Inferring Microbial Communities for City Scale Metagenomics Using Neural Networks

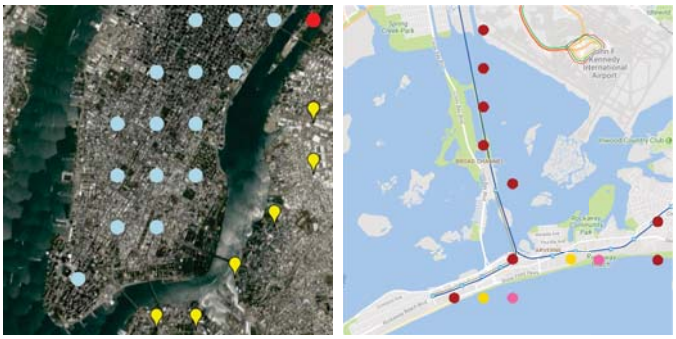
Guangyu Zhou, Jyun-Yu Jiang, Chelsea J.-T. Ju and Wei Wang*
Department of Computer Science, University of California, Los Angeles
Los Angeles, CA, USA, 90095
Email: *weiwang@cs.ucla.edu

***Abstract*—Microbes play a critical role in human health and disease, especially in cities with high population densities. Understanding the microbial ecosystem in an urban environment is essential for monitoring the transmission of infectious diseases and identifying potentially urgent threats. To achieve this goal, researchers have started to collect and analyze metagenomic samples from subway stations in major cities. However, it is too costly and time-consuming to achieve city-wide sampling with fine-grained geo-spatial resolution. In this paper, we present MetaMLAnn, a neural network based approach to infer microbial communities at unmeasured locations, based upon information from various data sources in an urban environment, including subway line information, sampling material, and microbial compositions. MetaMLAnn exploits these heterogeneous features to capture the latent dependencies between microbial compositions and the urban environment, thereby precisely inferring microbial communities at unsampled locations. Moreover, we propose a regularization framework to incorporate the species relatedness as prior knowledge. We evaluate our approach using the public metagenomics dataset collected from multiple subway stations in New York and Boston. The experimental results show that MetaMLAnn consistently outperforms five conventional classifiers across several evaluation metrics. The code, features and labels are available at https://github.com/zgy921028/MetaMLAnn**

***Keywords*—Urban metagenomics; Multi-label classification; Neural network**

I. Introduction

Metagenomics is the study of the genomic content obtained from a human body site or an environment to understand the extent and role of microbial diversity. The microorganisms presented in our environment play an important role in health and disease. While human microbiome studies have allowed us to analyze the microbial diversity within the human body [1], environmental metagenomics has also become increasingly important due to its impact on public health, especially in densely populated urban areas [2, 3, 4, 5, 6, 7, 8]. Therefore, understanding and inferring the fine-grained metagenomics composition throughout a city is vital in helping long-term disease surveillance and health management. Recent studies have made numerous efforts to establish city-scale metagenomic profiles [9, 10]. For example, Afshinnekoo *et al.* [9] collected samples from various surfaces across the entire New York subway system and created a city-wide metagenomic profile. They performed read alignment to generate taxonomic assignments and computed the relative abundances at the species level. Such a profile describes the metagenomic communities and shows how humans interact with new microbes or dangerous pathogens. In addition to the profiling in New York, Hsu *et al.* [10] provided a comprehensive metagenomic profile of microbial communities across multiple surface types in the Boston transportation system. While their works provided initial datasets for further analysis of urban microbiome diversity, it is costly and time-consuming to collect, sequence, and analyze the metagenomics data at every station. Based on these previous urban metagenomic sequencing efforts, we are interested in developing a model to automatically infer the microbial communities for unsampled locations.



(a) Geographical topology affects the microbial distribution

(b) Subway system network influences the microbial distribution

Fig. 1: In (a), there are three clusters of subway stations based on the microbial community abundance in each location, by using the Pearson correlation. The East River is a clear boundary that separates the three districts: Manhattan (blue dots), Brooklyn (yellow marks), and the Roosevelt Island (the red dot at top right). In (b), the line passing the Broad channel conserves its own microbial community cluster (red dots), whereas samples in stations on Rockaway Park Island are more diverse (different colors).

