Host-Variable-Embedding Augmented Microbiome-Based Simultaneous Detection of Multiple Diseases by Deep Learning

Shunyao Wu et al.

Published: Advanced Intelligent systems, September 21, 2023

Woobeen Jeong
November 2, 2023

Interdisciplinary Program in Bioinformatics, Seoul National University

Contents

- **Introduction**
- **Materials and Methods**
  - Host-variable-embedding
  - MMoE: Microbiome-disease association
  - Cross network: Feature interaction
  - MSI : Feature importance based on SHAP
- **Result**
- **Discussion**

Introduction

- Microbiome is a promising predictor of human disease.

- Previous studies have presented only one definitive status of each specimen from microbiome cohorts, either healthy or with a specific disease.

- To address these issues, a highly explainable deep learning (DL) method based on deep neural network (DNN) called Meta-Spec is proposed

Problem1. Single-label classification

- In classifier, simply predicting only one disease or status (as a single-label) from microorganisms has a significant limitation.

- It ignores prevalence of comorbidities in actual cohort.

- American Gut Project (AGP) ~61% patients were diagnosed with at least two disease.

- Even if the microbiome of the single disease and that of comorbidities share common biomarkers, they may have different microbial pattern.

- To provide interpretation considering the combination of diseases from microbial data, through a multi-label classifier.

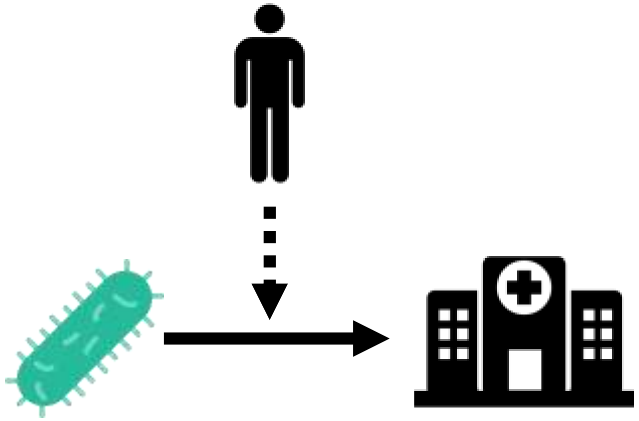
Problem2. Host phenotypes drive changes in microbiota

- Host phenotypes such as physiological traits, lifestyle, etc. have not been fully utilized in models

- This information may disrupt the microbiome-based disease prediction.

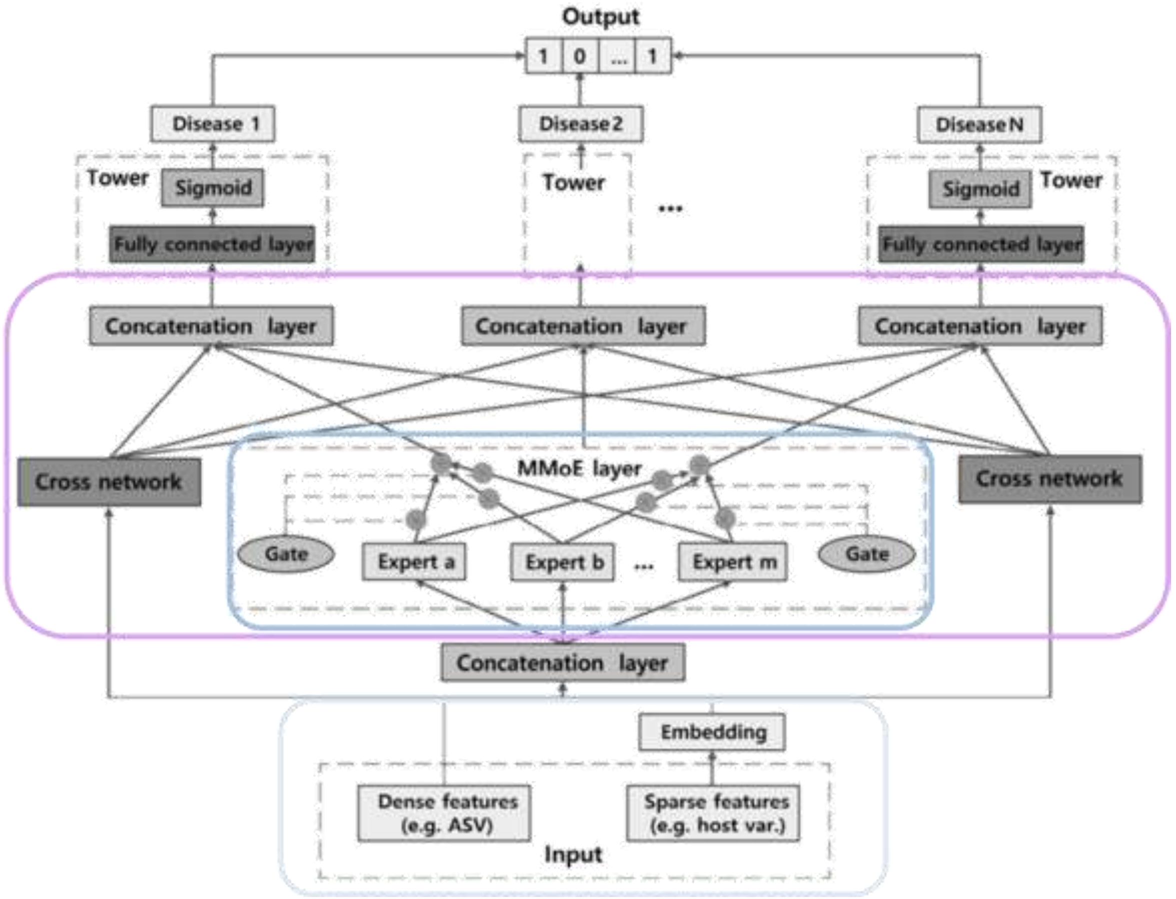
- For an example, Age is one of the major risk factors for cardiovascular disease and is also associated with Crohn’s disease.

- Therefore, even if a person have a normal microflora, she can still develop the disease due to her age. and it can interferes with prediction.



Meta-Spec

- Quantifying the relative contributions of each status include microbial and host features.
- Allow us to interpret the confounding factors.



1. concatenation

2. Multi-gate mixture of Expert model

(Association between microbiome and diseases)

3. Cross network

(Interaction between variables)

4. Update weight
(Loss function for train set)

Prediction
(Combination of comorbidities)

Multi-label classification

	ASV1	ASV2	ASV3	ASV4
Sample1	0.001	0.45	0	0
Sample2	0.2	0	0	0.003
…	…	…	…	…

- Microbial feature (genotype)

  - ASV (amplicon sequence variants)
    = High abundance reads from removing similar low ones.
  - OTU (operational taxonomy units)
    = Reads with 97% similarity.

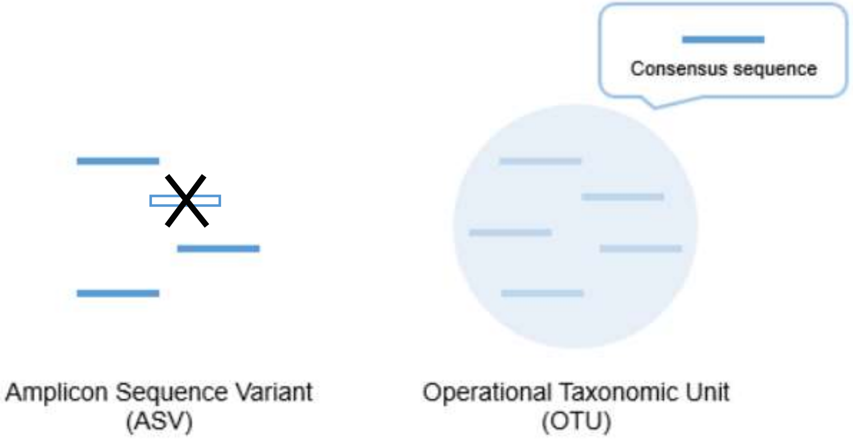
	Age	BMI	…
Sample1	3	2	…
Sample2	1	2	…
…	…	…	…

- Host variables (phenotype)

  - categorized

	Diabetes	Thyroid	…
Sample1	0	1	…
Sample2	1	1	…
…	…	…	…

- Disease Label


Consensus sequence
Amplicon Sequence Variant (ASV)
Operational Taxonomic Unit (OTU)

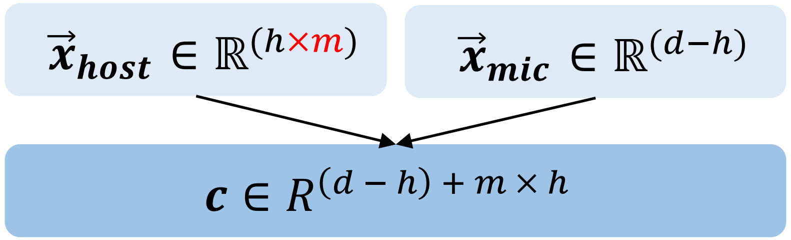
Host variable embedding

- $vector\ x = (x_1, \dots, x_h, x_{h+1}, \dots, x_d)$

- Since $(h \ll d - h)$, Imbalanced feature numbers cause dilution of host variable features.

