Machine Learning Analysis of Microbiome Data to Predict Cocaine Addiction

Anthony B Song*

Edwin O Smith High School, Mansfield-Storrs, CT 06269, USA

Corresponding author email: anthony.b.song@gmail.com

Abstract

Cocaine use disorder is a major public health problem in the US and worldwide. The microbiome may influence behavioral response to cocaine via gut-brain interactions. Microbial and behavioral variables are measured on mice to test whether differences in the microbiome account for the inter-subject variation in cocaine use. However, complex data requires sophisticated data science methods to effectively analyze. This study aimed to take a novel machine learning approach to first reduce data dimension, then cluster mice according to their cocaine use patterns, and finally classify mice into the resultant clusters based on microbial features. Both linear and nonlinear dimensionality reduction methods were employed and compared to generate new coordinates. Based on the best coordinates, K-means identified three clusters of mice: high risk mice that became addicted quicker and administered more cocaine; low cocaine use mice that used less doses in response to different dosages; low risk mice that required much more infusion sessions to eventually acquire. An artificial neural network used to differentiate these behavioral groups based on abundance of 68 bacterial genera produced micro- and macro-averaged AUC of 0.73 and 0.67 respectively. Model interpretation of this network helped identify risk and protective factors. An increased abundance of *Escherichia Shigella* and *Enterococcus* was associated with high-risk mice whereas an increased abundance of *Akkemensia* and *Erysipelotrichaceae Incertae Sedis* was shown to be protective factors.

Keywords

Microbiology; Applied Microbiology, Gut Bacteria, Machine Learning, Cocaine Addiction

Introduction

The microbiome plays a key role in human health and is a predictor of various diseases[1]. Substance use and addiction are major public health problems in the United States and worldwide[2]. Particularly, cocaine abuse is associated with substantial morbidity and mortality, and cocaine overdose deaths are increasing, and in certain populations, outnumber deaths due to heroin and opiate overdose[3]. To date, studies of the neural basis underlying substance use disorders have focused on the neurobiology of reward processing, cognitive control, and emotion regulation in the central nervous system (the brain and spinal cord)[4, 5]. An emerging research direction starts to investigate the microbiome's participation in drug reward via the gut brain axis[6, 7]. The gut-brain axis (GBA) is the bidirectional communication between the central nervous system and the virus and bacteria within a host's gut together with the enteric nervous

system[8]. There has been a critical gap in data science research to integrate gut microbiome data with substance use behavior to understand how and what microbiota are associated with the development and severity of substance addiction[9]. Understanding the relationship between gut microbiota and cocaine use behavior will help explain the biological mechanism of cocaine addiction.

Although the gut and brain are separate organs, they communicate with each other via trillions of intestinal bacteria that collectively make up the gut microbiome. Findings from both humans and animals support the theory that the gut microbes play a critical role in regulating brain function and mood[10]. Cocaine addiction reflects dysregulation of motivational, reward, and stress circuits[11]. The gut microbiome influences the same neural circuitry, suggesting gut-brain interactions in cocaine use disorders. For instance, in human studies significantly different gut bacterial microbiomes between cocaine users and non-cocaine users were found[12, 13]. Alterations of the gut microbiome also affect behavioral responses to cocaine in mice[14]. Changes in the gut microbiome and its metabolites may not only be a consequence of cocaine use but may in fact participate in mediating behavioral responses to cocaine[6]. Cocaine use disorder is characterized by highly heterogeneous behaviors and symptoms[15]. Microbiome opens a new pathway to examine the individual variation in cocaine use. Mouse data on fecal microbiome and cocaine administration behavior have been acquired in the Jackson Laboratory for Genomic Medicine and analyzed in this study to examine whether bacteria present in feces are associated with different cocaine use behavior and the risk of developing an addiction. Mice are an attractive animal model for this study because many variables can be controlled in their experiments such as environment, diet, and lifestyle that are confounders in humans.