To infer the microbial communities for unsampled locations is challenging. Firstly, microbial communities could vary tremendously in a complicated urban system due to multiple factors, such as geographical topology and public transit network. Recent studies have discussed the effects of line connectedness on the similarity of microbial communities. Leung *et al.* [2] have conducted a Mantel test of Hong Kong subway line (MTR), indicating that closely connected MTR lines shared more similar microbial communities than pairs that are further apart, possibly by distance-dependent dispersal and transferring commuters. To further evaluate this assumption, we conduct the correlation analysis between the microbial abundance distributions in New York subway stations, as shown in Figure 1. Different microbes can be separated by geographical boundaries, and the same microbe can be spread along the same subway line. Secondly, the

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

surface material type, from which the samples are collected, may also affect the formation and transmission of microbial communities [10]. Lastly, within each community, the relatedness among individual microorganisms also need to be considered. As the microbial community is affected by mixed signals from various factors, a simple model predicting along the same subway line for the station is not sufficient.

To address these challenges, we formulate the problem of inferring the microbial communities of unsampled locations as a multi-label classification (MLC) task. Given a set of heterogeneous features extracted from the urban environment, we aim to predict the presence or absence of a list of microbes at a nearby location. For MLC, each location is considered as an instance where each label represents a microbe. MLC is suitable for solving our microbes inference problem since different class labels have to be predicted simultaneously [11], and their dependencies need to be exploited. These properties reflect the nature of microbial communities.

In this paper, we present MetaMLAnn (**M**etagenomic **M**ulti-**L**abel **A**rtifical **n**eural **n**etwork) to infer the microbial community for urban metagenomics. MetaMLAnn is based on the widely-used feed-forward neural network model, but instead of predicting each label individually, it incorporates an extra shared structure to capture the dependencies among different labels (microbes). To train MetaMLAnn, we integrate features constructed from different data sources. Manifold regularization is used to incorporate domain knowledge, which makes our model robust to the sparse samples with limited labeled data. Finally, to further improve our model, we present an ensemble model, MetaMLAnn+. To our best knowledge, our work is the first attempt to infer the microbial community for urban metagenomics with the neural network.

The contribution of this paper lies in the following aspects:

- This is the first in-depth study of inferring metagenomic communities for unsampled locations. We formalize the inference task as a multi-label classification problem and propose a neural network learning technique (MetaMLAnn) to solve it.
- We leverage manifold regularization to guide the training of MetaMLAnn by using domain knowledge of microbial evolutionary relationships.
- We extract useful features from various data sources. Transit features are constructed by using network embedding techniques and types of surface materials are encoded as categorical features.
- Experiments are conducted on the public available metagenomic samples collected from the subway stations in New York and Boston. The results demonstrate that MetaMLAnn outperforms five baseline methods. We further boost our performance by using the ensemble model.

II. Related Work

In the field of urban computing, statistical model, such as the regression tree [12] has been employed in atmospheric science to do a real-time prediction of air quality. More recently, there has been a trend of applying the big data approach to solve urban challenge [13]. For example, in U-Air [14], the authors aim to infer the fine-grained air quality throughout a city. Their model is a semi-supervised learning approach, based on the air quality data reported by existing monitor stations and a variety of data to infer the air quality for other locations. The spatial classifier for their model is based on an artificial neural network (ANN). Yet, this model predicts a single value (i.e. the air quality index) for each location and is inadequate to address the MLC task we formulated.

Several computational models, such as BioMiCo [15] and NMF [16] have been developed to infer microbial community structures. To model the composition of each sample given the abundance profile, BioMiCo uses a supervised Bayesian model while NMF leverages the matrix factorization. Nevertheless, these works are not directly applicable to infer the microbial community for unsampled locations in the urban environment. This is due to the fact that they are not able to incorporate spatial information.

The aforementioned models are either not suitable for understanding the complicated environmental situations or not powerful enough to model the complex relationships between microbial compositions and the urban environment. This motivates our neural network based model: MetaMLAnn.

III. Methods

In this section, we formalize the notations and the problem definition. Then we describe our proposed framework and how we can leverage other models to further improve our model.

A. Preliminaries and Problem Definition

Definition 1. Microbe Index: Microbe Index (I) is defined as an ordered list of identified microbial organisms. Each element in I is a taxonomic name at the species level.