- m – dimensional embedding vector for each host variable feature. $(m = 128)$

- $concatenated\ vector\ \boldsymbol{c} \in \boldsymbol{R}^{(d-h)+m \times h}$



$\vec{\boldsymbol{x}}_{\boldsymbol{host}} \in \mathbb{R}^{(h \times m)}$

$\vec{\boldsymbol{x}}_{\boldsymbol{mic}} \in \mathbb{R}^{(d-h)}$

$\boldsymbol{c} \in R^{(d-h)+m \times h}$

Microbial features

$x_{h+1}, \dots, x_d$

Host variable features

$x_1, \dots, x_h \quad (\times\, m)$

By one-hot encoding

	ASV1	ASV2	ASV3	ASV4	…	Age	BMI	…	a	b	c	c	…
Sample1					…	a	c	…	1	0	1	1	…
Sample2					…	b	c	…	0	1	1	1	…

$= \boldsymbol{c}$

Basics of DNN

- Deep Neural Network:
  Network consisting of 3 or more layers with
  2 or more hidden layers

$$\hat{y} = g(z_i) = \mathrm{g}\left(\sum_{\mathrm{i=1}}^{\mathrm{n}} \vec{w_i}\,\vec{x_i} + \vec{b}\right)$$

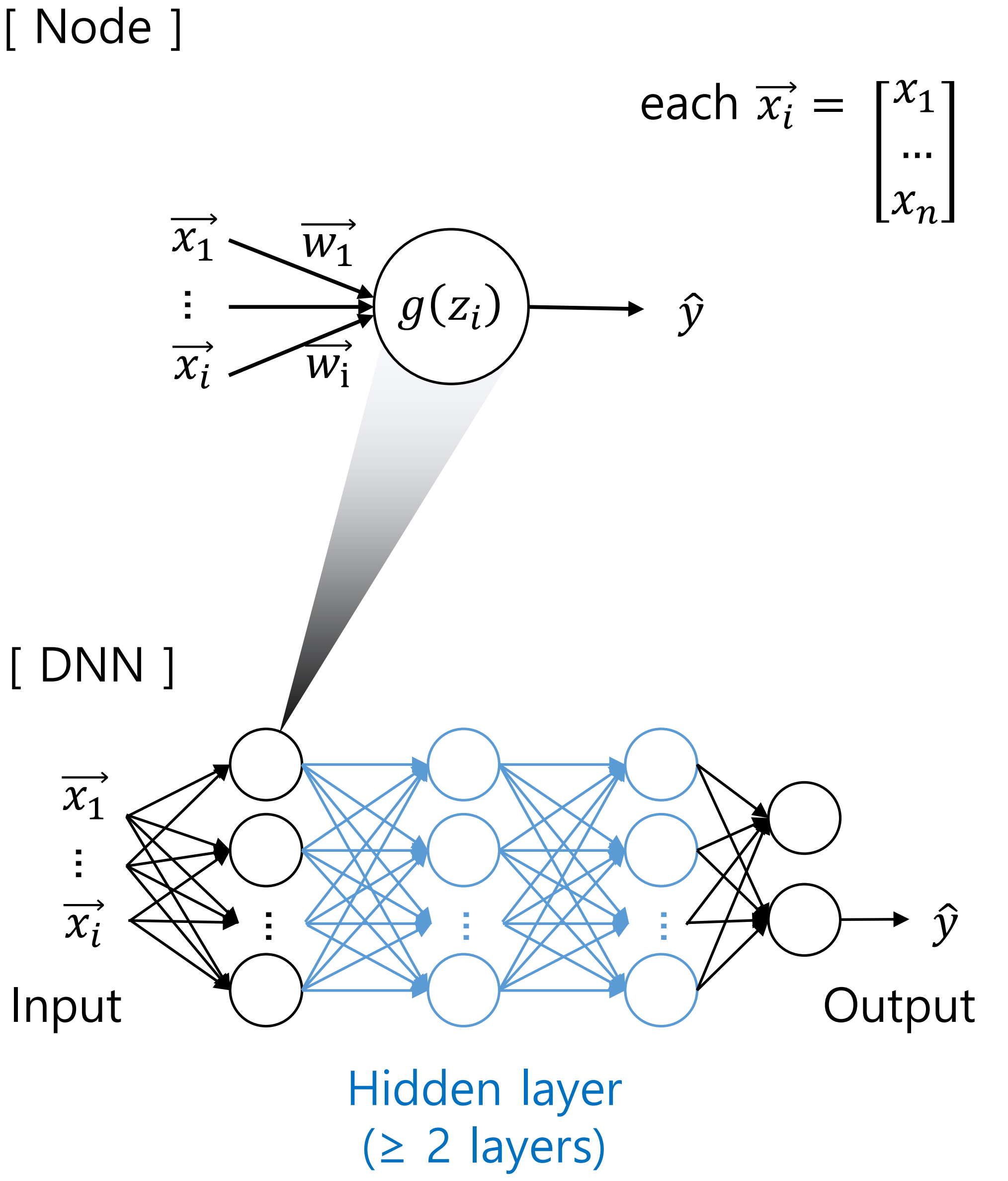
$g(z) = activation\ function,$ $\quad \vec{w_i}\ from\ loss\ function$

- Example: Multi-class classification on DNN

$$g(z)_i = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}} = softmax(z)_i$$

$$for\ i = 1, \dots, n\ and\ z = (z_1, \dots, z_n) \in \mathbb{R}^n$$

$$L(\hat{y}, y) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)\,, \qquad cross\ entrophy$$