There have been limited analytic tools available to jointly analyze the multivariate data at both the microbiome and behavior levels. Correlation and association analysis is widely used to examine concurrent patterns between gut dysbiosis and substance use[16]. However, it can be challenging to co-analyze the different data types, such as *binary* behavioral indicators versus real-*valued* microbial parameters that measure diversity and stability. Directly merging the different types of data may not maintain data integrity. Alpha and beta diversity[17] measure respectively the diversity of bacteria within one sample (a mouse) and the dissimilarity between two samples (two mice), but these diversity parameters reflect the overall inter-host difference. Thus, analyzing diversity parameters cannot capture the effects of *individual* bacteria on a host's behavior such as cocaine use.

Further, both microbial and behavioral data can be complex and high dimensional, requiring dimension reduction to reveal essential structures of data and relationships. Deep learning has become a choice of method to model complex data[18, 19]. In the microbiome study field, it has been used to mainly learn representations of microbial features, such as deepMicro[20], a set of autoencoders used to represent microbiome data by lower-dimensional vectors. Deep learning models can also be trained to predict disease as a function of microbiome variables[21], but often the first question is to determine which disease indicators and symptoms are the prediction target when the disease is multi-faceted. It can be important to design a deep learning-based framework that reduces microbiome data dimension and simultaneously predicts heterogeneous animal or human behavior.

Data have been acquired on gut microbiomes and cocaine use behaviors of diversity outbred (DO) mice. DO mice can best mimic the human populations to generate a random assortment of genetics. The dataset contains counts of the copies of fecal bacteria in specific genera. Each

mouse is characterized by these microbial features and behavioral parameters. Nine cocaine self-administration (using catheters) behavioral parameters are assessed and recorded. In this work, a pipeline of analytic steps is designed to first reduce the dimension of behavioral data which helps reveal clusters of mice with more homogeneous behavioral patterns, and then microbial features are used in a classifier to predict the different clusters. Linear dimension reduction and nonlinear dimension reduction methods are used to reveal the essential dimension of the cocaine use behavior followed by the k-means clustering method to identify clusters of mice that correspond to a cocaine use index. Then an artificial neural network (ANN) is designed to take microbial features of a mouse as inputs and calculate through layers of connected ANN nodes to predict the identified cluster assignment of the mouse.

This paper is organized as follows: Section 2 describes the methods used in the proposed analysis, and the results and discussions are presented in Section 3. Section 4 concludes the paper with observations and future directions.

Methods

Figure 1 illustrates the proposed analytic pipeline. A dimension reduction method first projects the behavioral data into new coordinates. Using the projected data, the k-means clustering method identifies subgroups of mice that exhibit different patterns in cocaine administration. A cluster validity index is employed to evaluate the consistency within the subgroups, which helps choose the best number of clusters. An ANN classifier is then constructed to assign mice to specific subgroups based on their microbial features. The study sample is first split into a training set and a test set. The ANN is trained on the training set and tested for classification performance on the test set. Although deep neural networks are generally considered black boxes, researchers have explored ways to interpret the networks. A visualization method, class activation map (CAM)[22], is employed to identify the bacterial genera that play critical roles in the classification of behavioral groups. This analytic pipeline allows us to better classify cocaine addiction and identify the gut bacteria that differentiate cocaine responses.

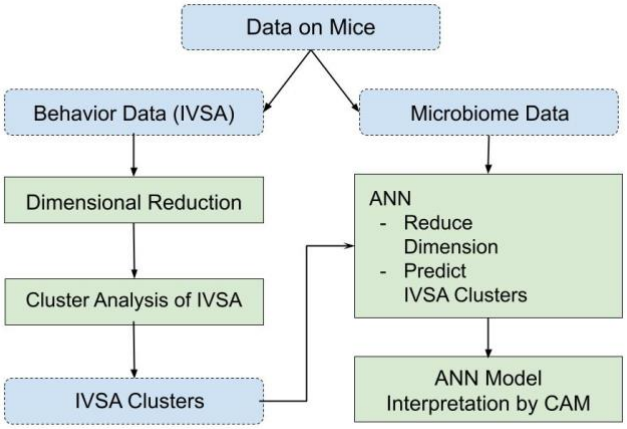


Figure 1. Pipeline of the proposed analytic steps (IVSA: intravenous self-administration of cocaine; ANN: artificial neural network; CAM: class activation map).

