3. 2D microbiomeprints of Disease-Set1 generated by random and manifold embedding algorithms and disease-prediction performances of the corresponding AggMapNet models

(A) The example of the different embedding algorithms for generating the 2D signatures in the IBD dataset. The random uniform embedding (RUE) and five manifold embedding approaches, including multi-dimensional scaling (MDS), isometric mapping (ISOMAP), locally linear embedding (LLE), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), are used for the embedding of the microbes to generate the microbial grid map and 2D microbiomeprint Fmaps. The metagenomic-based hierarchical clustering is used to group the microbes into ten subgroups. Each color of microbes is one cluster from the clustering of the microbial species.

(B) The average ROC AUC performance of the AggMapNet models that are trained on the four metagenomic Disease-Set1 (cirrhosis, IBD, obesity, and T2D binary task) with the RUE, MDS, ISOMAP, LLE, t-SNE, and UMAP 2D microbiomeprint Fmaps as inputs. The model performances were evaluated by the average AUC of 10-fold cross-valuation repeated 10 times (total 100 data points), and the SD error bars of 10 repeats are shown. The paired t test (100 pairs of RUE versus each ME method) was used to test the significance of the difference between the random embedding RUE and ME performance. p values for the significant levels: ****$p \leq 0.0001$, ***$0.0001 < p \leq 0.001$, **$0.001 < p \leq 0.01$, *$0.01 < p \leq 0.05$; not significant (n.s.), $0.05 < p \leq 1$.







(legend on next page)

AggMapNet outperformed RF with selected features in classifying stages S0 and SIII/IV of CRC-Stage with AUC improved from 0.73 to 0.83 to 0.84 and 0.89. It additionally classified the CRC stages MP and SI/II at 0.83 and 0.86 AUC (Figure S12D). With a lower bacterial species abundance cut-off 1e-8, AggMapNet was trained based on 7,278 species (Figure S12A) compared with dozens of selected species in training RF and LASSO.[4] The improved performance of AggMapNet may be partly attributable to the efficient learning of more diverse species by MEGMA feature restructuring.

It should be noted that the 2D microbiomeprints of different datasets can be in different sizes due to the differences in the number of metagenomic features. The number of features determines the width and height of the 2D microbiomeprints. For example, the size of MEGMA-generated 2D microbiomeprints for Disease-Set1 is smaller than (24, 24) in width and height, because the numbers of features are from 443 to 606. For comparison, the size of 2D microbiomeprints for CRC-Nation and CRC-Stage are (30, 29) and (86, 85), respectively. Our results showed that the MEGMA-AggMapNet pipeline outperformed conventional ML methods on 2D microbiomeprints of different size scales, demonstrating the robustness of the proposed approach.

MEGMA outperformed Met2Img in disease prediction based on the generation and DL of image-like Fmaps

We further compared our MEGMA method with the existing Met2Img method on disease-prediction performances of DL models trained with image-like Fmaps generated by each method. In the study with Met2Img,[16,17] the proposed Fill-up approach has produced better performance on the Disease-Set2 datasets than that of the ME algorithm in generating the "synthetic images" as the input for the DL models. In the Fill-up approach, features (i.e., the microbial species) are arranged into a square matrix of a given size by a specific order of phylogenetic classification for the known species, while the ME algorithms in Met2Img[16,17] have overlap issues because the features are not properly arranged on the 2D grid space.

From the methodology perspective, our MEGMA method combines the advantages of the Fill-up approach and ME algorithms in the Met2Img[16,17] and overcomes the overlap issues of the ME algorithms by the "position mapping" operation (Figure 2B). More importantly, MEGMA generates multichannel information by grouping microbes at different dimensions and scales, which is in contrast to Met2Img algorithms whereby microbial abundance values are repeatedly projected into color maps. Consequently, Met2Img[16,17] can only generate Fmaps of up to 3 channels, while MEGMA can generate any integer number of channels (e.g., 60 channels) (Table S2). We further compared the performances of AggMapNet models on all the datasets of Disease-Set2 trained with Met2Img Fmaps and MEGMA Fmaps, respectively (Table S2). The models with MEGMA Fmaps (default UMAP embedding and MG channel c = 5) produced significantly better performances ($p < 0.0001$ by paired t test) than those with Met2Img[16,17] Fmaps, under almost all metrics on all datasets of Disease-Set2. Increasing the channel number of MEGMA (c = 60) can further improve AggMapNet performances in some of the datasets, such as cirrhosis, obesity, and CRC of Disease-Set2.

MEGMA-AggMapNet explainable module exposes disease-associated microbes as potential biomarkers

To explain MEGMA-AggMapNet models, the GFI score *S* was calculated for Disease-Set1 (cirrhosis, IBD, T2D, obesity, and CRC) from a 10-fold cross-validation. The explanation microbial saliency map and the top 20 IMs for each dataset are provided in Figure S10 and Tables S3–S7, respectively. The GFI scores for all features are provided in Data S1. The IMs of every dataset are concentrated in a certain hot zone of the explanation saliency map (Figure S10). Hence, AggMapNet GFI selected IMs are highly correlated with each other in terms of microbial abundance distribution across the subjects (organized by MEGMA).

More importantly, by comparing the identified 100 IMs (the top 20 microbes for each disease by AggMapNet-GFI) with the literature one by one, 84 of the 100 IMs are highly consistent with the disease-relevance reports of 74 distinct literature (of the top 20 microbes for the 5 diseases of cirrhosis, IBD, T2D, obesity, and CRC, there are 16, 16, 16, 17, and 19 microbes that have been mentioned in the literature as key markers or IMs to the specific disease, respectively) (Tables S3–S7). These results further indicate that AggMapNet-identified IMs are potential biomarkers in high consistency with literature reports. The GFI-based saliency map signatures are useful for identifying and displaying these key biomarkers for specific diseases.

The top 20 IMs (Avg. GFI > 2.2) of cirrhosis are mostly up-signatures (Table S3). The top-ranked *Clostridium symbiosum* and *Haemophilus parainfluenzae* are reportedly increased in cirrhosis patients[28–30] and universal up-signatures of cirrhosis.[30] The IMs also include four *Clostridium* species (*C. citroniae*, *C. hathewayi*, *C. asparagiforme*, and *C. nexile*), *H. parainfluenzae*, and three *Veillonella* species (*V. dispar*, *V. parvula*, and an unclassified species). *Clostridium* species are enriched in liver cirrhosis[31] and