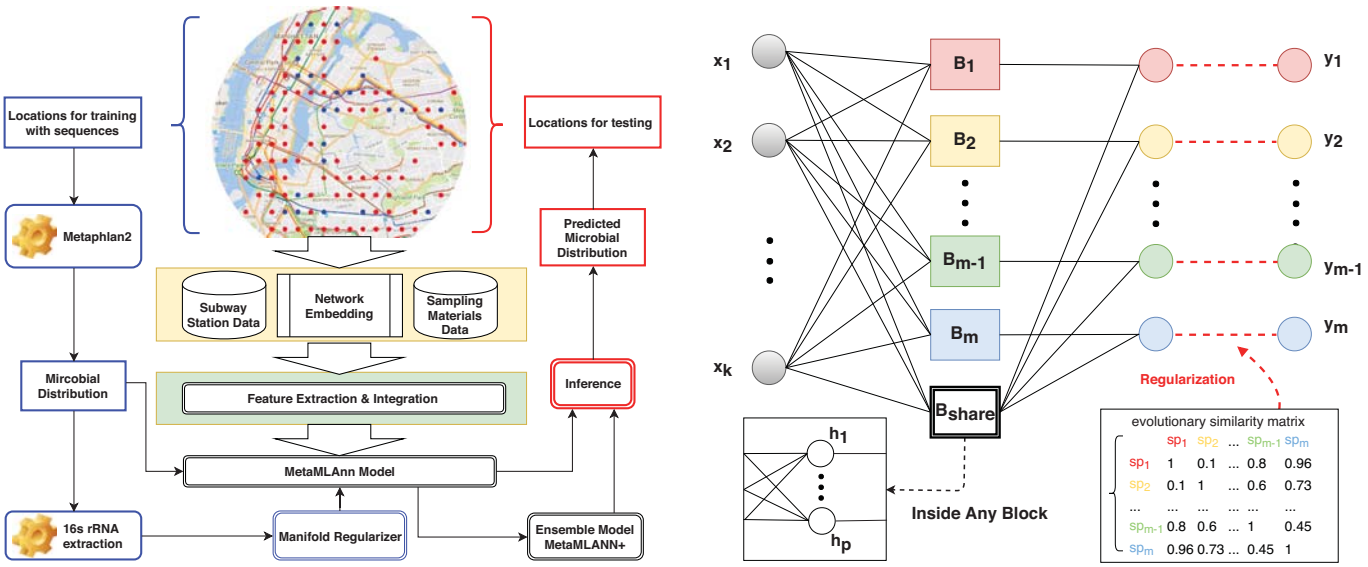
Definition 2. Microbial Distribution Matrix: All samples at different locations are represented as a matrix $Y \in R^{n \times m}$, where n is the number of sampling locations, and m is the total number of microbes in the Microbe Index. Each row Y_i represents the microbial distribution vector of location i. Each element Y_{ij} represents whether the j^{th} microbe exists (i.e. its relative abundance meets a threshold γ) in the i^{th} location.

Now, we demonstrate how we model the problem of interring microbial distribution based on the definition of Multi-Label Classification [17].

Definition 3. Inferring microbial distribution as Multi-Label Classification: A sampled location can be represented by a k-dimensional feature vector, where k is the number of predefined features. Given a sampled location and its features, we aim to infer a microbial distribution vector $\mathcal{Y} = \{y_1, y_2, ..., y_m\} = \{0, 1\}^m$ so that y_i indicates if the relative abundance of the i-th microbe in I is greater than a threshold γ.

Problem Statement. Given a set of locations $S = S_1 \cup S_2 = \{s_1, s_2, ..., s_n\}$, where S_1 and S_2 are the sets of sampled and unsampled locations, respectively. Each location $s_i \in S_1$ is sampled and associated with a microbial distribution vector Y_{s_i} while we aim to infer the microbial distribution vector Y_{s_j} of each unsampled location $s_j \in S_2$.

The overall framework is shown in Figure 2a. It consists of the MetaMLAnn model with two major flows: the learning flow (blue units), and the inference flow (red units).



(a) Framework **(b)** Architecture of MetaMLAnn

Fig. 2: The left figure (a) illustrates the proposed framework in this paper, and the right figure (b) shows the detailed architecture of the proposed MetaMLAnn.