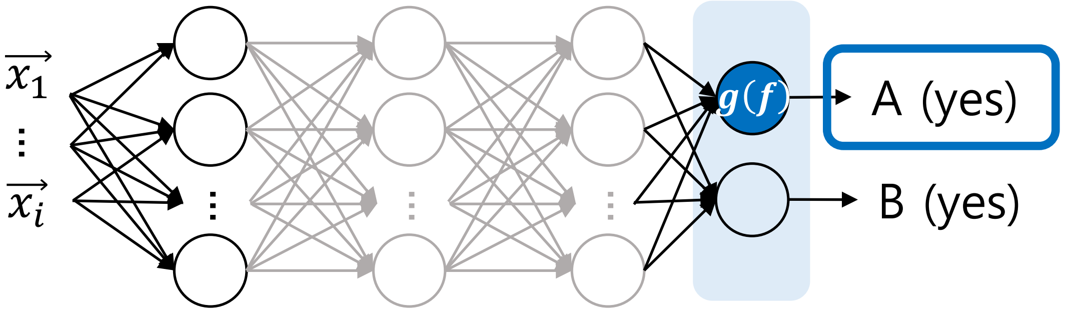


Single-label vs. Multi-label classifier

$f(x)$ = Result from former layer

- Single-label classification

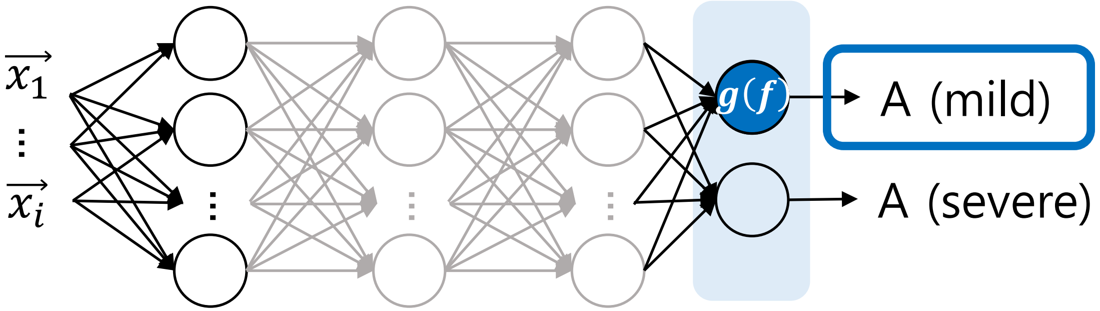
- Binary classification



$$g(f(\vec{x}_i)) = sigmoid(f(\vec{x}_i))$$

0 1

- Multi-class classification

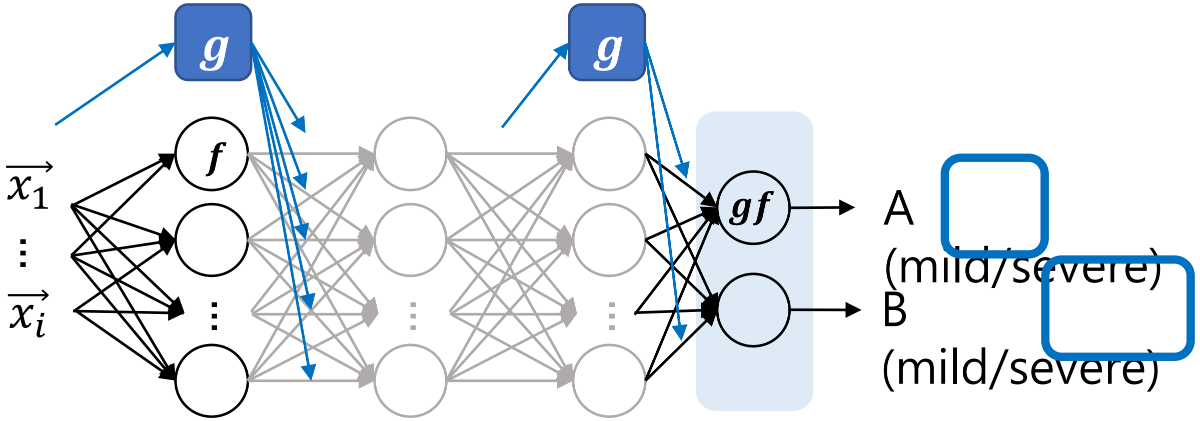


$$g(f(\vec{x}_i)) = softmax(f(\vec{x}_i))$$

0 1

- Multi-label classification

- MoE (Mixture of expert) classification



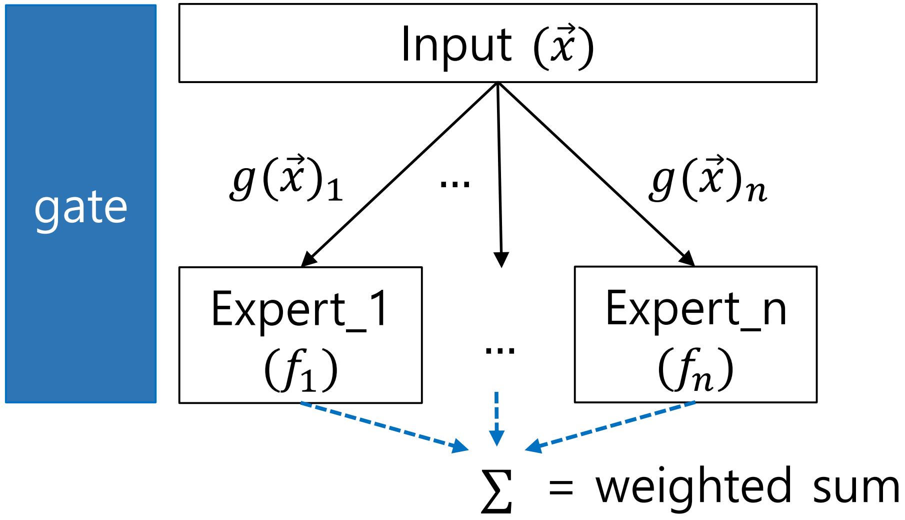
$$\sum \underbrace{g(\vec{x})_i}_{\text{gate}} \underbrace{f(\vec{x}_i)}_{\text{expert}} = \sum softmax(\vec{x})_i f(\vec{x}_i)$$

Multi-label classifier

- Original Mixture-of-Expert (MoE) model:

$$y = \sum_{i=1}^{n} g(\vec{x})_i \, f_i(\vec{x}), \qquad \sum_{i=1}^{n} g(\vec{x})_i = 1$$

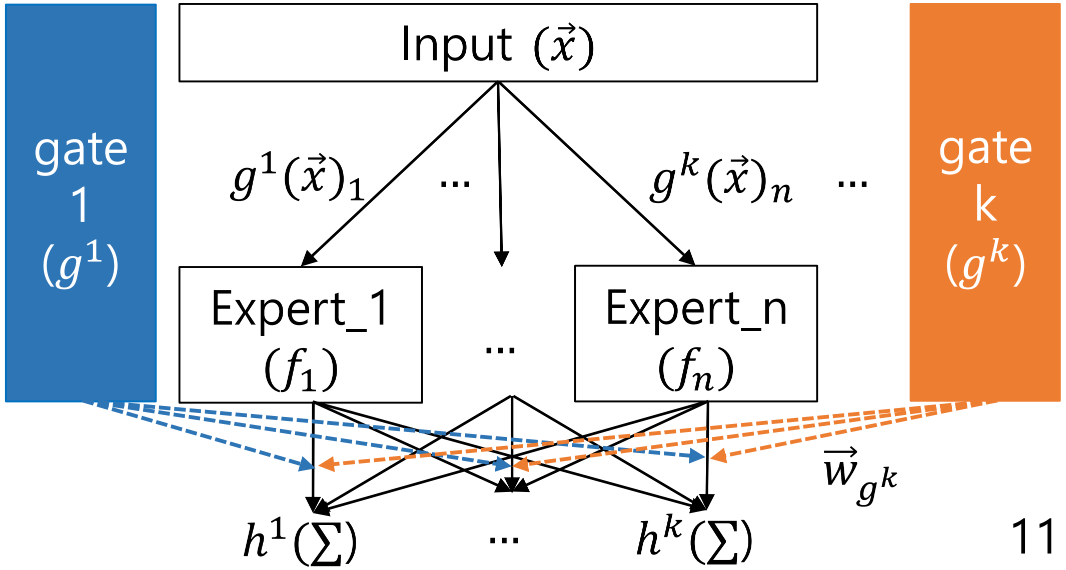
$$\underline{g(\vec{x})_i} = softmax(\vec{x})_i = probability\ for\ expert\ f_i$$



- Multi-gate Mixture-of-Expert (MMoE) model:

$$y_k = h^k\left(f^k(\vec{x})\right)$$

$$f^k(\vec{x}) = \sum_{i=1}^{n} g^k(\vec{x})_i f_i(\vec{x}), \qquad \underline{g^k(\vec{x})} = softmax(\vec{w}_{g^k}\,\vec{x})$$



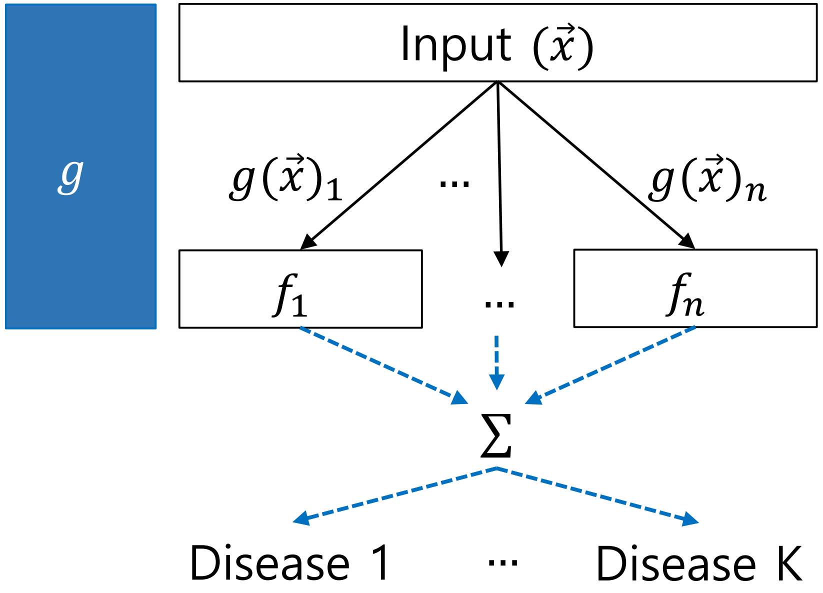
MMoE layer: Associations between microbiome and diseases

- The role of the gate in the MoE layer is to calculate the probability of going through the n th expert as a weighted sum.

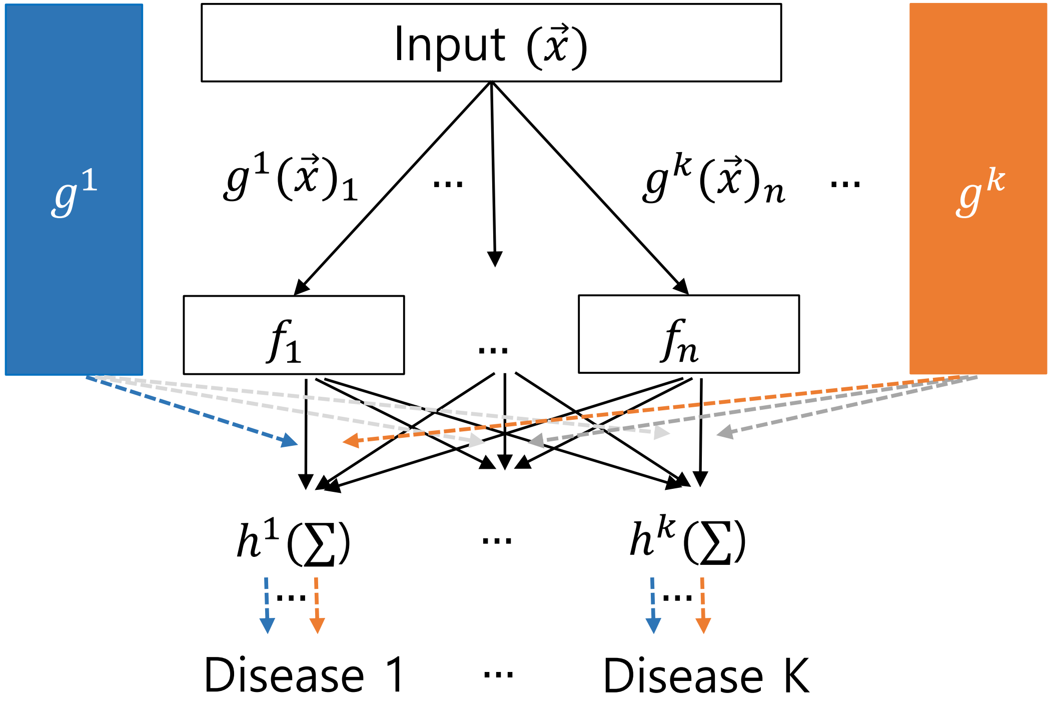
- Therefore, there is only one gate for the n th expert, which does not select any expert.

- In the MMoE layer, k gates selectively select the n th expert, each summing to 1.