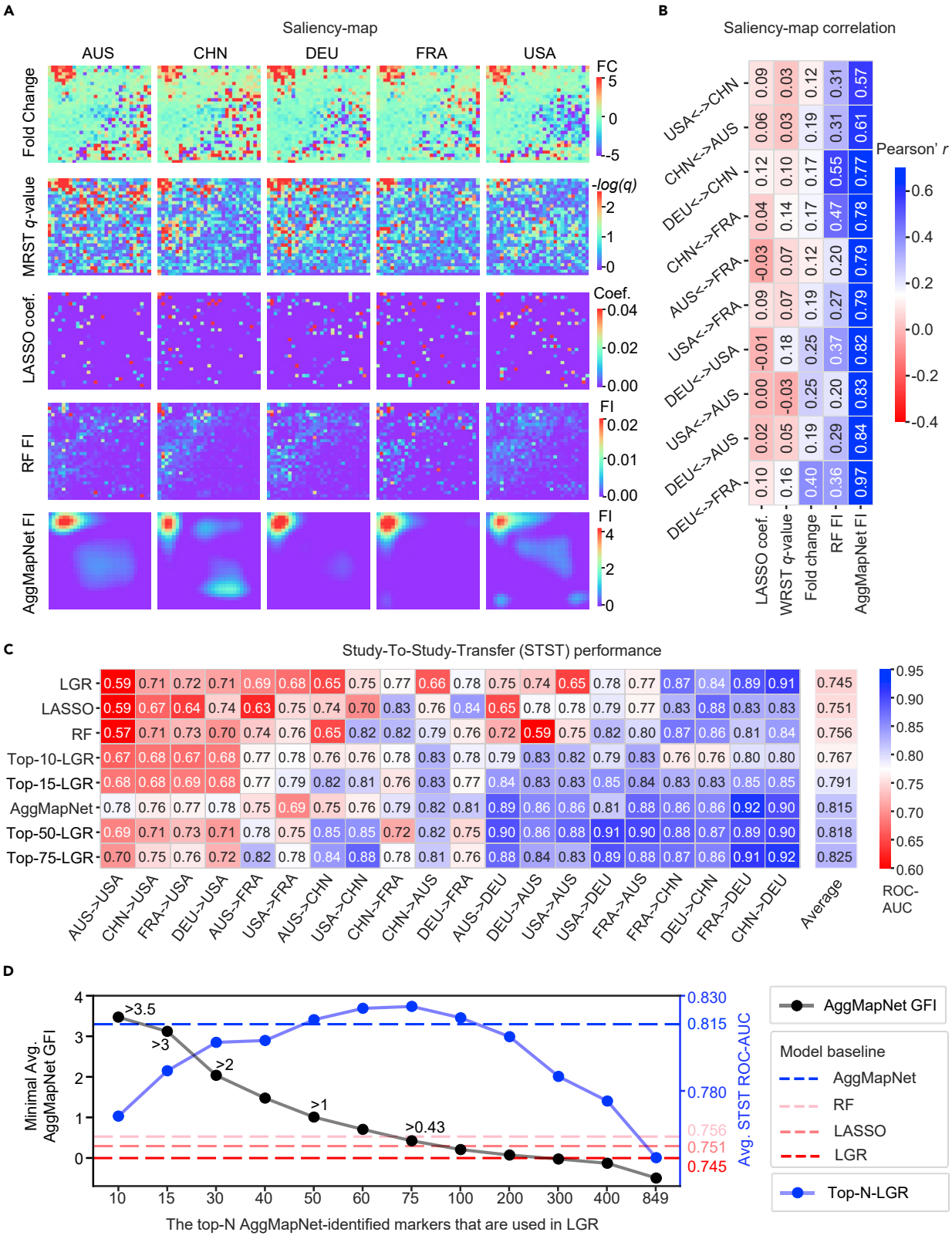
Figure 4. The microbial taxonomic grouping and metagenomic grouping for generating the multichannel characteristics of 2D microbiomeprints and their disease-predictive performances as inputs of AggMapNet on the IBD dataset of Disease-Set1

(A) The taxonomic grouping of microbes by truncating taxonomic levels in the phylogenetic tree (the 1, 3, and 10 clusters are generated by truncating the Control, Kingdom, and Phylum levels, respectively. c is the number of the channels (e.g., c = 10 means that the number of channels of the 2D microbiomeprint is 10).

(B) The metagenomic grouping by microbes specifying the number of clusters in the hierarchical clustering tree. To make a fair comparison with the taxonomic-based grouping, the same number of clusters are specified in the metagenomic-based grouping (i.e., c = 1, 3, and 10, respectively).

(C) The average performance (10-fold cross-validation and repeat 10 times by different random seeds) of the AggMapNet models that are trained on the taxonomic-based and metagenomic-based grouping multichannel 2D microbiomeprint, respectively. The taxonomic levels of Control, Kingdom, Phylum, Class, Order, Family, and Genus are used to truncate the phylogenetic tree of the IBD dataset, and the 1, 3, 10, 18, 23, 49, and 68 clusters are generated accordingly. The metagenomic-based grouping specifies the same number of clusters (clades) as the taxonomic-based grouping.

(D) Pairwise comparison of ROC AUC performance for taxonomic-based and metagenomic-based groupings. The significant level of paired t test with Bonferroni correction: ****$p \leq 0.0001$, ***$0.0001 < p \leq 0.001$, **$0.001 < p \leq 0.01$, *$0.01 < p \leq 0.05$; not significant (ns), $0.05 < p \leq 1$.



(legend on next page)

altered in the progression of cirrhosis to hepatocellular carcinoma.[32] Commensal *Veillonella* microbes are correlated with liver diseases, such as cirrhosis and autoimmune hepatitis.[31] *V. parvula* is a biomarker for clinical outcomes of cirrhosis patients.[33] The top 20 IMs of IBD (GFI > 1.6) are mostly down-signatures (Table S4). The top-ranked IM is *Alistipes finegoldii*. Another IM is *A. putredini*. The reduced abundance of these species is characteristic of IBD.[34] In addition, we also identified the correlated species, such as *Coprococcus* sp. ART55/1 as an important species that is markedly reduced in IBD. The top 20 IMs of T2D (GFI > 2.0) are primarily down-signatures (Table S5), the top-ranked IM is an unclassified *Butyrivibrio* species, and another IM is *Faecalibacterium prausnitzii*. The reduction of these anti-inflammatory bacteria likely contributes to the development of T2D and are potential targets for T2D therapeutics.[35] Two other IMs Lachnospiraceae species (Lachnospiraceae *bacterium* 1_1_57FAA and 3_1_46FAA) are of interest. Earlier human and mouse metagenomic studies have suggested that Lachnospiraceae species are associated with T2D.[36] Some of the top 20 IMs (GFI > 0.7) for obesity (Table S6) are also experimentally implicated. The top-ranked IM *Ruminococcus bromii* is implicated by a study of the Japanese obese and non-obese population.[37] Two IMs *Oxalobacter formigenes*[38] (5, 15) and *Butyrivibrio crossotus*[39] (7, 14) are also known obesity-associated gut microbes that are top-ranked microbes. The top 20 IMs (GFI > 2.4) of CRC are mostly up-signatures (Table S7). Five IMs *Peptostreptococcus stomatis*, *Parvimonas micra*, *Fusobacterium nucleatum* unclassified *Parvimonas*, and *Gemella morbillorum*, are known CRC-related gut microbes and the first three microbes are part of a biomarker panel of CRC.[40]

The identified biomarkers are more reproducible than those of the established methods in cross-nation validations

MEGMA-AggMapNet model explanation was also used for identifying the CRC biomarkers of each nation of the cross-nation CRC-Nation[3] data. The reproducibility of these biomarkers was evaluated by the study(nation)-to-study transfer validations across the five nations, in comparison with those identified by three established methods. These include two statistical methods, the generalized FC[3] and the two-sided WRST with FDR-adjusted p value (WRST q value),[3] and two ML methods, the coefficient of the LASSO models (LASSO coef.) and the FI score of the RF models (RF FI). Figure 5A presents the microbial 2D saliency maps of each nation identified by every method, which shows that the IM hot zone derived by AggMapNet GFI is highly concentrated in the upper left corner of the saliency maps while those derived by the three established methods are more scattered across multiple parts of the saliency maps. Importantly, the FC saliency maps reveal that the microbes on the upper left corner are mostly of higher abundance in CRC than in controls. In addition, there is also a distinct difference in the upper left corner of the MEGMA-generated color 2D microbiomeprints of CTR and CRC groups (Figures 2B and 2C), suggesting that AggMapNet GFI has identified the true associations between the microbes and CRC.

Compared with the biomarkers identified by the four established methods, the biomarkers identified by AggMapNet GFI (AggGFI-biomarkers) are substantially more correlated across five nations (Figure 5B) and have higher levels of cross-nation overlaps (i.e., the biomarkers of the patients of one nation overlap with those of another nation). The top 10 AggGFI-biomarkers in the hot zones of the saliency maps, averaged over five nations (five nation average GFI > 3.5, Figure S10; the GFI scores for all features are provided in Data S1), are unknown *Dialister*, *F. nucleatum* s*. vincentii*, *Anaerococcus obesiensis/vaginalis*, *Parvimonas* sp., *Parvimonas* sp., *Peptostreptococcus anaerobius*, *Alloprevotella tannerae*, *F. nucleatum* s*. animalis*, *Prevotella oris*, and *Porphyromonas somerae.* These microbes have been reported as potential CRC-associated biomarkers.[40–44] The diagnostic relevance of the AggGFI-biomarkers was further evaluated by a simple linear diagnostic indicator, which measures disease-prediction performances of logistic regression (LGR) models developed by the AggGFI-biomarkers on the STST tests of the cross-nation datasets (a total of 20 tests) (Figures 5C and 5D). The LGR models of the Top-N AggGFI-biomarkers outperformed the LGR models of all microbes in these tests. In particular, the LGR models of top 10 AggGFI-biomarkers (avg. GFI > 3.5) outperformed the LASSO and RF LGR models constructed by all 849 microbes. The LGR models of the top 50 AggGFI-biomarkers (avg. GFI > 1.0) even outperformed AggMapNet models of all microbes (average ROC-AUC of the 20 tests: 0.818 versus 0.815). Generally, the performances of the LGR models gradually increase with the increasing number of AggGFI-biomarkers up to 75 biomarkers (avg. GFI > 0.43) (Figure 5D). Beyond this point, the performances decrease with

Figure 5. Comparisons of AggMapNet trained on MEGMA-generated 2D microbiomeprints with commonly used methods for CRC detection and key biomarker identification by cross-nation testing on the CRC-Nation data

(A) The five microbial saliency maps for each nation generated by five different methods of the generalized fold change (FC), WRST q value (FDR-adjusted Wilcoxon rank-sum test p value, transformed by negative logarithms), LASSO model coefficient (coef.), the random forest feature importance (RF FI), and AggMapNet model GFI. The corresponding pixels are the same microbes as shown in the 2D microbiomeprints of Figures 2B and 2C. AUS, Australia; CHN, China; DEU, Germany; FRA, France; USA, United States of America.