Dataset

A set of 175 DO mice between the ages of 8-12 weeks were sequenced on the 16S ribosomal RNA (rRNA) gene[23]. DO mice were derived from eight mouse founders and were inter-crossed in several generations to create a population of genetically heterogeneous mice, similar to natural selection such as in human communities[24]. The 16S rRNA gene consists of nine hyper-variable regions (V1 – V9), each exhibiting appreciable sequence diversity among varying bacteria[25]. As no region fully encapsulates all possible bacteria, the mice underwent high-density 16S rRNA sequencing on the V1 – V3 regions of fecal pellets[26]. The sequence reads were already processed at the Jackson Laboratories following rigorous protocols[27]. Ninety-six different genera were identified in this mouse population, and the gene read count for each genus was reported. Microbiome data were zero inflated. The microbial features that exhibited constant zeros for over 99% of the mice were removed because they were too sparse to create stable classifiers. As a result, 68 microbial features - counts of bacteria copies in 68 genera - remained.

Behavioral variable	Description of the variable
B1	Number of sessions taken to meet acquisition criteria during the acquisition stage.
B2	Number of sessions to stabilization for 1 mg/kg dose of cocaine.
B3	AUC calculated based on the numbers of infusions from the B4, B5, B6, and B7.
B4	Number of infusions at 1.0 mg/kg during the dose response stage.
B5	Number of infusions at 0.32 mg/kg during the dose response stage.
B6	Number of infusions at 0.1 mg/kg during the dose response stage.
B7	Number of infusions at 0.032 mg/kg during the dose response stage.
B8	Number of active lever presses during sessions of extinction.
B9	Number of inactive lever presses vs the number of saline active lever presses during reinstatement.

Table 1. Description of the nine variables of cocaine intravenous self-administration (IVSA).

These mice were also assessed with their behavior of intravenous self-administration (IVSA) of cocaine. In the first of these IVSA experimental sessions, mice pressed an active lever to receive an intravenous infusion of cocaine, along with an illumination stimulus of lights for five seconds. After an initial acquisition period at a 1.0 mg/kg dosage of cocaine, mice fulfilled the criterion for stabilization if the number of received infusions through the active lever did not vary by more than twenty percent for two consecutive sessions. Following the acquisition stage, the mice progressed through the following dosages: 0.32, 0.10, 0.032, and 1.0 mg/kg/infusion. To proceed to the subsequent dosage in the series, the mice had to meet the same stabilization criteria, or proceed at the same dosage until five days had passed. This period of varying dosages is known as the Dosage Response stage. The numbers of infusions received at each

of the dosage response stages were then plotted to form a curve and the area under the curve (AUC) was computed to characterize the overall dosage response trend for a mouse. At the conclusion of this phase, an extinction period was executed in which active lever presses did not provide any infusion of cocaine, nor activate the stimulus lights. Finally, a reinstatement period concluded the cocaine IVSA assessment, in which active lever presses were recorded, which would result in drug-paired stimuli while still lacking an infusion of cocaine. All these experiments resulted in 9 IVSA features. Table 1 lists these 9 behavioral features.

Dimension reduction

The linear dimension reduction - principal component analysis (PCA)[28]- and nonlinear dimension reduction - uniform manifold approximation and project (UMAP)[29, 30]- methods were used and compared. PCA maps high-dimensional data into fewer new coordinates to display sample patterns. It constructs these uncorrelated coordinates by finding principal components which are the linear combinations of original features. The principal components successively maximize the explained data variance. UMAP, on the other hand, first constructs a high-dimensional graph representation that connects each data point to its nearest neighbors within a variable radius. Then UMAP runs an iterative process to identify new coordinates which allow the same topology and proximity to be preserved in the new coordinate system. In other words, if sample A is closer to sample B than to sample C in the original high dimensional space, than the relationship should still hold in the lower dimensional space.

In this study, both methods were used to project the 9 IVSA features into 2 or 3 coordinates. Then the coordinates from each method were used in a cluster analysis to cluster mice into more homogeneous groups. The appropriate dimension reduction method and the best number of coordinates will be determined according to the quality of the clusters.