B. Model: MetaMLAnn

Recall that we aim to infer m labels from k dimensional input features. We first introduce the neural network model [18] with a single hidden layer which contains p hidden units. Input layer $x \in R^{k\times 1}$ is connected to the hidden layer $h \in R^{p\times 1}$ with weights $W^{(1)} \in R^{p\times k}$ and biases $b^{(1)} \in R^{p\times 1}$. The hidden nodes are then connected to output nodes $o \in R^{m\times 1}$ via weights $W^{(2)} \in R^{m\times p}$ and biases $b^{(2)} \in R^{m\times 1}$. The feed-forward neural network $f_\theta : x \to o$ can be represented as follows:

$$f_\theta(x) = f_o(W^{(2)} f_h(W^{(1)}x + b^{(1)}) + b^{(2)}), \quad (1)$$

where $\theta = \{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$. f_o and f_h are the activation functions in the output layer and the hidden layer, respectively.

Our goal is to find a parameter vector θ that minimizes the cost function $J(\theta; x, y)$, which measures the discrepancy between predictions and given targets y. Here we use the cross-entropy [19] as the cost function as follows:

$$J_{CE}(\theta; x, y) = -\sum_i (y_i \log o_i) + (1 - y_i) \log(1 - o_i), \quad (2)$$

where o_i and y_i are the predicted scores and the ground truth for label i respectively. We use the sigmoid activation function in the output layer, i.e., $o = \sigma(z) = f_o(z) = 1/(1+exp(-z))$.

This naïve form of the feed-forward neural network model is used as our baseline. In MetaMLAnn, we propose a heterogeneous neural network architecture. The detailed architecture of MetaMLAnn is shown in Figure 2b. In the hidden layers, we define two different types of sub hidden layers which we call blocks (B). For the first type, each block corresponds to an individual output label, denoted as the individual blocks B_i, $1 \le i \le m$, where m is the number of labels. The second type is a shared block, B_{share}, connecting to all output labels. With the extra structure B_{share}, MetaMLAnn is able to capture the latent dependencies among different labels. Each block has the same structure, where the number of neurons is specified as a parameter p. Each block has a single layer of p hidden neurons. In the output layer, instead of using a fully connected structure, each output neuron is only connected to the corresponding individual block and the share block. Then, the output for label y_i is the linear combination of two blocks: B_i and B_{share}. Before generating the final outputs, we have one more regularization layer discussed in the next section.

Based on the above definition, the input layer remains the same as the basic neural network model. In the hidden layers, instead of p hidden units, we have $m+1$ blocks B. Within each block is a hidden layer with p hidden neurons. For each i, the input layer $x \in R^{k\times 1}$ is connected to each block $B_i \in R^{p\times 1}$ with weights $W_i^{(1)} \in R^{p\times k}$ and biases $b_i^{(1)} \in R^{p\times 1}$. Then, the blocks B_i and B_{share} are connected to output node $o_i \in R$ via weights $W_i^{(2)} \in R^{1\times p}$ and biases $b^{(2)} \in R$. The cost function is the same as Equation 2.

To efficiently optimize the above objective function, we use stochastic gradient descent (SGD) [20]. For the individual block, we randomly sample a location i and a unit from y_i to compute B_i. For B_{share}, we randomly sample a location i and a unit from all the classes among y_1 and y_m to capture the global properties shared by all microbes. The updating rules W and b can be derived by taking the derivatives of the above objective function and applying SGD. We omit the details here due to the space limit.

C. Manifold Regularization

Neural networks are known to work the best in big data scenarios with many training examples. Since we only have access to a limited number of examples with few instances of each class label, incorporating prior knowledge can potentially compensate for data sparsity. Since evolutionary relationships are expected to be associated with patterns of community composition [21], we assume that some groups of the microbes, which are closely related to each other in the taxonomy, tend to

co-occur in the same community. Here, the taxonomy refers to the identification, naming, and classification of organisms [22]. Since taxonomy is usually richly informed by the evolutionary relationships among microbes (i.e., phylogenetic), we choose to use the evolutionary similarity as the domain knowledge feeding into our regularizer. We build our regularization frameworks based on the graph Laplacian regularizer [23, 24, 25] to incorporate the microbial similarity.