(B) The nation-to-nation Pearson correlation coefficient for the five saliency maps in (A). The pairwise Pearson's r was calculated by the vectors that converted from the saliency maps (i.e., simply reshape the 2D saliency maps into the 1D vectors).

(C) The study-to-study transfer (STST) (a total of 20 cross-nation tests) performance for the four models of LGR, LASSO, RF, and AggMapNet using all 849 microbial species in CRC-Nation datasets. The Top-N-LGR model means the LGR model with AggMapNet-identified Top-N key microbial markers (i.e., the microbes in the hot zones of AggMapNet GFI saliency map) as input features. LGR is used as a simple linear diagnostic indicator in STST to test the biomarkers discovered by AggMapNet GFI. LGR, logistic regression; LASSO, least absolute shrinkage and selection operator; RF, random forest.

(D) The performance of LGR models that are constructed on the Top-N AggMapNet-identified biomarkers, the average ROC-AUC of the 20 STST tests are reported. The baselines for the four models of AggMapNet (0.815), RF (0.756), LASSO (0.751), and LGR (0.745) using all microbial species (849) are shown as dashed lines. For each Top-N marker set, the minimal five-nation average AggMapNet GFI value is shown (e.g., the minimal average AggMapNet GFI for top 30 markers is greater than 2, top 50 > 1, and top 75 > 0.43).

increasing number of AggGFI-biomarkers. These results consistently indicate the exceptional ability of MEGMA in selecting diagnostically relevant and reproducible biomarkers; this ability is realized and presented by the AggMapNet microbial saliency maps and MEGMA-generated 2D microbiomeprints.

MEGMA-AggMapNet reveals disease-stage-specific and shifting patterns of microbial compositions

The biomarker analysis capability of the explainable artificial intelligence model AggMapNet with MEGMA-generated 2D microbiomeprint as input was further evaluated on the disease-stage sets of CRC-Stage.[4] This dataset contains a large number of microbes but a low sample size of multiple stages (MP) (stage 0, SI/II, SIII/IV, and healthy controls, Table 1),[4] which is useful for evaluating MEGMA and explainable AggMapNet. The average GFI (calculated based on the 10-fold cross-validation) saliency map for each stage is presented in Figure S12. Compared with the controls, the hot zones are highly distinct across stages but nonetheless share a common zone (zone A). Zone A is particularly important for distinguishing the disease stages from controls, which are dominated by microbes from several genera *Streptococcus* (78 species occupying 36.8% of zone A), *Enterococcus* (27 species, 12.7%), *Lactobacillus* (18 species, 8.5%), *Listeria* (12 species, 5.7%), and *Vagococcus* (9 species, 4.3%). *Streptococcus*, *Enterococcus*, *Lactobacillus*, and *Vagococcus* are lactic acid bacteria involved in the prevention of CRC.[45] Zone B is important for both MP and SI/II, which contain genera *Clostridium* (33 species, 8.7%), *Actinomyces* (28 species, 7.4%), *Bacteroides* (16 species, 4.2%), *Streptomyces* (13 species, 3.4%), and *Blautia* (10 species, 2.6%). Notably, *Actinomyces odontolyticus* has been found to significantly increase only in MP and/or S0.[4] Zone C is mainly important for S0 or SI/II, and contains genera *Thauera*, *Prevotella*, *Undibacterium*, *Vogesella*, and *Sphingobacterium*. Zones D and E are specific for stage SIII/IV. Specifically, zone D includes genera *Lactobacillus* (36, 28.1%), *Clostridium* (24, 18.7%), *Desulfotomaculum* (17, 13.3%), *Selenomonas* (8, 6.3%), and *Fusobacterium* (7, 5.5%). Co-occurrence of these anaerobic bacteria has been reported in colorectal carcinomas.[46] Zone E includes genera *Xenorhabdus* (17, 18.1%), *Vibrio* (14, 14.9%), *Yersinia* (12, 12.8%), *Pantoea* (7, 7.5%), and *Serratia* (6, 6.4%). These distinguished and shifting patterns of microbial composition across disease stages reflect dramatic physiological alterations in the colonic microenvironment during tumorigenesis, which can be captured by AggMapNet.

DISCUSSION

The high-dimensional sparse features and low sample size of metagenomic data hinder in-depth data analysis and learning. Motivated by the characteristics of natural image data and ConvNet-based DL architectures, through data representation and algorithm development, we developed the MEGMA-AggMapNet-GFI pipeline of metagenomic DL for accurate and explainable detection of diseases and for the identification of key signatures.

In relation to accurate disease detection, MEGMA leverages a unique multichannel 2D microbiomeprint transformation to enhance the downstream tasks of DL-based disease prediction and key signature identification. Through the transformation algorithm of MEGMA, high-dimensional, noisy, and sparse metagenomic data can be restructured into highly structured data of multichannel 2D microbiomeprints (3D tensor, where the highly correlated microbes are aggregated into multiple channels in the third dimension for amplifying the abundance signals). MEGMA enhances the downstream tasks by combining algorithms of manifold-guided microbial embedding (unordered 1D data re-arranged into locally coherent and connective 2D image data) and MG/TG-based microbial grouping (to form the multichannel character according to phenotype/genotype distances). This pipeline of ME and grouping largely improves the performance of DL models for disease prediction.

For reproducible signature discovery, the pipeline leverages the GFI and 2D microbiomeprint saliency map of AggMapNet models for robust and replicable key signature identification. The GFI revealed that important signatures (or biomarkers) are aggregated in specific regions of the MEGMA-generated 2D microbiomeprints, which allows the finding of the disease-related microbes more intuitively and precisely. Drawing on these advantages, we applied MEGMA and AggMapNet for the systematic identification of the key microbes for the early detection of CRC in different nations and the progression of CRC in different stages. The study not only revealed the consistency of the microbial markers among different populations from different nations but also identified the shift of microbes during the CRC disease progression of patients at different stages.

While the MEGMA-AggMapNet-GFI pipeline exhibits several distinguished advantages, it nonetheless has two limitations. The first limitation is in the MEGMA 2D representation. Although the 2D regular grid representation is capable of generating 2D spatially correlated Fmaps (i.e., the Euclidean data) and can be learned by ConvNet-based DL models, it may not accurately preserve the correlation information between FPs because the correlation distance between two FPs in the regular grid is not an identity map to the original correlation distances between the two FPs. The FPs are uniformly distributed in a regular grid, while the FPs are not uniformly distributed after 2D embedding. Therefore, further research using a non-Euclidean data representation (such as a weighted topological graph in Figure 2B) and a graph neural network may achieve better learning performances. The second limitation is in the GFI scores. The GFI scores may be affected by the first layer kernel size of the ConvNet; a larger convolutional kernel may lead to false-positive signatures, because a larger kernel size has a larger receptive field[47] on the input data. Such false-positive signatures may be partly removed by combining the MEGMA-AggMapNet-GFI pipeline with generalized FC and statistical methods.

Despite these disadvantages, the results of this work suggested that the distinguished advantages of our MEGMA-AggMapNet-GFI pipeline in metagenomic DL may potentially find broad applications in various biomedical fields. Clinically, individualized MEGMA-derived 2D microbiomeprints may help to make "stratified" medicine more "precise." The further developed MEGMA and explainable AggMapNet tools together with other advanced methods are useful for facilitating non-invasive diagnostic, prognostic, and theragnostic tasks and biomarker discovery.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Yu Zong Chen (chenyuzong@sz.tsinghua.edu.cn).