[MoE]



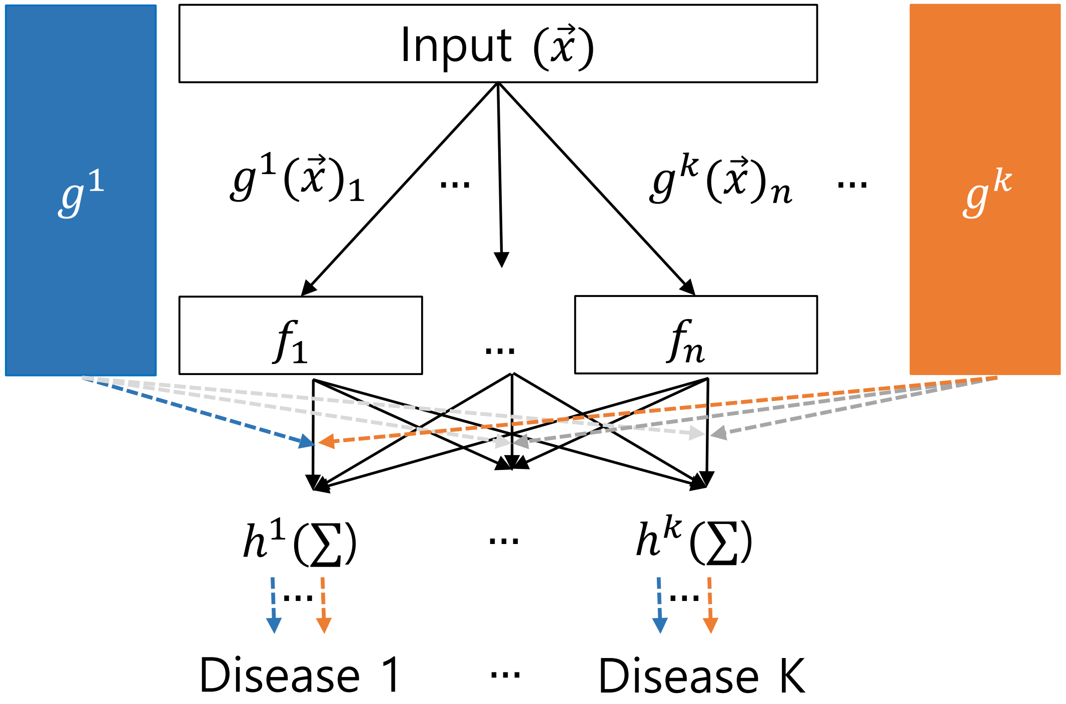
[MMoE]



MMoE layer: Associations between microbiome and diseases

- The MMoE layer can detect selective combinations of relationships between the gates of each expert.

- The output can be described as a probability combination of certain gates, just like the disease.



Input $(\vec{x})$ ↓	g^1	g^2	g^3
f_1	0.8	0	0.5
f_2	0	1.0	0.4
f_3	0.2	0	0.1

given $f_1(\vec{x})$, gate1 and gate 3 affects
= combination of Pr(diseases state)

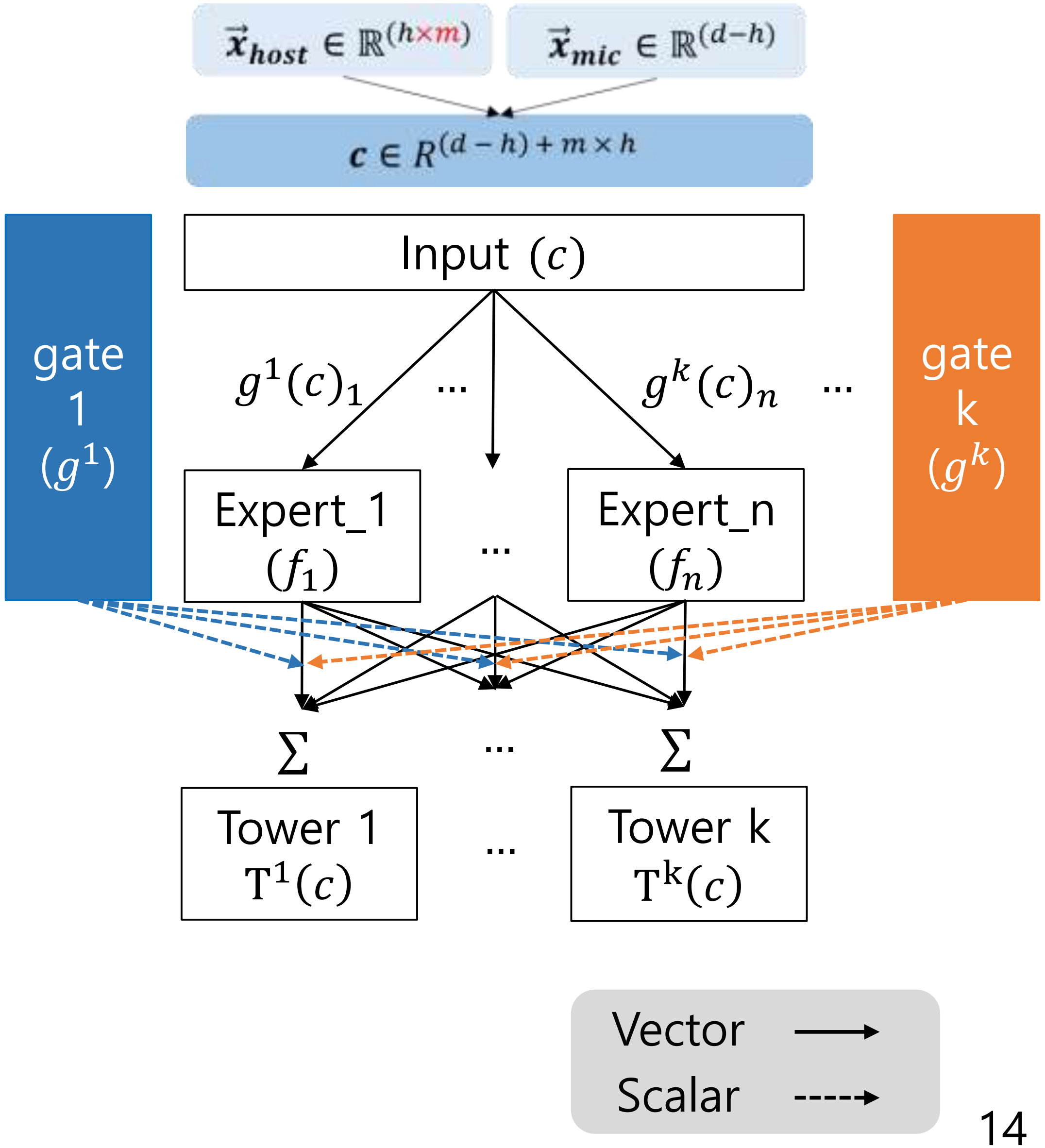
MMoE layer: Associations between microbiome and diseases

- By applying MMoE layer: $(k = \#\, of\ disease)$

$$\underline{f^k(x)} = \sum_{i=1}^{n} \underline{g^k(x)}_i \underline{f_i(x)}, \quad g^k(x) = softmax(w_{g^k}\, x)$$

- Output of $n\ th\ expert = f_n(c)$ (for each k)
- Weight of $k\ th\ gate = w_g^k \in \mathbb{R}^L$
- Tower network (for disease k):

$$\mathrm{T}^k(c) = \sum_{n=1}^{N} \underbrace{w_g^k}_{\text{weights from k th gate}}\ \underbrace{f_n(c)}_{\text{selective expert}}$$


$\vec{x}_{host} \in \mathbb{R}^{(h\times m)}$
$\vec{x}_{mic} \in \mathbb{R}^{(d-h)}$
$c \in R^{(d-h)+m\times h}$
Input (c)
gate 1 (g^1)
$g^1(c)_1$
…
$g^k(c)_n$
…
gate k (g^k)
Expert_1 (f_1)
…
Expert_n (f_n)
Σ
…
Σ
Tower 1 $\mathrm{T}^1(c)$
…
Tower k $\mathrm{T}^k(c)$
Vector
Scalar

Cross network: Microbial interaction and host variable interaction

- Microbial interaction :

$$T_{cross}(\vec{X}_{mic}) = \vec{X}_{mic} \odot (\vec{w}_{mic}\vec{X}_{mic} + \vec{b}_{mic}) + \vec{X}_{mic}$$

- Host variable interaction :

$$T_{cross}(\vec{X}_{host}) = \vec{X}_{host} \odot (\vec{w}_{host}\vec{X}_{host} + \vec{b}_{host}) + \vec{X}_{host}$$

- Simply, Cross network(X_i) = $\boldsymbol{X_i} \times (\underbrace{W_i\boldsymbol{X_i}+b)}_{\boldsymbol{X'_i}} + \boldsymbol{X_i}$

- Concatenate cross networks with MMoE

$$Output = sigmoid(\,[\overset{0}{\mathrm{T}^{\mathrm{k}}(c)} \quad \overset{1}{T_{cross}(\vec{X})}]\,)$$

$\odot$ = Hadamard product

$$A \odot B = (A)_{ij}(B)_{ij}$$

$$\begin{bmatrix} a1 & b1 \\ a2 & b1 \end{bmatrix} \odot \begin{bmatrix} a'1 & b'1 \\ a'2 & b'2 \end{bmatrix} = \begin{bmatrix} a1 \times a'1 & b1 \times b'1 \\ a2 \times a'2 & b2 \times b'2 \end{bmatrix}$$

T_{cross} = Cross network
mic = Microbial features (ASV, OTU)
$host$ = Host variables (categorized)

$\vec{X}_{mic} \in \mathbb{R}^{(d-h)}$, $\vec{b}_{mic} \in \mathbb{R}^{(d-h)}$
$\vec{w}_{mic} \in \mathbb{R}^{(d-h)\times(d-h)}$

$\vec{X}_{host} \in \mathbb{R}^{(h\times m)}$, $\vec{b}_{mic} \in \mathbb{R}^{(h\times m)}$
$\vec{w}_{host} \in \mathbb{R}^{(h\times m)\times(h\times m)}$