Cluster analysis

Dimension reduction often helps reveal sample grouping structures. The k means clustering method was used to create k (=2, 3, or 4) clusters of mice. Due to the limited sample size, smaller values of k were considered. For each clustering solution, Silhouette score[31] was used to measure how well the clusters were separated. The score equals (b a)/max(a,b) where a is the average of the distances between each pair of points within a cluster, and b is the average distance between each pair of clusters. Silhouette score ranges from -1 to 1, where a higher score indicates better quality of a clustering solution. Silhouette score was employed as a metric to compare and select the number of projected coordinates and the number of clusters. Clustering solutions were also visualized by a scatter plot for comparison.

After a clustering solution was selected, a violin plot was generated to visualize the cluster means, quartiles, and density curves of each behavioral feature. The violin plot helped detect plausible patterns in the behavioral clusters.

Classification

An ANN[32] was created to classify the mice into the identified clusters as a function of 68 microbial features. This ANN first learned a lower dimensional representation for microbial features using one or two hidden layers and then in the last layer performed classification. The ANN used fully connected layers, each consisting of a linear mapping followed by a ReLu

transfer except the output layer. The ReLu transfer takes in the output of the linear mapping and calculates a piecewise linear function called the rectified linear activation function. The output layer replaced the ReLu function with the softmax function which computes the probability of each mouse belonging to a cluster. A mouse was assigned to the cluster for which the highest probability was reported by the ANN.

A hold-out set was first created to contain 20% of the data sampled with stratification (i.e., the numbers of mice in each cluster in this set were proportional to those in the full sample). This hold-out set was used to evaluate the classification performance of the ANN model. A three-fold cross validation (CV) process[33] within the remaining 80% (training) data was used to tune the model parameters, such as, the number of hidden layers and the number of hidden nodes in each hidden layer, which eventually determined the number of reduced dimensions. During the ANN training, there were also tuning parameters such as the mini-batch size and learning rate which were also determined during this process. The training data was split into three even subsets and each time a model was trained using two subsets and tested on the remaining subset. We used the average validation classification performance to select the best model parameters. The Area Under the receiver operating characteristic Curve (AUC)[34] was used to measure the classification performance. AUC ranges from 0 to 1 and higher values indicate better classification performance. Using the selected model parameters, a final ANN was trained on all training data and its performance on the hold-out set was reported. Important microbial features were identified by model interpretation of this final ANN.

Model interpretation

Although ANN is generally considered as a black box, recent research has attempted to interpret its decision-making process. In particular, the class activation map[22, 35] is a powerful technique used in image classification to explain how an image is classified into a specific category. In this study, CAM visualized not only the class predicted for each mouse by the ANN, but also the specific gut bacteria that were critical for the classification. It calculated the gradient of the model prediction by backpropagating the network output back to the input and then multiplied the gradient with the input to produce an importance weight for each input microbial feature. Using each classifier (corresponding to a cluster), an importance weight vector was computed for a mouse. Then for each cluster, the Wilcoxon rank sum test[36] was used to test whether a microbial feature significantly differentiated a behavioral cluster from the remaining clusters by comparing the importance weights from all mice. The top four features identified by the rank sum tests were shown in a bar plot for comparison.

Results and Discussions

Dimension reduction and cluster analysis

The 9 behavioral variables were first normalized to have a mean of 0 and standard deviation of 1. PCA was used to map the 9 behavioral variables into 2 or 3 principal components, explaining respectively over 65% or 75% of the total variance in the behavior data. The K-means method was applied to the two sets of reduced data to obtain 2, 3, or 4 clusters, for which Silhouette scores were calculated respectively. These scores were all lower than the counterparts obtained by the nonlinear dimension reduction method UMAP. Table 2 summarizes and compares the clustering validity scores for the different numbers of projected coordinates and clusters. From Table 2, we observe that the 3-cluster solution based on the 3 reduced dimensions from UMAP

produced the best Silhouette score. Figure 2 shows the scatter plots of the resultant clusters in the respective coordinates calculated by PCA and UMAP for all solutions with 3 clusters, which visually confirms the same best clusters.