Materials availability

This study did not generate new unique materials.

Data and code availability

The Disease-Set1, Disease-Set2, CRC-Nation sets, and CRC-Stage sets are available in the GitHub repository https://github.com/shenwanxiang/metageomic_datasets, and were released in Zenodo (https://doi.org/10.5281/zenodo.7146901).

MEGMA and AggMapNet were released in the omics data representation and learning package bidd-aggmap (https://pypi.org/project/aggmap/1.1.7/). The source code and jupyter notebooks on MEGMA are available at GitHub (https://github.com/shenwanxiang/aggmap-megma/), a tutorial PAGE is available at: https://shenwanxiang.github.io/aggmap-megma/. To run the code and test the repeatability, please follow the link on Google Colab: https://colab.research.google.com/drive/1T4nAtK-CT_1IxfYHd1kgLMjVeR59_III.

Human gut microbial metagenomic datasets

The datasets used in this study are summarized in Table 1. All datasets used in our study were directly obtained from the processed data and results in previous studies. The human gut microbial taxa of all datasets are all at the species level from fecal samples and with relative abundance values for the specific disease and control groups. These datasets cover a different number of species, and the taxonomic profiles of the two disease sets (Disease-Set1[8,9] and Set2[16,17]), the CRC-Nation[3] and the CRC-Stage[4] were generated from the MetaPhlAn2[19] pipeline, the mOTU2 pipeline,[20] and the alignment to the All-Species LTP of the SILVA database (SILVA LTP),[21] respectively.

The Disease-Set1 benchmark datasets were directly obtained from Reiman et al.'s studies of PopPhy-CNN[8] and Meta-Singer,[9] while the Disease-Set2 benchmark datasets were directly obtained from Met2Img of Nguyen et al.[16,17] Although both Disease-Set1 and Set2 contain the same original datasets of the same five diseases (cirrhosis, obesity, T2D, IBD, and CRC),[9,29,36,41,48–50] they are nevertheless different in the numbers of cases/controls, microbes, and features because of different data processing and curation methods used in Nguyen et al.'s studies[16,17] and Reiman et al.'s studies.[8,9] For example, the number of cases and controls for cirrhosis datasets in disease sets 1 (114 versus 118) and 2 (118 versus 114) are reversed. In addition, the T2D in disease sets 1 and 2 have a different number of samples and features. Moreover, the CRC dataset of Disease-Set1 is a multi-class dataset that covers 507 gut microbial species from fecal samples of 48 cancers, 39 adenomas, and 47 healthy controls, while the CRC dataset of Disease-Set2 contains 503 gut microbial species from fecal samples of 48 cancers and 73 controls. In this study, to eliminate the differences in model performance caused by different datasets and to make a fair comparison with existing methods, such as PopPhy-CNN[8] or Met2Img,[16,17] we used both data groups of Disease-Set1 and Disease-Set2 (Table 1). All these Disease-Sets are logarithm transformed (after adding a pseudo-count of 1×10^{-8} to avoid non-finite values resulting from log(0))[51] and then with Z score standardization or min-max normalization, and the performance of the disease prediction model was evaluated by stratified 10-fold cross-validation repeated 10 times (different seed was used for each repeat).

The two sets of CRC metagenomic datasets are the across-nation CRC data (CRC-Nation sets) and CRC stage-specific data (CRC-Stage sets), which are obtained from recent studies by Wirbel et al.[3] and Yachida et al.,[4] respectively. The CRC-Nation data covers 849 gut microbial species from fecal samples of 575 subjects (CRCs or healthy CTRs) compiled by Wirbel et al.[3] The taxonomic profiles were generated with the mOTU profiler v.2.0.0[20] in the study of Wirbel et al.[3] It contains five metagenomic datasets from five separate studies in five nations: France, Australia, Germany, China, and the US). All the datasets are logarithm transformed (after adding a pseudo-count of 1×10^{-8} to avoid non-finite values resulting from log(0))[51] then with Z score standardization, and the STST model (trained on one nation data and tested on the rest nations) was used for the evaluation of the performance of the metagenomic prediction models and the consistency of the across-nation biomarker discovery.

The CRC-Stage data are each of the four CRC stages in pairwise comparisons against healthy controls, which was adapted from the study of Yachida et al.[4] It consists of 317 subjects from the Japanese cohort with both metagenomic and metabolomic profiles, where 127 subjects are in the healthy stage, 40 subjects are in stage MP (multiple polypoid adenomas with low-grade dysplasia), 27 subjects are in stage S0 (intramucosal carcinoma polypoid adenoma with high-grade dysplasia), 69 subjects are in stage SI/II CRC, and 54 subjects are in stage SIII/IV CRC.[4] It originally covers a total of 8,367 gut microbial species by the alignment from All-Species LTP of the SILVA database (SILVA LTP),[21] we removed the species with very low abundance values (smaller than 1e−8) by calculating the average relative abundance in each stage, leading to 7,278 species. All the datasets are logarithm transformed (after adding a pseudo-count of 1×10^{-8} to avoid non-finite values resulting from log(0))[51] then with Z score standardization, and the stratified 10-fold cross-validation (repeated 5 times with different random seeds) was used for the evaluation of stage prediction model and identification of key microbial signatures of shifts in different stages of CRC.

Manifold/random-guided embeddings

To form the 2D microbiomeprints in MEGMA, the first step is to embed the microbes into 2D space based on their abundant correlations. Randomized embedding was used as the control embedding method and ME was used to increase the local spatial coherence of the 2D microbiomeprints. Because ME can expose the topological connections of the microbes, which enables the correlated microbes in their abundance to be the neighbors in the embedding 2D space. Specifically, RUE and five ME approaches, including MDS, ISOMAP, LLE, t-SNE, and UMAP, have been explored in this study.

RUE

RUE was used to make a random-guided but uniform embedding and distribution in 2D space: first, two probability density functions (PDFs) ρ and φ were built and the coordinates (x and y) of the 2D embedding were transformed by the cosine and sine functions based on the PDFs:

$$\rho(x) = \sqrt[2]{P(0 \le x \le 1)}$$

$$\varphi(x) = P(0 \le x \le 4\pi)$$

$$x = \rho * \cos\varphi$$

$$y = \rho * \sin\varphi.$$

Each microbe will be assigned the 2D coordinates in REU. RUE generates the totally unordered and randomized but uniform embedding, the neighbors of each microbe will be the random microbes instead of highly correlated microbes in abundance. The 2D microbiomeprints generated by RUE are similar to images with randomly permutated pixels, that is, the spatial coherence of the 2D microbiomeprints is fully destroyed.

ME

Different from RUE, five ME methods were used for microbe embedding by considering the pairwise abundance correlation of microbes. Two microbes (i.e., the feature points or data points) are considered similar if they are highly correlated in their abundances, where Pearson correlation coefficient $r(i,j)$ was used for measuring the distance d_{ij} of microbial pairs i and j:

$$d_{ij} = 1 - r(i,j) = 1 - \frac{\sum_{a=1}^{n}(i^a - \bar{i})(j^a - \bar{j})}{\sqrt{\sum_{a=1}^{n}(i^a - \bar{i})^2 \sum_{a=1}^{n}(j^a - \bar{j})^2}}.$$

Based on this distance, highly correlated microbes can be drawn closer to each other for data restructuring embedding in a 2D space by using ME methods. Five ME methods are used in this study: MDS, LLE, ISOMAP, t-SNE, and UMAP.

MDS. The metric MDS method attempts to model the similarity/dissimilarity of data by calculating distances between each pair of points using their geometric coordinates. The MDS maps the precomputed distance matrix d_{ij} to 2D space while preserving those distances between points (microbes) as well as possible. It minimizes the cost function called "Stress," which is a residual sum of squares:

$$Stress_D(x_1, x_2, \ldots x_n) = \left(\sum_{i \neq j = 1,\ldots,N} \left(d_{ij} - \left\| x_i - x_j \right\| \right)^2 \right)^{1/2},$$

where d_{ij} is the precomputed correlation distance matrix, the $\|x_i - x_j\|$ term is the distance between the two corresponding data points (microbes) in their lower-dimensional space (2D space). The closer the value of $\|x_i - x_j\|$ is to d_{ij}, the smaller will be the value of the stress loss.