Meta-spec

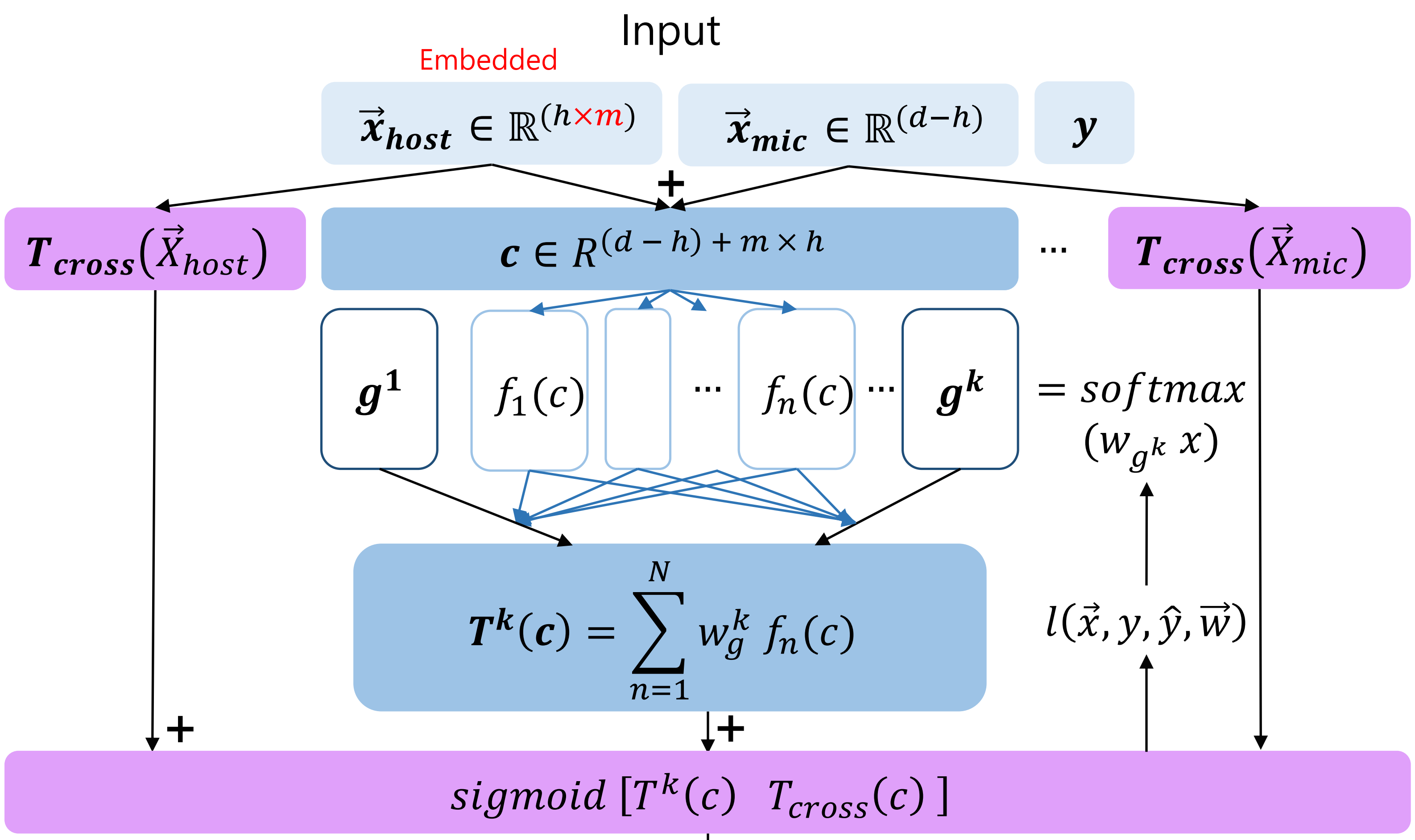
1. concatenation

2. MMoE
(Association $\vec{x}_{mic}$ & w_g^k)

3. Cross network
(Interaction between variables)

4. Update weight
(Loss function for train set)

Prediction (interpret, validate)



Input
Embedded
$\vec{x}_{host} \in \mathbb{R}^{(h \times m)}$
$\vec{x}_{mic} \in \mathbb{R}^{(d-h)}$
y
+
$T_{cross}(\vec{X}_{host})$
$c \in R^{(d-h)+m \times h}$
...
$T_{cross}(\vec{X}_{mic})$
g^1
$f_1(c)$
...
$f_n(c)$
...
g^k
$= softmax(w_{g^k}\, x)$
$T^k(c) = \sum_{n=1}^{N} w_g^k\, f_n(c)$
$l(\vec{x}, y, \hat{y}, \vec{w})$
+
+
$sigmoid\ [T^k(c)\ \ T_{cross}(c)\ ]$

$disease$	1	...			k
$\hat{y}$	0	0	1	0	1

Output

Loss function for

- Combined multi-task loss function:

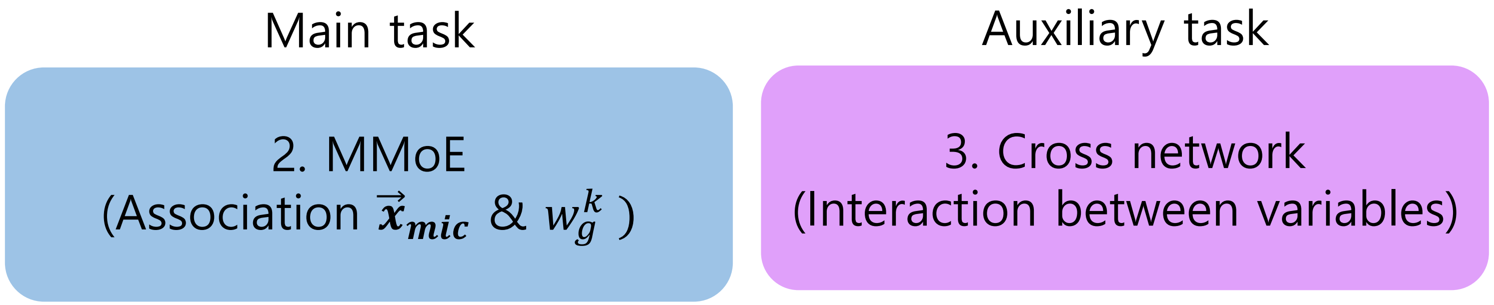
$$l(\vec{x}, y, \hat{y}, \vec{w}) = \sum_{k=1}^{K} \frac{1}{2\,c_k^2}\, l_k(\vec{x}, y_k, \hat{y}_k, \vec{w}_k) + ln(1 + c_k^2)$$

C_k = Trainable weight

$l_k(\vec{x}, y_k, \hat{y}_k, \vec{w}_k)$ = loss function for k th task

$\vec{w}_k$ = Network parameter

- To enable multi-task learning, we have to find a common representation in the earlier layers of the network.



Main task

2. MMoE
(Association $\boldsymbol{\vec{x}_{mic}}$ & w_g^k)

Auxiliary task

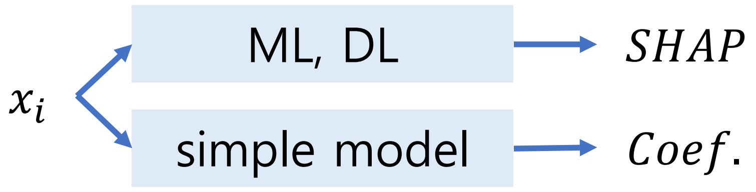
3. Cross network
(Interaction between variables)

- This way, it helps the network to be applicable to both auxiliary and main tasks.
- It can also act as a regularizer by optimizing the parameter space.

Meta-spec feature importance (MSI)

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_{12}(x_1 + x_2) + \cdots + \beta_{1\ldots i}(x_1 + \cdots + x_i) + \cdots$$



- MSI based on SHAP
  (Shapley additive explanation)

$$SHAP = \phi_i(f) = \frac{1}{\text{n}} \sum_{S\subseteq\{x1,\ldots,xni\}\backslash\{xi\}} \binom{n-1}{|S|}^{-1} (f(S\cup\{i\} - f(S))$$

$$\phi_i(\hat{y}) = \beta_i x_i - E(\beta_i X_i) = \beta_i x_i - \beta_i E(X_i)$$

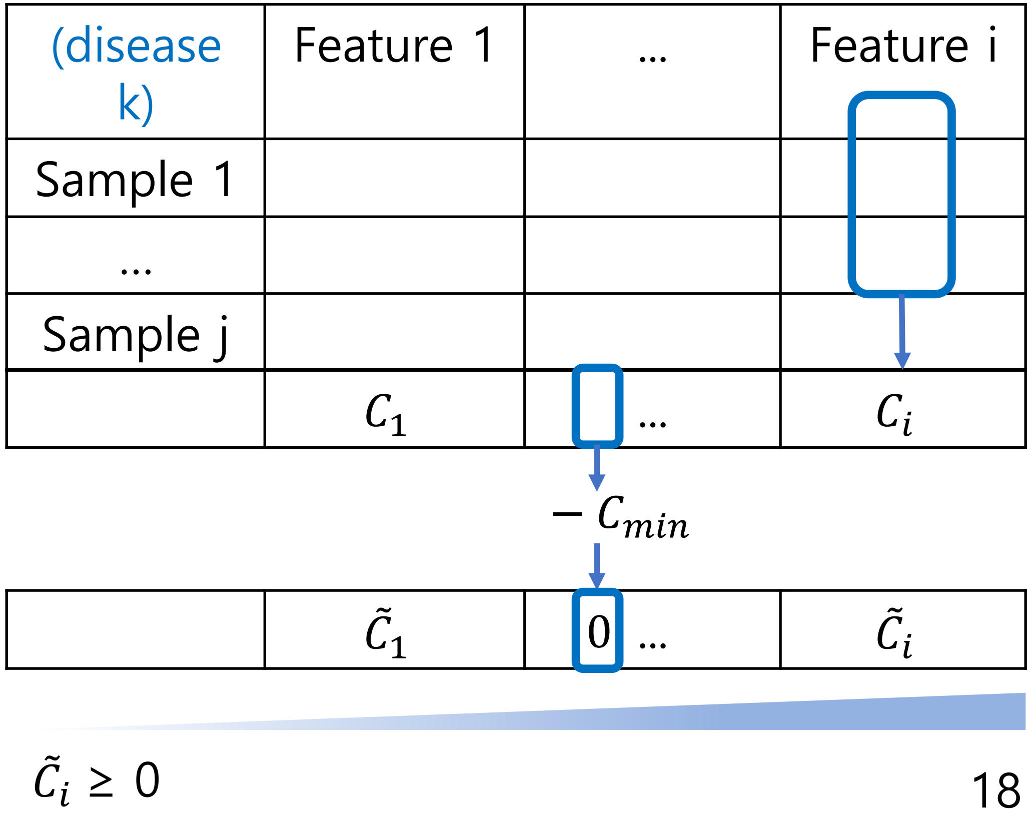
$$\sum_{i=1}^{N} \phi_i(\hat{y}) = \hat{y}(x) - E(\hat{y}(X))$$