	# of Coordinates	# of Clusters	Silhouette Score
UMAP	2	2	0.62
		3	0.66
		4	0.36
	3	2	0.54
		3	**0.83**
		4	0.66
PCA	2	2	0.39
		3	0.41
		4	0.42
	3	2	0.33
		3	0.33
		4	0.34

Table 2. The silhouette scores calculated for the clustering solutions obtained by k-means using the reduced data from UMAP and PCA.

The three clusters from the selected solution consisted of 89, 53 and 33 mice respectively. Figure 3 shows the three violin plots each for a cluster. Each violin plot illustrates the cluster mean, quartile, and probability density curve of each of the nine behavioral features. Because all features B1-B9 have been normalized to have zero mean, when a behavioral feature has a smaller mean (<0) for a cluster, it indicates that mice in the cluster endorse the specific behavior less than an average mouse does; or otherwise, mice in the cluster endorse the behavior more than an average mouse. For the first and the largest cluster, because B1 and B2 have negative mean values, mice in this cluster used less numbers of infusion sessions than the sample average to meet the acquisition criteria and to stabilize on the initial dosage of 1mg/kg, indicating that these mice were quicker and easier to get addicted to cocaine. These mice had relatively larger numbers of infusions during the dosage response (DR) stages. Especially, the high tails of the violin bars show that this cluster has some mice reaching very high numbers of DR infusions and taking more lever presses during extinction sessions with easier reinstatement. Thus, this cluster corresponds to a "*high risk*" group of mice. In the second cluster, mice used the average numbers of sessions to reach acquisition and stabilize, but they took less cocaine (i.e., the numbers of infusions) in response to the different dosages and especially much less for smaller dosages. Hence, this cluster is named as the "*low use*" group. The third cluster shows an opposite trend to the first cluster on variables B1 and B2. Mice in the third cluster needed substantially more sessions to meet acquisition and stabilize on the initial dosage. They acquired average DR infusions but for lower dosage of 0.032 mg/kg they had a lower mean number of infusions. Thus, this cluster is overall a "*low risk*" group.