ISOMAP. ISOMAP can be viewed as an extension of MDS, it seeks a lower-dimensional embedding that maintains geodesic distances between all points and enables to preserve the non-linear relationships between data points.[52] The algorithm comprises three steps:

1) Find the k-nearest neighbors of every data point based on the distance matrix d_{ij} by kNN approach, where the k is the hyperparameter.
2) Construct the neighborhood graph where points are connected to each other if they are each other's neighbors, and then compute the shortest path between each pair of data points (nodes), i.e., the geodesic distance between points.
3) MDS is used to compute lower-dimensional embedding. Given distances between each pair of points are known, MDS places each object into the 2D space such that the between-point distances are preserved as well as possible.

LLE. The LLE seeks a lower-dimensional projection of the data, which preserves distances within local neighborhoods.[53] It can be thought of as a series of local principal-component analyses, which are globally compared to find the best non-linear embedding. Similar to ISOMAP, LLE combines three major steps to produce the lower-dimensional embedding. These are:

1) Find the k-nearest neighbors of every data point based on the distance matrix d_{ij} by the kNN approach, where k is the hyperparameter.
2) Compute the weights W_{ij} that best linearly reconstruct point X_i from its k-nearest neighbors (every point is a linear combination of its neighbors), solving the constrained least-squares problem (i.e., to minimize the cost function $\varepsilon(W)$):

$$\varepsilon(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2, \quad \text{such that} \sum_j W_{ij} = 1$$

3) Perform the low-dimensional embedding:
4) Compute the low-dimensional embedding vectors Y_i best reconstructed by W_{ij}, minimizing $\varnothing(Y)$ by finding the smallest eigenmodes of the sparse symmetric matrix: $\varnothing(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2$

Note that although the weights W_{ij} and vectors Y_i are computed by linear methods, the constraint that points are only reconstructed from neighbors can result in highly non-linear embeddings.[53]

t-SNE. t-SNE[54] converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler (KL) divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The algorithm contains three steps:

1) Converting the input distance matrix d_{ij} into conditional probabilities $p_{j|i}$ that represent similarities:

$$p_{j|i} = \frac{\exp\left(-d_{ij}/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-d_{ik}/2\sigma_i^2\right)}.$$

Here, σ_i is the variance of the Gaussian and is a tuning parameter, which is usually determined based on a certain perplexity measure and a simple binary search. The conditional probabilities $p_{j|i}$ has to satisfy the symmetry condition:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, 1 \leq i, j \leq n.$$

2) Similarly, the joint probabilities q_{ij} in a 2D map are represented by Student's t distributions with one degree of freedom:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}.$$

3) To minimize the KL divergence between the two joint probability distributions in the original space p_{ij} and the embedded space q_{ij} by gradient descent:

$$KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

While Isomap and LLE are best suited to unfold a single continuous low-dimensional manifold, t-SNE will focus on the local structure of the data and will tend to extract clustered local groups of samples. This ability to group samples based on the local structure might be beneficial to visually disentangle a dataset that comprises several manifolds at once.

UMAP. Compared with t-SNE, UMAP constructs two weighted graphs in high-dimensional space (input) and low-dimensional space (embedding) and minimizes the differences of the two graphs. Similar to t-SNE, the UMAP algorithm also contains three major steps for the embedding:

1) To build a weighted topological k-neighbor graph P by exponential probability distribution using correlation distance d_{ij}:

$$p_{i|j} = exp(-(d_{ij} - \rho_i) / \sigma_i),$$

where ρ_i represents the distance from i-th point to its first nearest neighbor; σ_i is a (smoothed) normalization factor. The adjacency matrix $p_{i|j}$ has to satisfy the symmetry condition:

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j} p_{j|i}.$$

2) To build a weighted graph Q in low dimension:

$$q_{ij} = \left(1 + a(y_i - y_j)^{2b}\right)^{-1},$$

where q_{ij} is the weight matrix of the low-dimensional neighbor graph Q, y_i and y_j are the initial embedding coordinates, and a and b are hyperparameters.

3) To minimize the error between the two topological representations P and Q. The graph layout for Q is force directed, the forces are derived from gradients optimizing the edgewise cross-entropy:

$$CE(X, Y) = P\left(X\right) \log\left(\frac{P(X)}{Q(Y)}\right) + \left(1 - P(X)\right) \log\left(\frac{1 - P(X)}{1 - Q(Y)}\right),$$

where $CE(X, Y)$ is the total cross-entropy loss over all the edge existence probabilities between weighted graphs P and Q. Minimization of $CE(X, Y)$ will let the low-dimensional representation settle into a state that relatively accurately represents the overall topology of the source data.

Among these embedding methods, UMAP was chosen as the default embedding method in our aggmap package (https://pypi.org/project/aggmap/) because of its effectiveness in aggregating similar data points while preserving their relative proximities for both local and global data structures. The hyperparameters in these ME methods have impacts on the embedding results. For example, the N neighbors UMAP provides the trade-off for persevering the local/global structure. Technically, the hyperparameters in ME will affect the constructed 2D microbiomeprint, which in turn affects the learning performance of AggMapNet models. To avoid the impacts of different parameters in ME methods, the default settings in all

ME methods are used for the 2D embedding of microbes. In the test of the AggMapNet model performance of 2D microbiomeprints generated by various embedding methods, the default architecture parameters and training control parameters are used, and 50 epochs that reached convergence were used by all experiments to eliminate the effects of the AggMapNet hyperparameters.

MG/TG-based groupings

A color image is the multichannel data (e.g., the RGB channels) that contains more abundant information than a grayscale image, and multichannel networks are helpful for learning complex data by separately learning feature subsets.[55] Their representational richness often allows capturing of non-linear dependencies at multiple scales.[56] The shape of the multichannel 2D microbiomeprint is (−1, w, h, c), where c is the number of channels that can be greater than 3 and generated by a grouping method. In this study, we used each color to represent one channel. The multichannel (i.e., the number of microbial subgroups or colors) can be generated by truncating the phylogenetic tree (taxonomic-based grouping) or the hierarchical clustering tree (MG-based grouping) at given levels. The truncated clusters (the channel parameter c) will be separately embedded into different channels to generate the multichannel character of the 2D microbiomeprints, making the inputs more information abundant.

For the MG-based grouping of the microbes, a hierarchical clustering tree is constructed based on the microbial metagenomic abundance correlation distance matrix d_{ij}, the bottom-up agglomeration algorithm was used for group merging. The complete linkage method is used to measure inter-cluster distances. The number of clusters (parameter c) is a hyperparameter, a higher number of clusters will generate fine-grained separations of the microbes.

For the TG-based grouping of the microbes, the phylogenetic tree of the microbes is used. The phylogenetic tree is constructed based on the known taxonomy classification of the microbes, i.e., the kingdom, phylum, class, order, family, and genus levels. Different from the MG-based grouping to directly decide the number of the clusters, taxonomic-based grouping truncates different taxonomy levels (e.g., the genus level) as grouping, and different levels will generate different numbers of clusters, i.e., the numbers of the channels for the 2D microbiomeprint Fmaps.

Taken together, MG is based on the phenotype distance of microbes, while TG is based on the genotype distance of microbes. To compare the performance of the models that are trained on MG and TG 2D microbiomeprint in Disease-Set1, we used the different levels (e.g., the genus level) of the taxonters (parameter c) was used in MG to generate the clusters. To eliminate the effects of the AggMapNet hyperparameters, the default architecture parameters and training control parameters are used, and 50 epochs that reached convergence were used to train the models.