Definition 4. Graph Laplacian matrix L: Given a matrix $P \in R^{m \times m}$ representing pairwise similarities, the Graph Laplacian matrix is defined as $L = D - P$, where D is a diagonal matrix with the j^{th} diagonal element $D_{j,j} = \sum_{j'=1}^{m}(P_{j,j'})$.

Given the trace operator $tr(\cdot)$, the local geometrical structure of a vector β of length I can be preserved by minimizing

$$\Omega(\beta) = \frac{1}{2} \sum_{1 \leq i,i' \leq I} P_{i,i'} ||\beta_i - \beta_{i'}||_2^2 = tr(\beta^T L \beta), \quad (3)$$

From Equation 3, β_i and $\beta_{i'}$ are enforced to be similar, resulting in the following regularized loss function, where o_i and y_i are the predicted score and true label for the sample i:

$$J_{CE_{reg}}(\theta; x, y) = -\sum_i \left[(y_i \log o_i) + (1 - y_i) \log (1 - o_i)\right] + \lambda tr\left(\beta^T L \beta\right) \quad (4)$$

To obtain the Laplacian matrix L, we first constructed the pairwise evolutionary similarity matrix (P) of different microbes, the details of which are discussed in Section IV-C. After the neural network model generates the predicted microbial distribution vector Y_i^* for the given location i, we can regularize it by feeding Y_i^* into Equation 3, where β refers to the predicted vector Y_i^* and β_i, β_j refers to the microbe i and the microbe j at this location, respectively.

D. Feature Extraction

We define a k dimensional feature vector as $F : R^k$. Each dimension represents an individual feature extracted from data sources. The feature vectors of instances then form a feature matrix and are used to train the model. In our task, we specifically construct the following feature vectors: subway station information, inter-station connections, and sampling surface materials.

Subway station features (F_s): By obtaining the MTA (for New York) and MBTA (for Boston) subway station data, we associate each location with the nearest stations within a predefined radius $r = 0.01$ miles. This radius value is an empirical parameter and can be tuned. The feature vector is then created based on the lines that pass through the current station. This station information is then used as the node in the subsequent network construction process.

As the number of DNA collected in a station has a positive correlation with the number of riders [9], we also obtained the public MTA data regarding the usage of turnstiles in the subway system at each station. We computed the average number of riders within the DNA collection date at each station and then weigh the corresponding node vector.

Interconnection features (F_c): With each location associated with a subway station, we construct the subway system network as follows: each node represents a subway station, and an edge between two nodes represents the interconnection between two stations. We use the minimum number of stops from station i to station j to compute the weight on edge(i, j). We assign 1 as the weight if there exist express trains directly connecting two stations. After obtaining the station network, the network embedding algorithm Node2Vec [26] is applied to embed each node into a low dimensional vector based on the generated graph. The embedding vector represents the interconnection features of each node.

Surface materials features (F_m): As mentioned in [10], there is a strong correlation between the surface materials and the microbial communities. Since each data sample includes information on the type of materials it was collected from, we construct another set of vectors to capture such signal. For the New York dataset, the vectors are of length 5, where each element represents one type of material: 'concrete', 'metal', 'plastic','water' or 'wood'. For the Boston dataset, there are 4 types of materials: 'glass', 'polyester', 'PVC', and 'steel'.

Finally, all features are concatenated into a feature vector for each sample. The feature vectors for all the instances form a feature matrix and be fed into the model.

E. Ensemble with Hybrid Prediction

When the amount of training data is insufficient, the performance can be compromised. To alleviate this issue, we propose to construct an ensemble of MetaMLAnn with any other model that needs fewer training samples.

For each label i, let o_i be the predicted score of MetaMLAnn. Given the score from the other model $\mathcal{M}$ as $o_i^{\mathcal{M}}$, we conduct a linear hybrid prediction for the ensemble as: $o^h = \alpha \cdot o_i + (1 - \alpha)\, o_i^{\mathcal{M}}$, where $0 \leq \alpha \leq 1$ is a parameter to decide the weights of two models. When $\alpha = 1$ the prediction is MetaMLAnn, and when $\alpha = 0$ the prediction is the model m. We denote the ensemble approach as MetaMLAnn+.

IV. Data

A. Data Description

We focus on inferring the microbial communities in densely populated urban areas. We evaluate our model using the New York and Boston datasets obtained from the MetaSUB Inter-City Challenges track of the 2017 CAMDA Contest[1]. They both contain raw metagenomic reads and sample descriptions.