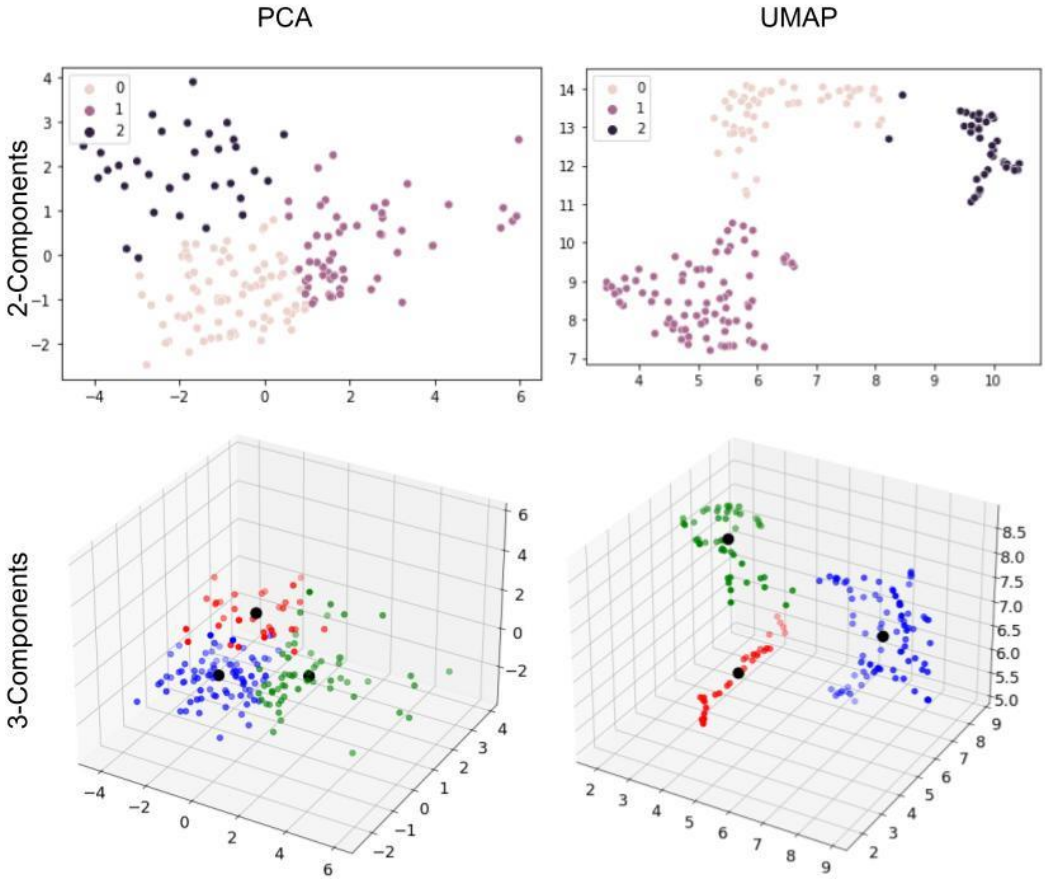


Figure 2. Comparison of clusters (n=3) obtained from the 2 or 3 projected coordinates of PCA and UMAP in scatter plots.

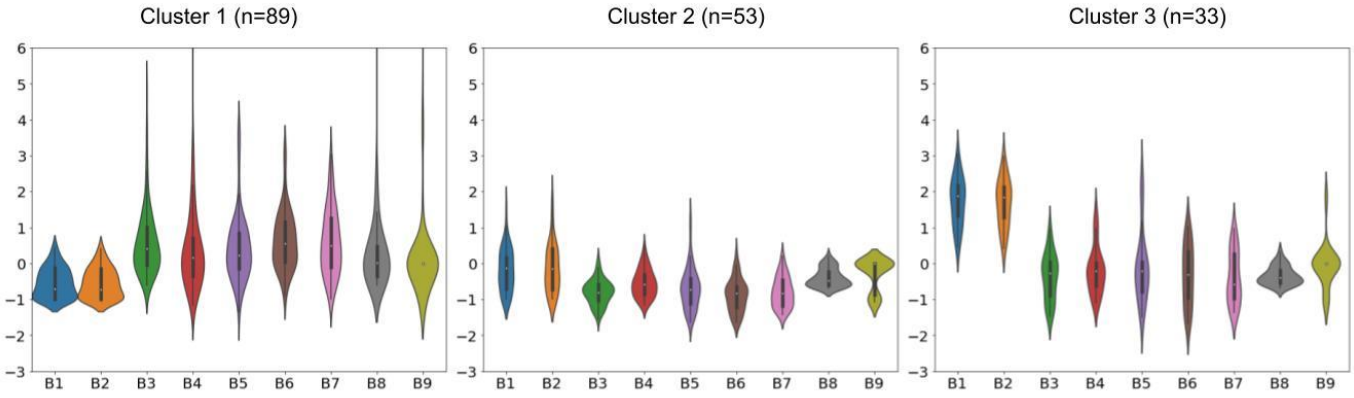


Figure 3. Violin plots of the three clusters, showing the cluster mean, quartile, and the distribution density of each behavioral feature for the clusters.

Classification and model interpretation

An ANN was trained to classify the mice into the respective *high risk*, *low use*, and *low risk* clusters based on the microbial features. Several choices for the number of hidden layers (L=1 or 2) and the number of hidden nodes were compared to determine the ANN architecture. Because dimension reduction was also a target of study when training the ANN, the numbers of hidden nodes were set to be smaller than the number of microbial features. When a single hidden layer was employed, 3, 5, and 8 hidden nodes were tested. When two hidden layers were used, a gradual reduction of the dimension was desirable, so the first hidden layer had more hidden nodes than the second hidden layer. Table 3 compares the different choices in terms of the AUC values each of which was calculated by averaging over all three clusters and all three cross validation splits. From this table, the choice of 2 hidden layers with 20 and 5 hidden nodes in the first and second layers produced the best classifier with an average AUC of 0.65, which was then selected. During the tuning of the model parameters, mini-batch size of 30, 50, and 60, and learning rate of 0.01 and 0.001 were also tested. For the selected ANN model, mini-batch size = 60 and learning rate = 0.01 were the best choices.