Mapping and transformation

The mapping is to form the 2D grid and the transformation is to form the 2D microbiomeprint Fmaps. The mapping operation includes position mapping and channel mapping. In the position mapping, the 2D embedded microbes were further mapped into regular 2D grids by the linear assignment algorithm. The Jonker-Volgenant (J-V) algorithm[57] was used for the assignment, which preserves the 2D embedded neighborhood relationships while the microbes are assigned into the grid points. Specifically, we first calculated the pairwise squared Euclidean distance between two microbes from the matrices of 2D embedding and 2D regular mesh grid:

$$d = \left(x^{embed} - x^{grid}\right)^2 + \left(y^{embed} - y^{grid}\right)^2,$$

where (x^{embed}, y^{embed}) is the 2D coordinates of an FP, (x^{grid}, y^{grid}) is the 2D coordinates of a point in the mesh grid. The squared Euclidean distance matrix is the N*N size matrix (where N is the number of the microbes), which was further used as the cost matrix to solve the linear assignment problem (LAP) using the J-V algorithm. The J-V algorithm finds an optimal solution to the global nearest neighbor assignment problem by finding the set of assignments that minimize the total cost of the assignments. In channel mapping, the microbes are assigned to different channels based on the subgroups. Finally, each microbe has an optimized position in the microbial 2D regular grid map, and its neighbors are the highly correlated microbes by ME.

Based on the 2D grid location map and the grouping information of the microbes, the microbial abundance 1D vector data of each sample will be further transformed into a multichannel 2D microbiomeprint (3D data), which is used as the input of the AggMapNet DL models.

ConvNet-based AggMapNet modeling

As shown in Figure 1B, we designed the ConvNet-based AggMapNet[18] with MEGMA-generated 2D microbiomeprints as inputs. AggMapNet consists of three major parts, the input Fmaps are the multichannel microbiomeprints (the 3D tensor data), the CNN-based feature extraction layers, and pyramid fully connected layers. This network structure has a relatively small number of trainable parameters (~0.3 million) but with two inception blocks. In the first convolutional layer, we used a larger kernel number (48) for increased data dimension. The max-pooling layer (kernel size = 3) was used for down-sampling with a stride size of 2 after each convolution layer; thus, the spatial resolution of the Fmaps is reduced more aggressively for lowering computation cost. Choosing the right kernel size for the convolution operation is difficult because different tasks may favor different kernel sizes. To improve AggMapNet performance for general tasks, we adopted the naive inception layer of GooLeNet[58] (a top-performer of the ILSVRC-2014 classification task). The inception layer in AggMapNet only consists of three small parallel kernels (sizes of 1×1, 3×3, and 5×5) to enhance the local perception. Subsequently, the global max-pooling layer was used before the dense layer instead of a flattened layer, which significantly decreases the number of parameters, followed by dense layers for improved non-linear transformation capability.

The hyperparameters AggMapNet include the network architecture parameters (NAPs) and the training-control parameters (TCPs). The NAPs are the kernel size of the first convolutional layers (conv1_kernel_size), the number of dense layers, and corresponding units (dense_layers). Larger kernel sizes or dense layer units increase the model parameters, and a larger kernel size allows for more expressive power and a global perception.[59] The TCPs include the number of epochs, learning rate, and batch size. The cross-entropy loss was used for both multi-task and binary tasks. During the training, AggMapNet has two important parameters (the monitor and patience) for early stopping. The monitor is the metric performance of the validation set, and the patience parameter is the number of epochs with no improvement on the monitor after which training will be stopped. The early stopping strategy was used in the nested cross-validation to find the best number of epochs.

Model evaluation metrics

Due to imbalanced classification labels for some datasets, using only the accuracy metric cannot measure the performance of the model. Therefore, we use multiple metrics for performance measurement, including F1 score for unbalanced situations. Specifically, the following metrics were used for the model evaluation in the benchmark Disease-Set1 and Disease-Set2: the AUC (binary classification only), Matthew's correlation coefficient (MCC), accuracy (ACC), Precision, Recall, and F1 score. Here,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ (TPR) = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$AUC = \int_0^1 TPR\ \mathrm{d}\ (FPR)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Input: Trained model $\boldsymbol{f}$, feature matrix $\boldsymbol{X}$, feature size $\boldsymbol{N}$, training sample size $\boldsymbol{M}$, target true label vector $\boldsymbol{y}$, error measure $\boldsymbol{L}(\boldsymbol{y}, \boldsymbol{f})$. To estimate this error $\boldsymbol{L}$, the log loss(cross-entropy) is used for the classification model

(1) Estimate the original model error: $\boldsymbol{e^{orig}} = \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{X})) = -\frac{1}{\boldsymbol{M}}\sum_{\boldsymbol{j}=1}^{\boldsymbol{M}} \boldsymbol{y_j} * \boldsymbol{log}(\boldsymbol{f}(\boldsymbol{X})_{\boldsymbol{j}}) + (1-\boldsymbol{y_j}) * \boldsymbol{log}(1-\boldsymbol{f}(\boldsymbol{X})_{\boldsymbol{j}})$. The model error $\boldsymbol{e}$ is calculated by the log loss of the prediction values versus true labels for one class by many samples $\boldsymbol{M}$.

(2) For each feature point $\boldsymbol{i} = 1, \ldots, \boldsymbol{N}$ do:

(i) Generate feature matrix $\boldsymbol{X_i^{pert}}$ by replacing feature $\boldsymbol{i}$ with the minimal value (background value) in the data $\boldsymbol{X}$. This breaks the association between feature $\boldsymbol{i}$ and true outcome $\boldsymbol{y}$.

(ii) Estimate error $\boldsymbol{e_i^{pert}} = \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{X_i^{pert}}))$ based on the predictions of the perturbed data.

(iii) Calculate perturbation FI score: $\boldsymbol{S_i} = \boldsymbol{e_i^{pert}} - \boldsymbol{e^{orig}}$

(3) Sort features by descending FI score $\boldsymbol{S}$.

(4) Optional corrections: applying a Z score standard scaling or logarithm transformation on $\boldsymbol{S}$

$$MCC = 2 * \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP(R), FP(R), TN, and FN are the true positive (rate), false positive (rate), true negative, and false negative, respectively.

Comparison of MEGMA-AggMapNet with existing methods on Disease-Set1 and Set2

To compare MEGMA-AggMapNet with existing ML models reported by Reiman et al.,[8,9] the AggMapNet models were trained using stratified 10-fold cross-validation on Disease-Set1.[8,9] For each training set of the fold, the nested cross-validation was used to find the best parameters. For MEGMA, the default UMAP embedding method was used and the number of channels was tuned. For AggMapNet, the training parameters, such as batch size and the number of epochs, were tuned. All these parameters were tuned only in the first fold of the nested cross-validation and were subsequently used in other folds. The average and SD of each evaluating criteria from 10 times (repeated with 10 different random seeds of 8, 16, 32, 64, 128, 256, 1,024, 2,048, 4,096, and 8,192 for the splitting training and test set) cross-validations are reported. The average performance and SD were calculated to compare the performance, and the paired t test (100 pairs) was used to test the significance of the difference between the performance of ML models and MEGMA-AggMapNet performance. The comparison results are in Table S1.

For the comparison of Met2Img[16,17] and MEGMA by the performance AggMapNet models on Disease-Set2, the Met2Img[16,17] and MEGMA methods were used to generate the inputs of AggMapNet with the same hyperparameters. For the Met2Img[16,17] method, the deepmg (https://pypi.org/project/deepmg/1.0.33) python package was used for generating the "synthetic images," the best method Fill-up with Species Bin and grays (fillup-spb-gray) or jet color map (fillup-spb-jet) were used as described in the paper by Nguyen et al.[16,17] For MEGMA, the aggmap (https://pypi.org/project/aggmap/1.1.7) python package was used for generating the inputs, and the following parameters were used: the default embedding method "umap," the default grouping channel number "5" (umap-c = 5), and higher grouping channel number "60" (umap-c = 60). The AggMapNet models were trained on Disease-Set2[16,17] by using the same default parameters with 50 epochs. The performances of Met2Img and MEGMA were tested by 10-fold cross-validation repeated 10 times with 10 different random seeds (8, 16, 32, 64, 128, 256, 1,024, 2,048, 4,096, and 8,192), a total of 100 runs. The average performance was reported, and the SD was shown in parentheses. The paired t test (100 pairs) was used for evaluating the significance of the difference between Met2Img and MEGMA performance. The comparison results are in Table S2.

Model explanation and microbial saliency map

In computer vision, a saliency map displays the degree of importance of every pixel. The saliency map of an input image specifies parts of it that contribute the most to the activity of a specific layer in the network, or the decision of the network as a whole.[22] In this study, the revealed important microbial features by AggMapNet model explanation were presented as a saliency map. The 2D grid location map of microbial species can be generated by MEGMA for each dataset. Based on the location map, the microbial saliency map can be generated from the importance score of each microbial species (i.e., the GFI score).