- Proportion of a feature's contribution to the prediction

$$C_i = log(|median_{\,j}\,(SHAP_{ij})|$$

$$\widetilde{C_i} = C_i - min(C_i)$$

$$\text{MSI}_{\text{ i}} = \frac{\widetilde{C_i}}{\sum_i \widetilde{C_i}} \times 100(\%)$$



(disease k)	Feature 1	…	Feature i
Sample 1			
…			
Sample j			
	C_1	…	C_i

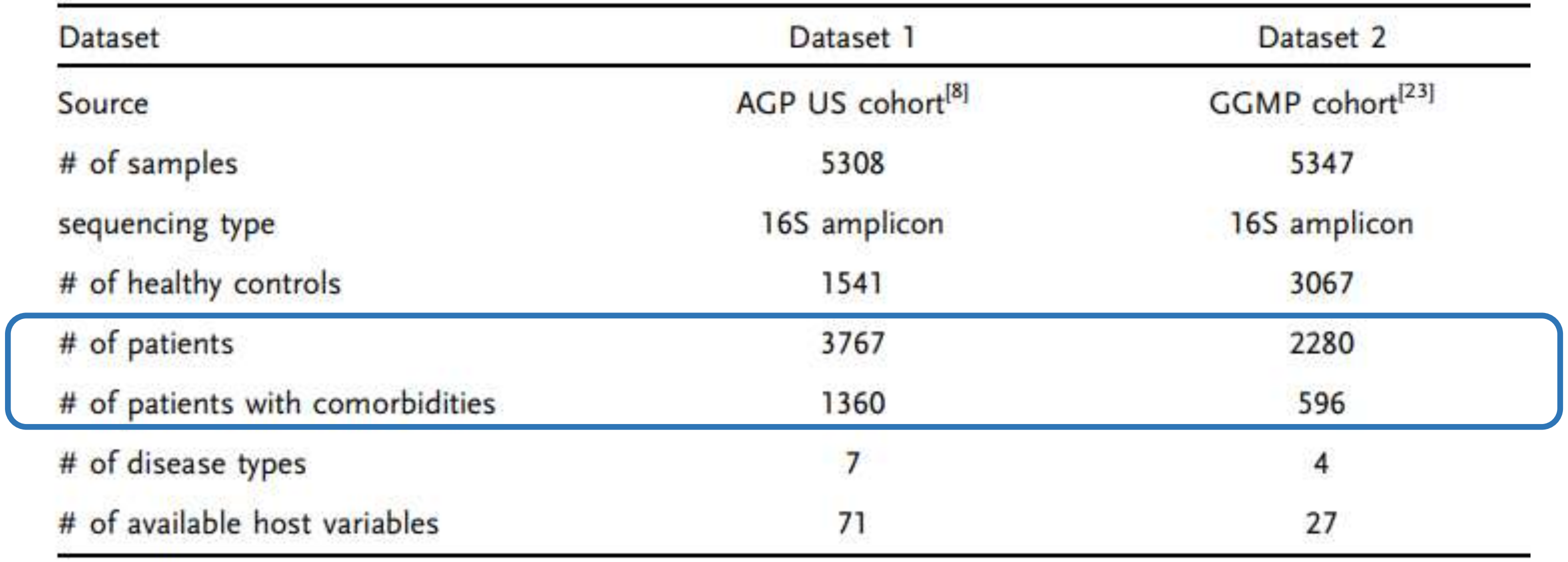
$- C_{min}$

	$\tilde{C}_1$	0 …	$\tilde{C}_i$

$\tilde{C}_i \geq 0$

Dataset

- Produced by the American Gut Project (AGP) and Guangdong Gut Microbiome project (GCMP)
- datasets included patients with comorbidities.
- The 7 target diseases of AGP are autoimmune disease, lung disease, thyroid disease, cancer, IBD, Cardiovascular Disease and Autism Spectrum Disorder
- The 4 target diseases in Dataset2 are metabolic syndrome, gastritis, type 2 diabetes, and gout.



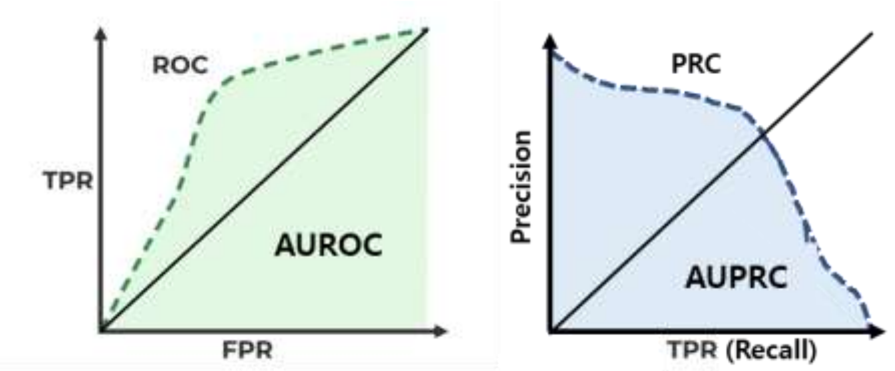
Dataset	Dataset 1	Dataset 2
Source	AGP US cohort[8]	GGMP cohort[23]
# of samples	5308	5347
sequencing type	16S amplicon	16S amplicon
# of healthy controls	1541	3067
# of patients	3767	2280
# of patients with comorbidities	1360	596
# of disease types	7	4
# of available host variables	71	27

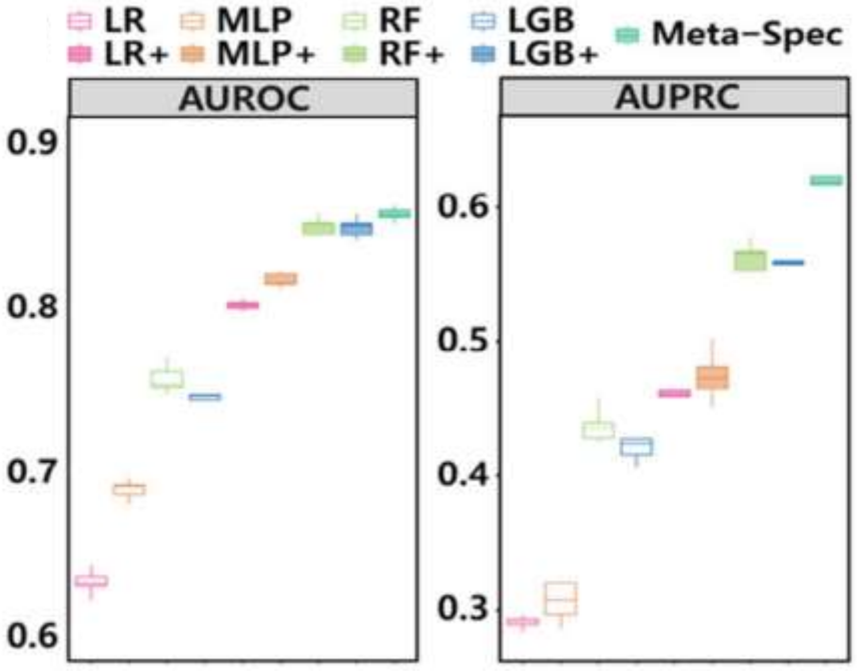
Performance comparison

- Overall performance on dataset1 (AGP US)
- Since the number of sample were highly Unbalanced, We also compare the AUPRC
- To check overall performance, 4 comparison models were set up.