There are 1,572 samples in the New York dataset. These samples are collected from open stations for all 24 subway lines of MTA. DNA samples collected from each site are sequenced using the Illumina platform, with a total of 10.4 billion paired-end reads, as reported by [9]. Aside from the raw reads, each sample is associated with meta information, including the latitude and longitude of the collection site, the collecting environment, and surface materials.

Similarly, there are 141 samples in the Boston dataset, consisting of 5 subway lines that extend from downtown into the surrounding suburbs. 16S ribosomal RNA (rRNA) gene amplification sequence data are generated from most samples, and a subset of which are subjected to shotgun metagenomic sequencing, as mentioned in [10]. Each sample is also supplemented with additional information, describing the collecting station information, surface type, and the collection date.

[1] http://camda2017.bioinf.jku.at/doku.php/contest_dataset

B. Data Preprocessing

For each sample with metagenomic sequencing reads in the New York and Boston datasets, we conduct the following preprocessing steps:

1) To be consistent with the processing procedure in [9], we use MetaPhlan2 [27] to perform microbial profiling. Each profile contains the relative abundances as a percentage from the kingdom level to the species level.
2) As mentioned in [9], 48.3% of the reads do not match to any known organism in the New York dataset. When constructing the microbial distribution vector, we remove those unknown microbes and recompute the relative abundances of the remaining known microbes.

C. Supplemental Data Sources

The subway station data from the MTA and MBTA website are used to construct the subway line features. They contain geographic locations, station names, and lines labels. We also obtain the turnstile data to quantify the busyness of all stations.

To capture the underlying microbial relationship, we construct a pairwise similarity matrix. From the NCBI [28] and the Silva [29] database, we retrieve the 16S rRNA sequence for bacteria and archaea, 5S rRNA for eukaryotes, and the whole DNA sequences for viruses. Within each kingdom, we perform pairwise sequence alignments to obtain the similarity between species from 0 to 1 based on the alignment score. We assign 0 for cross-kingdom species pairs. To compute the similarity matrix at the genus level, we take the mean of all species' similarity scores under that level and aggregate them as the new score for each genus pairs.

V. Experiments and Results

A. Baselines

As a multi-label classification problem, we adopt several widely used algorithms as baselines, including Inverse Distance Weighting (IDW) interpolation [30], k Nearest Neighbor (kNN) [31], Support Vector Machine (SVM) [32], Random Forest (RF) [33], and Neural Network [34].

B. Experimental Settings

We first remove outlier samples whose locations are outside the boundary of New York or Boston due to measuring errors. After the data processing mentioned in Section IV-B, each sample is associated with a vector of abundances. We assess the abundance at all levels and observe that many species are seriously under-represented (i.e. appearing at only one location). To alleviate this disparity issue, we focus on the abundance at the genus level. This removes the issues related to missing species-level taxonomy, under-represented microbes and close-related microbial species. Together with features extracted discussed in Section III-D, we obtain 46 features and 269 labels (232 Bacteria, 15 Eukaryotes, 8 Archaea and 14 Viruses) for the New York dataset and 43 features and 236 labels (209 Bacteria, 7 Eukaryotes, 5 Archaea and 15 Viruses) for the Boston dataset. We demonstrate the generalization of different models through k-fold cross-validation ($k = 3$).

Recall that MetaMLAnn+ is an ensemble of MetaMLAnn and the other model ($\mathcal{M}$) interpolated by α. α closer to 1 means more weight on MetaMLAnn and closer to 0 means more weight on $\mathcal{M}$. We choose IDW as $\mathcal{M}$ in the New York dataset and Random Forest (RF) as $\mathcal{M}$ in the Boston dataset. After parameter tuning, we use $\alpha = 0.7$ for both MetaMLAnn+IDW and MetaMLAnn+RF.

TABLE I: Evaluation of all the methods by cross-validation on New York dataset at the genus level.