The microbial saliency map is an image that highlights the region on which the host phenotype prediction models focus first. The goal of a microbial saliency map is to reflect the degree of importance of a microbe to the specific host phenotype prediction model. The microbial saliency map aggregates the IMs into hot zones by their abundance correlations, which helps us to pick up the correlated microbes and recognize a specific microbial correlation pattern that is critical for the host phenotype.

The global feature (microbe) importance score of the trained AggMapNet model was calculated based on the training set using the perturbation-based (model-independent post hoc feature attribution) method Simply-explainer. In Simply-explainer, GFI score $\boldsymbol{S}$ is calculated by replacing each feature point value (i.e., the abundance value of each microbe) with a background value, without retraining the model:

The AggMapNet inputs are 4D tensor (batch size, width, height, channels) data in multiple channels. The perturbation value is a background microbial abundance value (e.g., zero value for blank pixel) of the input Fmaps. The correction of the GFI of Simply-explainer includes the logarithm transformation and standard scaling of the GFI values to reveal the IMs. After scaling, those microbes with GFI score >0 are considered notable microbes in a saliency map.

Statistical tests and ML methods for key biomarker identifications

Various metagenomic studies use univariate and multivariable statistical tests to detect differentially abundant individual taxa between disease and control groups.[9] Linear model coefficients and RF FI are also commonly used methods for ranking and selecting key features. In this study, the following methods were used for making comparisons with the AggMapNet model GFI to discover replicable CRC-biomarkers in the CRC-Nation set of CRC disease.

q value

For univariate association testing between the abundances of microbial taxa and host phenotypes (e.g., CRC versus CTR), we used the nonparametric tests method of two-sided WRST.[3] Abundances of each species were determined to be significantly elevated or depleted in the disease cases by pairwise comparison with the healthy controls using two-sided WRSTs. $p < 0.05$ was considered statistically significant. Furthermore, the q value was estimated by a Benjamini-Hochberg FDR-corrected p value.

Generalized FC

We further used the generalized (logarithmic) FC that was developed and used in the study of Wirbel et al.[3] The generalized FC is designed to have better resolution for sparse microbiome profiles, which is calculated as the mean difference in a set of predefined quantiles of the logarithmic CTR and CRC distributions. The generalized FC extends the established (median-based) FC to provide higher resolution in sparse microbiome data.[3]

LASSO coefficients

The LASSO is a multivariate linear regression modeling method, it uses L1 regularization to shrink the coefficients of the variables. This type of regularization can result in sparse models with few coefficients; some coefficients can become zero and be eliminated from the model. LASSO is a widely used method to select predictive microbial features and eliminate uninformative ones in metagenomic disease classification models.[3,4] Therefore, LASSO was used as the third statistical method of identification of disease-associated

gut microbial species in this study. The LASSO models were built to classify the CRC or different stages from CTR similar to AggMapNet on the metagenomic datasets of CRC-Nation and CRC-Stage. The important parameter alpha (a constant that multiplies the L1 regularization term) of LASSO was optimized by nested cross-validations. The variable's coefficients of the well-trained LASSO model were used as the model-adapted FI, that is, the importance score of the microbial features.

RF FI

RF is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset, it uses averaging strategy to improve the predictive accuracy and control over-fitting problems. RF is also a commonly used metagenomic ML method,[4,9,15] the FI of the RF model has been used to identify the biomarkers for disease model prediction.[4,9] Therefore, RF FI was used as the fourth method of identification of disease-associated gut microbial species in this study. The RF models and FI were constructed and calculated by the Python sklearn (https://scikit-learn.org/) package. The hyperparameters of "n_estimators," "min_samples_split," "min_samples_leaf," "max_features," "max_depth," and "bootstrap" in RF were tuned by the nested cross-validations using the RandomizedSearchCV module of sklearn package.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100658.

ACKNOWLEDGMENTS

We would like to express our thanks to the anonymous reviewers for their contributions, whose constructive suggestions greatly improved the quality of the article. We appreciate the financial supports from the National Key R&D Program of China, Synthetic Biology Research (2019YFA0905900), Shenzhen Municipal Government grants (nos. 2019156, JCYJ20170413113448742, and 201901), the Department of Science and Technology of Guangdong Province (no. 2017B030314083), and Singapore Academic Funds R-148-000-273-114 and NUS Research Scholarships.

AUTHOR CONTRIBUTIONS

Conceptualization, W.X.S. and Y.Z.C.; methodology, W.X.S. and Y.Z.C.; software, W.X.S.; validation, S.R.L.; formal analysis, W.X.S. and S.R.L.; investigation, W.X.S.; resources, Y.Y.J. and Y.Z.C.; data curation, W.X.S. and S.R.L.; writing – original draft, W.X.S. and Y.Z.C.; writing – review & editing, W.X.S. and Y.Z.C.; visualization, W.X.S.; supervision, Y.Y.J. and Y.Z.C.; project administration, W.X.S. and Y.Z.C.; funding acquisition, Y.Y.J. and Y.Z.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: June 1, 2022
Revised: July 15, 2022
Accepted: November 15, 2022
Published: December 15, 2022

REFERENCES

1. Cho, I., and Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. Nat. Rev. Genet. *13*, 260–270. https://doi.org/10.1038/nrg3182.
2. Tjalsma, H., Boleij, A., Marchesi, J.R., and Dutilh, B.E. (2012). A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat. Rev. Microbiol. *10*, 575–582. https://doi.org/10.1038/nrmicro2819.
3. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., and Ponnudurai, R. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689. https://doi.org/10.1038/s41591-019-0406-6.
4. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976. https://doi.org/10.1038/s41591-019-0458-7.
5. Pasolli, E., Truong, D.T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput. Biol. *12*, e1004977. https://doi.org/10.1371/journal.pcbi.1004977.
6. Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. BMC Bioinf. *19*, 227. https://doi.org/10.1186/s12859-018-2205-3.
7. Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., and Furlanello, C. (2018). Phylogenetic convolutional neural networks in metagenomics. BMC Bioinf. *19*, 49. https://doi.org/10.1186/s12859-018-2033-5.
8. Reiman, D., Metwally, A.A., Sun, J., and Dai, Y. (2020). PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE J. Biomed. Health Inf. *24*, 2993–3001. https://doi.org/10.1109/JBHI.2020.2993761.
9. Reiman, D., Metwally, A.A., Sun, J., and Dai, Y. (2020). Meta-signer: metagenomic signature identifier based on rank aggregation of features. Preprint at bioRxiv. https://doi.org/10.1101/2020.05.09.085993.
10. Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu. Rev. Stat. Appl *2*, 73–94. https://doi.org/10.1146/annurev-statistics-010814-020351.
11. Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host variables confound gut microbiota studies of human disease. Nature *587*, 448–454. https://doi.org/10.1038/s41586-020-2881-9.
12. Teschendorff, A.E. (2019). Avoiding common pitfalls in machine learning omic data science. Nat. Mater. *18*, 422–427. https://doi.org/10.1038/s41563-018-0241-z.
13. Knights, D., Parfrey, L.W., Zaneveld, J., Lozupone, C., and Knight, R. (2011). Human-associated microbial signatures: examining their predictive value. Cell Host Microbe *10*, 292–296. https://doi.org/10.1016/j.chom.2011.09.003.
14. Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M.J., Gasbarrini, A., and Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat. Rev. Gastroenterol. Hepatol. *17*, 635–648. https://doi.org/10.1038/s41575-020-0327-3.
15. Prifti, E., Chevaleyre, Y., Hanczar, B., Belda, E., Danchin, A., Clement, K., and Zucker, J.D. (2020). Interpretable and accurate prediction models for metagenomics data. GigaScience *9*, giaa010. https://doi.org/10.1093/gigascience/giaa010.
16. Nguyen, T.H., Prifti, E., Chevaleyre, Y., Sokolovska, N., and Zucker, J.-D. (2018). Disease classification in metagenomics with 2d embeddings and deep learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1806.09046.
17. Nguyen, T.H., Prifti, E., Sokolovska, N., and Zucker, J.-D. (2019). Disease Prediction Using Synthetic Image Representations of Metagenomic Data and Convolutional Neural Networks (IEEE), pp. 1–6. https://doi.org/10.1109/RIVF.2019.8713670.
18. Shen, W.X., Liu, Y., Chen, Y., Zeng, X., Tan, Y., Jiang, Y.Y., and Chen, Y.Z. (2022). AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks. Nucleic Acids Res. *50*, e45. https://doi.org/10.1093/nar/gkac010.
19. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for

enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903. https://doi.org/10.1038/nmeth.3589.