- LR: logistic regression
- MLP: multi-layer perceptron
- RF: random forest
- LGB: light gradient boost

+ indicates one-hot coded metadata





TPR = TP / (TP+FN)
FPR = FP / (FP+TN)
Precison = TP / (TP+FP)

		Actual	
		P	N
Predict	P	TP	FP
	N	FN	TN

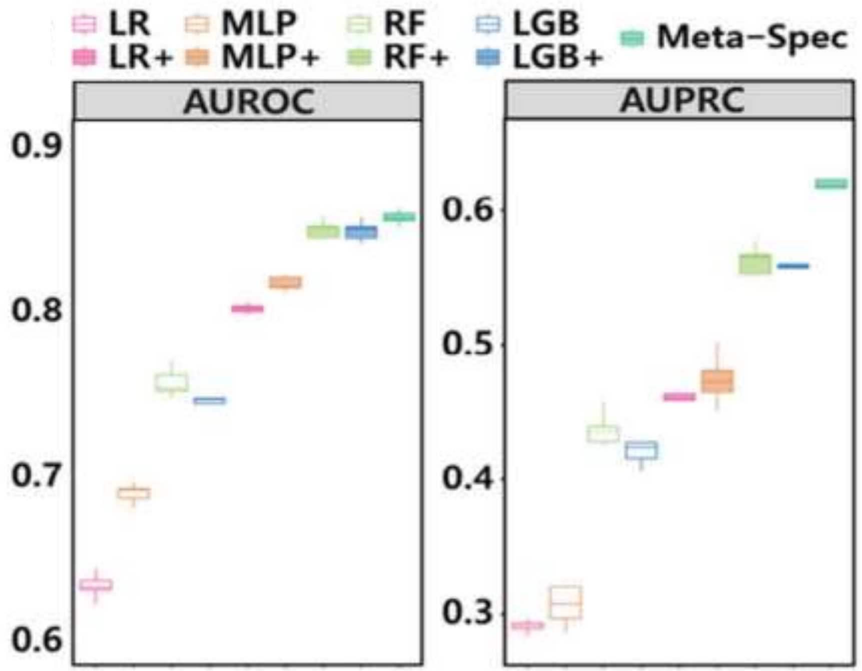
Performance comparison

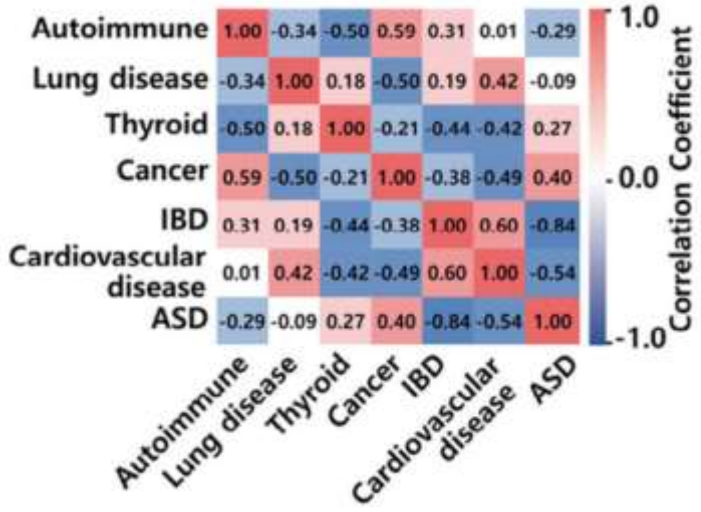
- The performance of all models improved with the use of metadata.

- The best performance of both was from Meta-Spec.

- The ML method resulted in a low AUROC, which may be due to confounding effects.

- Looking at the Pearson correlation coefficient, disease correlation information is well reflected through MMoE strategy.





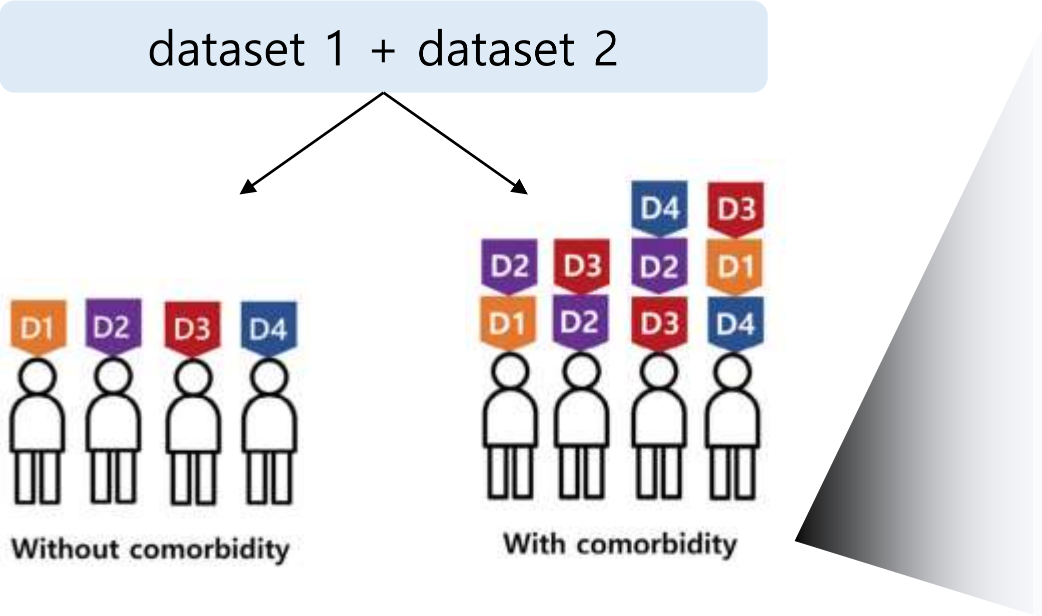
	Autoimmune	Lung disease	Thyroid	Cancer	IBD	Cardiovascular disease	ASD
Autoimmune	1.00	-0.34	-0.50	0.59	0.31	0.01	-0.29
Lung disease	-0.34	1.00	0.18	-0.50	0.19	0.42	-0.09
Thyroid	-0.50	0.18	1.00	-0.21	-0.44	-0.42	0.27
Cancer	0.59	-0.50	-0.21	1.00	-0.38	-0.49	0.40
IBD	0.31	0.19	-0.44	-0.38	1.00	0.60	-0.84
Cardiovascular disease	0.01	0.42	-0.42	-0.49	0.60	1.00	-0.54
ASD	-0.29	-0.09	0.27	0.40	-0.84	-0.54	1.00

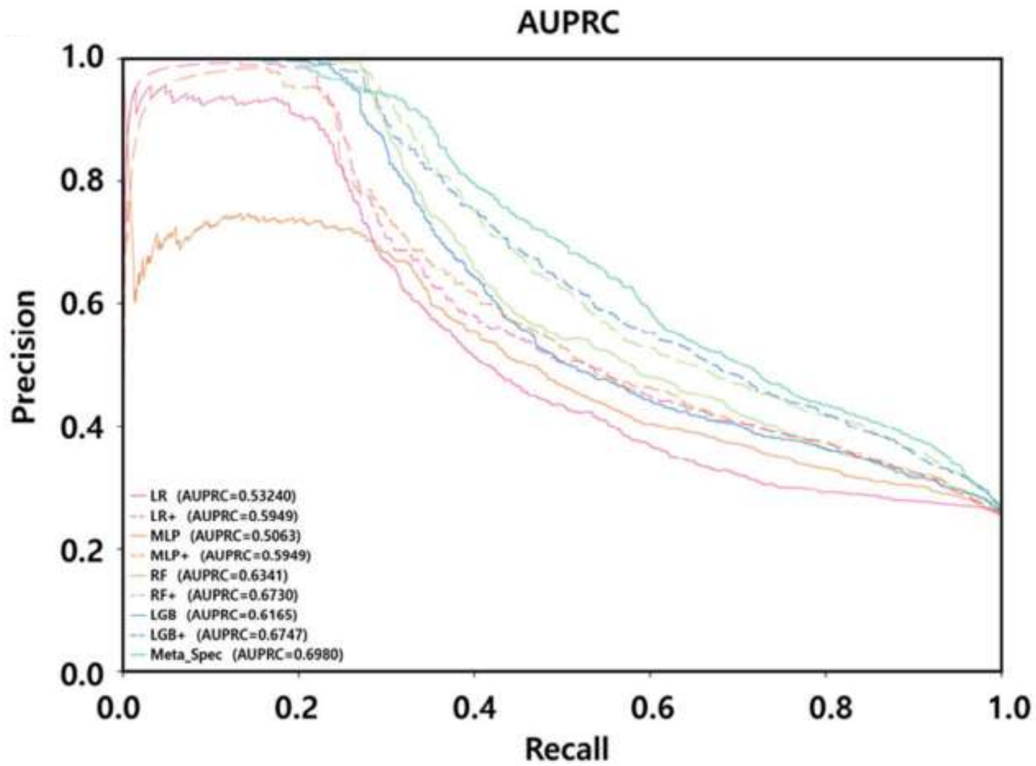
1.0
0.0
-1.0
Correlation Coefficient

Disease correlation for Comorbidity detection

- As a Fact, 1,360 among 3,767 patients on dataset 1 have 2 or more diseases.
  (596) (2,280) (dataset 2)

- To reveal that correlation between diseases helps explain these comorbidities, divided the patients into two groups, with and without comorbidities.