	Evaluation Metric			
Methods	precision	recall	F1 score	ranking loss
IDW	0.5669	**0.6686**	0.6129	0.1790
kNN	0.7203	0.5109	0.5977	0.1273
SVM	**0.7510**	0.4787	0.5845	0.0725
Random Forest (RF)	0.7288	0.5026	0.5941	0.1365
Neural Network	0.7419	0.5110	0.6050	0.0718
MetaMLAnn	0.7456	0.5325	**0.6212**	**0.0682**
MetaMLAnn+IDW	0.6578	0.6170	**0.6363**	**0.0688**

TABLE II: Evaluation of all the methods by cross-validation on Boston dataset at the genus level.

	Evaluation Metric			
Methods	precision	recall	F1 score	ranking loss
IDW	0.7288	0.5026	0.5941	0.1365
kNN	0.7401	0.6447	0.6812	0.1924
SVM	0.7583	0.5366	0.6282	0.1473
Random Forest (RF)	0.7397	**0.6746**	0.6991	0.2187
Neural Network	0.7228	0.5594	0.6214	0.1297
MetaMLAnn	**0.7674**	0.6706	**0.7095**	**0.1270**
MetaMLAnn+RF	**0.7744**	**0.6862**	**0.7229**	**0.1283**

C. Performance of MetaMLAnn and MetaMLAnn+

To assess the performance of our classifier, we report precision, recall, F1 score and ranking loss [35] as our evaluation metrics. Table I and II show the overall performance. As discussed in the experimental settings, we focus on the genus level inference.

In the New York dataset, our model performs the best in terms of F1 score and ranking loss, though the precision and recall rank the second among other baselines. MetaMLAnn presents the best balance of both precision and recall. In addition to MetaMLAnn, we also report the result of the ensemble model MetaMLAnn+IDW. Table I shows that the F1 score can be further boosted by more than 1%, which is better than either MetaMLAnn or IDW.

As for the Boston dataset, in addition to the F1 score and ranking loss, our model also outperforms all the baseline models in precision. Even though Random Forest achieves a slightly higher recall, it suffers from many false positives. However, as shown in Table II, after we leverage the Random Forest model as part of our ensemble model, MetaMLAnn+RF achieves the best performance in all metrics.

In both datasets, MetaMLAnn is conservative, i.e. predicting more 0s as labels, due to the sparsity of the datasets. IDW and Random Forest tend to predict more microbes with a high recall value in the New York and Boston dataset, respectively. By leveraging this property as orthogonal information, MetaMLAnn+ compensates the conservativeness of MetaMLAnn.

VI. Conclusion and Future Work

City-scale metagenomics profiling of microbial diversity is vital to a city for long-term disease surveillance and health management. While recent work has made great efforts to collect metagenomic samples in densely populated cities, it is still challenging to obtain the metagenomic profiles at fine-grained geo-spatial resolutions. In this paper, we identify the

problem of inferring microbial community for urban metagenomics and model it as a multi-label classification problem. We propose MetaMLAnn, a neural network based approach to infer microbial communities for unsampled locations based upon information from various data sources in the urban environment. The model captures the dependencies between microbes and the urban environment by a shared hidden layer. The ensemble technique further improves the performance of the model by leveraging signals from other models. The extensive experiments demonstrate the effectiveness of our approach. In the future, with the increasing amount of cities being sampled, we plan to extend our model to solve the inter-city metagenomic inference problem.

ACKNOWLEDGMENT

The authors thank Mr. Patrick Tan and Dr. Xiuli Ma for proof-reading. We also thank the reviewers for their helpful comments. The work was partially supported by NSF DBI-1565137 and NIH R01GM115833.

REFERENCES

[1] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[2] M. H. Leung, D. Wilkins, E. K. Li, F. K. Kong, and P. K. Lee, "Indoor-air microbiome in an urban subway network: diversity and dynamics," *Applied and environmental microbiology*, vol. 80, no. 21, pp. 6760–6770, 2014.

[3] C. E. Robertson, L. K. Baumgartner, J. K. Harris, K. L. Peterson, M. J. Stevens, D. N. Frank, and N. R. Pace, "Culture-independent analysis of aerosol microbiology in a metropolitan subway system," *Applied and environmental microbiology*, vol. 79, no. 11, pp. 3485–3493, 2013.

[4] C. Cao, W. Jiang, B. Wang, J. Fang, J. Lang, G. Tian, J. Jiang, and T. F. Zhu, "Inhalable microorganisms in beijing's pm2. 5 and pm10 pollutants during a severe smog event," *Environmental science & technology*, vol. 48, no. 3, p. 1499, 2014.