20. Milanese, A., Mende, D.R., Paoli, L., Salazar, G., Ruscheweyh, H.J., Cuenca, M., Hingamp, P., Alves, R., Costea, P.I., Coelho, L.P., et al. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. Nat. Commun. *10*, 1014. https://doi.org/10.1038/s41467-019-08844-4.

21. Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F.O. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Res. *42*, D643–D648. https://doi.org/10.1093/nar/gkt1209.

22. Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6034.

23. Albawi, S., Mohammed, T.A., and Al-Zawi, S. (2017). Understanding of a Convolutional Neural Network (IEEE), pp. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

24. Kobak, D., and Linderman, G.C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat. Biotechnol. *39*, 156–157. https://doi.org/10.1038/s41587-020-00809-z.

25. McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

26. Shen, W.X., Zeng, X., Zhu, F., Qin, C., Tan, Y., Jiang, Y.Y., and Chen, Y.Z. (2021). Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. Nat. Mach. Intell. *3*, 334–343. https://doi.org/10.1038/s42256-021-00301-6.

27. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. https://doi.org/10.48550/arXiv.1811.12231.

28. Ruuskanen, M.O., Åberg, F., Männistö, V., Havulinna, A.S., Méric, G., Liu, Y., Loomba, R., Vázquez-Baeza, Y., Tripathi, A., and Valsta, L.M. (2021). Links between gut microbiome composition and fatty liver disease in a large population sample. Gut Microb. *13*, 1888673. https://doi.org/10.1080/19490976.2021.1888673.

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

30. Oh, T.G., Kim, S.M., Caussy, C., Fu, T., Guo, J., Bassirian, S., Singh, S., Madamba, E.V., Bettencourt, R., and Richards, L. (2020). A universal gut-microbiome-derived signature predicts cirrhosis. Cell Metabol. *32*, 878–888.e6. https://doi.org/10.1016/j.cmet.2020.06.005.

31. Loomba, R., Ling, L., Dinh, D.M., DePaoli, A.M., Lieu, H.D., Harrison, S.A., and Sanyal, A.J. (2021). The commensal microbe V eillonella as a marker for response to an FGF19 analog in NASH. Hepatology *73*, 126–143. https://doi.org/10.1002/hep.31523.

32. Lapidot, Y., Amir, A., Nosenko, R., Uzan-Yulzari, A., Veitsman, E., Cohen-Ezra, O., Davidov, Y., Weiss, P., Bradichevski, T., Segev, S., et al. (2020). Alterations in the gut microbiome in the progression of cirrhosis to hepatocellular carcinoma. mSystems *5*. e00153–e00120. https://doi.org/10.1128/mSystems.00153-20.

33. Sung, C.M., Lin, Y.F., Chen, K.F., Ke, H.M., Huang, H.Y., Gong, Y.N., Tsai, W.S., You, J.F., Lu, M.J., Cheng, H.T., et al. (2019). Predicting clinical outcomes of cirrhosis patients with hepatic encephalopathy from the fecal microbiome. Cell. Mol. Gastroenterol. Hepatol. *8*, 301–318.e2. https://doi.org/10.1016/j.jcmgh.2019.04.008.

34. de Meij, T.G.J., de Groot, E.F.J., Peeters, C.F.W., de Boer, N.K.H., Kneepkens, C.M.F., Eck, A., Benninga, M.A., Savelkoul, P.H.M., van Bodegraven, A.A., and Budding, A.E. (2018). Variability of core microbiota in newly diagnosed treatment-naive paediatric inflammatory bowel disease patients. PLoS One *13*, e0197649. https://doi.org/10.1371/journal.pone.0197649.

35. Kulkarni, P., Devkumar, P., and Chattopadhyay, I. (2021). Could dysbiosis of inflammatory and anti-inflammatory gut bacteria have an implications in the development of type 2 diabetes? A pilot investigation. BMC Res. Notes *14*, 1–7. https://doi.org/10.1186/s13104-021-05466-2.

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

37. Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., Tameda, M., Shiraki, K., Ito, M., and Takei, Y. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. BMC Gastroenterol. *15*, 1–10. https://doi.org/10.1186/s12876-015-0330-2.

38. Duffey, B.G., Miyaoka, R., Holmes, R., Assimos, D., Hinck, B., Korman, E., Kieley, F., Ikramuddin, S., Kellogg, T., Moeding, A., and Monga, M. (2011). Oxalobacter colonization in the morbidly obese and correlation with urinary stone risk. Urology *78*, 531–534. https://doi.org/10.1016/j.urology.2011.01.011.

39. Grigor'eva, I.N. (2020). Gallstone disease, obesity and the Firmicutes/Bacteroidetes ratio as a possible biomarker of gut dysbiosis. J. Pers. Med. *11*, 13. https://doi.org/10.3390/jpm11010013.

40. Osman, M.A., Neoh, H.-m., Ab Mutalib, N.-S., Chin, S.-F., Mazlan, L., Ali, R.A.R., Zakaria, A.D., Ngiu, C.S., Ang, M.Y., and Jamal, R. (2021). Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer. Sci. Rep. *11*, 1–12. https://doi.org/10.1038/s41598-021-82465-0.

41. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Bohm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766. https://doi.org/10.15252/msb.20145645.

42. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat. Commun. *6*, 6528. https://doi.org/10.1038/ncomms7528.

43. Bullman, S., Pedamallu, C.S., Sicinska, E., Clancy, T.E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science *358*, 1443–1448. https://doi.org/10.1126/science.aal5240.

44. Wong, S.H., Kwong, T.N.Y., Chow, T.C., Luk, A.K.C., Dai, R.Z.W., Nakatsu, G., Lam, T.Y.T., Zhang, L., Wu, J.C.Y., Chan, F.K.L., et al. (2017). Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. Gut *66*, 1441–1448. https://doi.org/10.1136/gutjnl-2016-312766.

45. Zhong, L., Zhang, X., and Covasa, M. (2014). Emerging roles of lactic acid bacteria in protection against colorectal cancer. World J. Gastroenterol. *20*, 7878–7886. https://doi.org/10.3748/wjg.v20.i24.7878.

46. Warren, R.L., Freeman, D.J., Pleasance, S., Watson, P., Moore, R.A., Cochrane, K., Allen-Vercoe, E., and Holt, R.A. (2013). Co-occurrence of anaerobic bacteria in colorectal carcinomas. Microbiome *1*, 16. https://doi.org/10.1186/2049-2618-1-16.

47. Ding, X., Zhang, X., Han, J., and Ding, G. (2022). Scaling up your kernels to 31x31: revisiting large kernel design in cnns, pp. 11963–11975. https://doi.org/10.48550/arXiv.2203.06717.

48. Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Backhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103. https://doi.org/10.1038/nature12198.

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

50. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human

gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65. https://doi.org/10.1038/nature08821.

51. Costea, P.I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. Nat. Methods *11*, 359. https://doi.org/10.1038/nmeth.2897.
52. Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science *290*, 2319–2323. https://doi.org/10.1126/science.290.5500.2319.
53. Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323–2326. https://doi.org/10.1126/science.290.5500.2323.
54. Maaten, L.v.d., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html.
55. Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function, pp. 1335–1344. https://doi.org/10.1109/CVPR.2016.149.
56. Wainberg, M., Merico, D., Delong, A., and Frey, B.J. (2018). Deep learning in biomedicine. Nat. Biotechnol. *36*, 829–838. https://doi.org/10.1038/nbt.4233.
57. Jonker, R., and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing *38*, 325–340. https://doi.org/10.1007/BF02278710.
58. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions, pp. 1–9. https://doi.org/10.48550/arXiv.1409.4842.
59. Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. (2017). Large kernel matters–improve semantic segmentation by global convolutional network, pp. 4353–4361. https://doi.org/10.48550/arXiv.1703.02719.