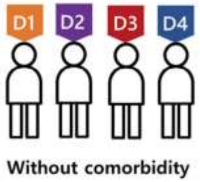


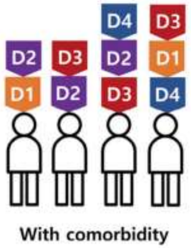


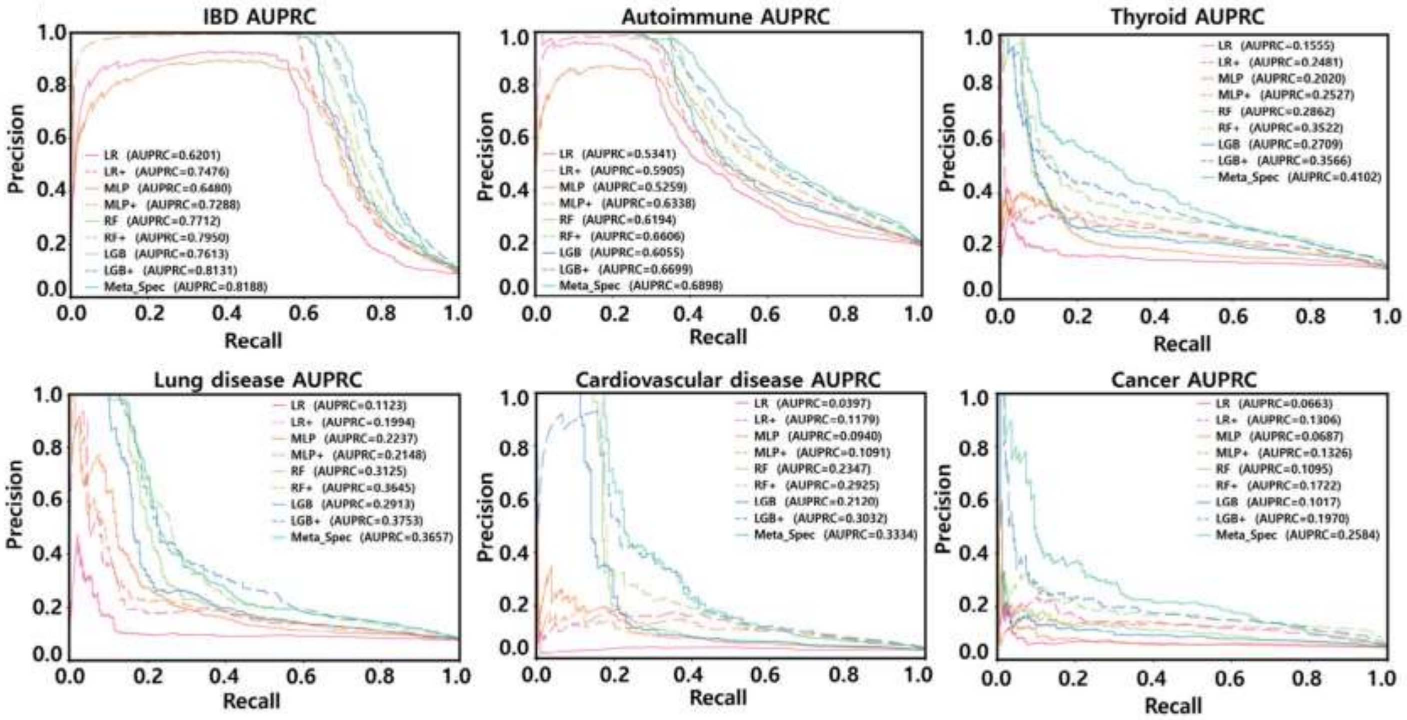
Disease correlation for Comorbidity detection

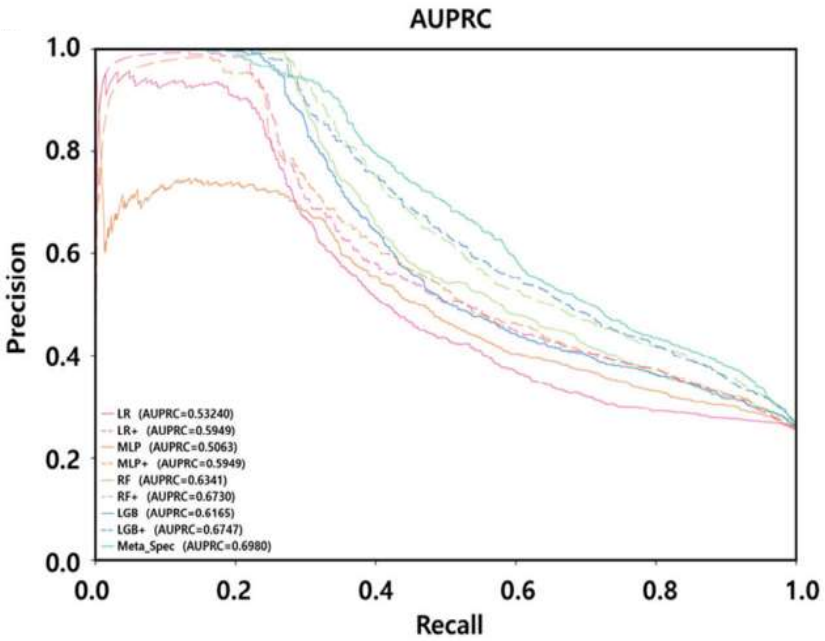
- Multi-label classification(Meta-Spec) shows higher performance than the other classifications.

- Models with single disease targets can miss comorbidity information, like thyoid or lung disease solely.









Feature selection for Multi-label disease screening

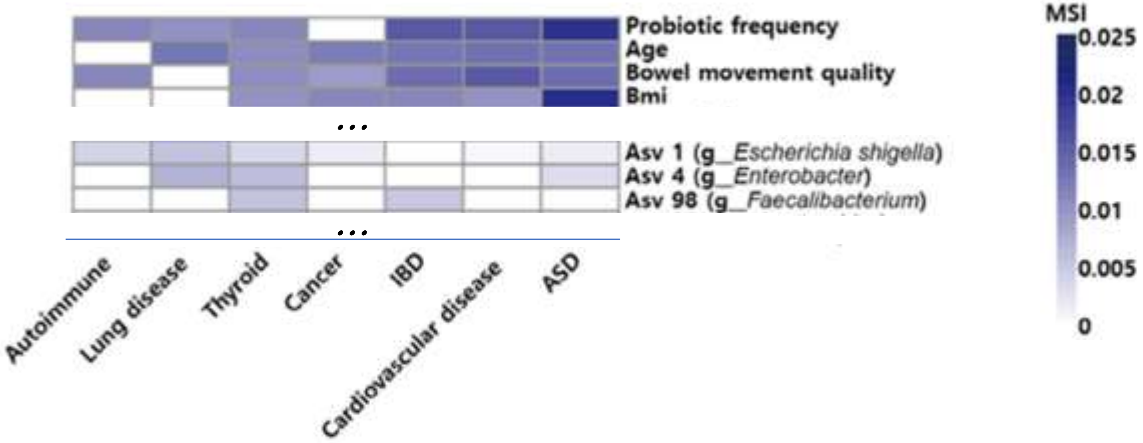
- Sorted by MSI (Meta-Spec Importance value), especially on cardiovascular disease.

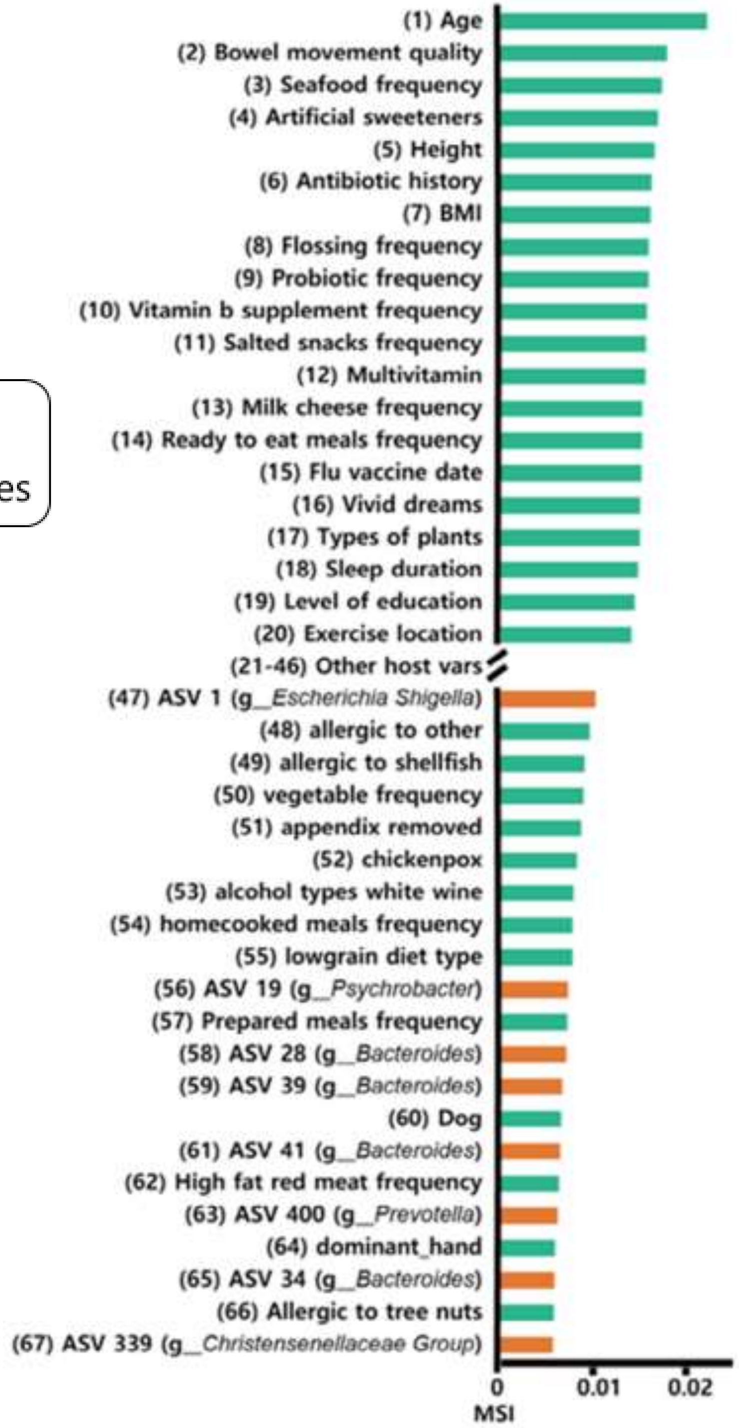
- Age was the most important feature.

- Some microbiome features (*E.Shigella, Bacteroidetes*) have been reported on cardiovascular disease

- There are common features with high rank across diseases.

🟩 Host variables
🟧 Microbial features





Discussion

- Predicting comorbidities considering the composite state of the host from microbial data using multiple datasets and cohorts is a challenging task.

- Using Meta-Spec enhances interpretability regarding the impact of microbial communities on diseases through the utilization of exclusive feature importance (MSI).

- Considering the influence of host data on disease screening is crucial, despite the well-established significance of the gut microbiota in human health.

Thank You