[5] S. Yooseph, C. Andrews-Pfannkoch, A. Tenney, J. McQuaid, S. Williamson, M. Thiagarajan, D. Brami, L. Zeigler-Allen, J. Hoffman, J. B. Goll *et al.*, "A metagenomic framework for the study of airborne microbial communities," *PLoS One*, vol. 8, no. 12, p. e81862, 2013.

[6] C. Firth, M. Bhat, M. A. Firth, S. H. Williams, M. J. Frye, P. Simmonds, J. M. Conte, J. Ng, J. Garcia, N. P. Bhuva *et al.*, "Detection of zoonotic pathogens and characterization of novel viruses carried by commensal rattus norvegicus in new york city," *MBio*, vol. 5, no. 5, pp. e01 933–14, 2014.

[7] T. Conceição, F. Diamantino, C. Coelho, H. de Lencastre, and M. Aires-de Sousa, "Contamination of public buses with mrsa in lisbon, portugal: a possible transmission route of major mrsa clones within the community," *PLoS One*, vol. 8, no. 11, p. e77812, 2013.

[8] A. T. Reese, A. Savage, E. Youngsteadt, K. L. McGuire, A. Koling, O. Watkins, S. D. Frank, and R. R. Dunn, "Urban stress is associated with variation in microbial species composition but not richness in manhattan," *The ISME journal*, vol. 10, no. 3, pp. 751–760, 2016.

[9] E. Afshinnekoo, C. Meydan, S. Chowdhury, D. Jaroudi, C. Boyer, N. Bernstein, J. M. Maritz, D. Reeves, J. Gandara, S. Chhangawala *et al.*, "Geospatial resolution of human and bacterial diversity with city-scale metagenomics," *Cell systems*, vol. 1, no. 1, pp. 72–87, 2015.

[10] T. Hsu, R. Joice, J. Vallarino, G. Abu-Ali, E. M. Hartmann, A. Shafquat, C. DuLong, C. Baranowski, D. Gevers, J. L. Green, X. C. Morgan, J. D. Spengler, and C. Huttenhower, "Urban transit system microbial communities differ by surface type and interaction with humans and the environment," *mSystems*, vol. 1, no. 3, 2016. [Online]. Available: http://msystems.asm.org/content/1/3/e00018-16

[11] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1-2, pp. 5–45, 2012.

[12] R. J. Lewis, "An introduction to classification and regression tree (cart) analysis," presented at annual meeting of the society for academic emergency medicine," in *Annual Meeting of the Society of Academic Emergency Medicine in*. Citeseer, 2000.

[13] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[14] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1436–1444.

[15] M. Shafiei, K. A. Dunn, E. Boon, S. M. MacDonald, D. A. Walsh, H. Gu, and J. P. Bielawski, "Biomico: a supervised bayesian model for inference of microbial community structure," *Microbiome*, vol. 3, no. 1, p. 8, 2015.

[16] Y. Cai, H. Gu, and T. Kenney, "Learning microbial community structures with supervised and unsupervised non-negative matrix factorization," *Microbiome*, vol. 5, no. 1, p. 110, 2017.

[17] E. Gibaja and S. Ventura, "Multi-label learning: a review of the state of the art and ongoing research," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444, 2014.

[18] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[19] L.-Y. Deng, "The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation, and machine learning," 2006.

[20] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[21] I. J. Lovette and W. M. Hochachka, "Simultaneous effects of phylogenetic niche conservatism and competition on avian community structure," *Ecology*, vol. 87, no. sp7, 2006.

[22] J. S. Wilkins, "What is systematics and what is taxonomy?" *Evolving Thoughts*, 2011.

[23] K. Q. Weinberger, F. Sha, Q. Zhu, and L. K. Saul, "Graph laplacian regularization for large-scale semidefinite programming," in *Advances in neural information processing systems*, 2007, pp. 1489–1496.

[24] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[25] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 507–516.

[26] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[27] D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata, "Metaphlan2 for enhanced metagenomic taxonomic profiling," *Nature methods*, vol. 12, no. 10, pp. 902–903, 2015.

[28] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2015.

[29] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[30] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & geosciences*, vol. 34, no. 9, pp. 1044–1055, 2008.

[31] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[35] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.