# of hidden layers	# of hidden nodes		Average AUC
L = 1	H1 = 3		0.50
	H1 = 5		0.56
	H1 = 8		0.53
L = 2	H1 = 10	H2 = 3	0.51
		H2 = 5	0.58
		H2 = 8	0.57
	H1 = 20	H2 = 3	0.51
		H2 = 5	**0.65**
		H2 = 8	0.60

Table 3. Comparison of average AUC values for the different choices of hidden layers and hidden nodes (where L denotes the number of hidden layers, H1 and H2 represent the number of hidden nodes in the first and second hidden layers respectively).

The full training set was then utilized to train an ANN model with the selected parameter setting. The resultant model was tested on the hold-out set, producing AUC values of 0.66 (in classifying the *high-risk* mice from the rest), 0.72 (*low use* mice versus the rest), and 0.59 (*low risk* mice versus the rest), and the micro- and macro-averaged AUC over the three clusters of 0.73 and 0.67, respectively. The classification performance for separating the *low-risk* mice from the rest had the worst performance partly due to small cluster size (n=33). Figure 4 shows the test ROC curves for the three class outputs of the ANN model and the related micro- and macro- average ROC curves. These results provide evidence that variation in the abundance of gut bacteria can explain to some level the difference in the behavioral response to cocaine because the microbial features show predictive power to the different behavioral clusters.

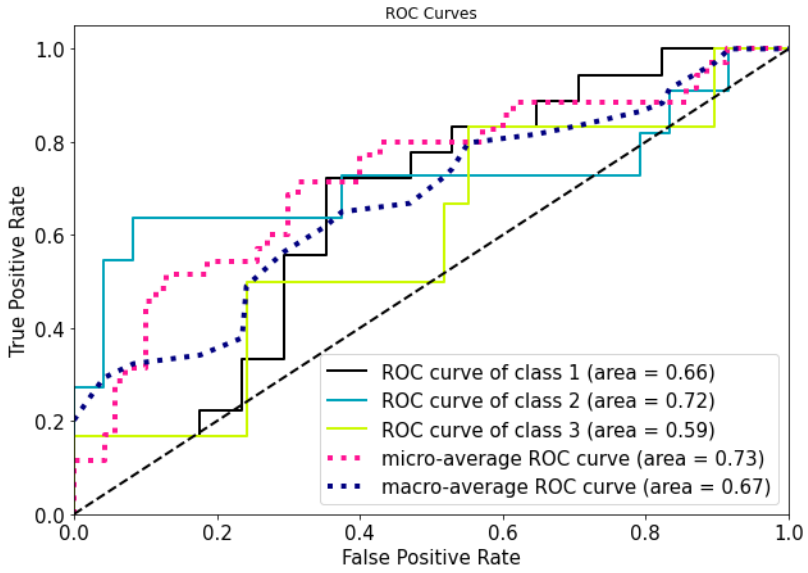


Figure 4. Receiver operating characteristics (ROC) curves for each behavioral cluster and the related macro- and micro-ROC curves.

For every mouse that was correctly classified by this ANN model, a CAM was computed using the backpropagation of the ANN output so one could evaluate which features were responsible for the correct class assignment. Then these CAMs were gathered for each cluster to apply the

Wilcoxon rank sum test to each microbe genus. This process identified quite a few microbial features that showed significant statistical power in distinguishing one cluster from the rest in a nonlinear interplay with other features specified by the ANN. Figure 5 lists the top four microbial features identified for each cluster. For the *high-risk* cluster, the strongest features identified had p values reaching 10^{-10}. For the *low-risk* cluster, the strong features had p values reaching 10^{-8}. A conjecture was that the *low use* cluster might be positioned in between the *high risk* and *low risk* clusters. Thus, it was more difficult to differentiate the *low use* mice from both the *high risk* and *low risk* mice, so the strongest feature to separate this cluster from the other two clusters received the best p value ~ 0.005 only.

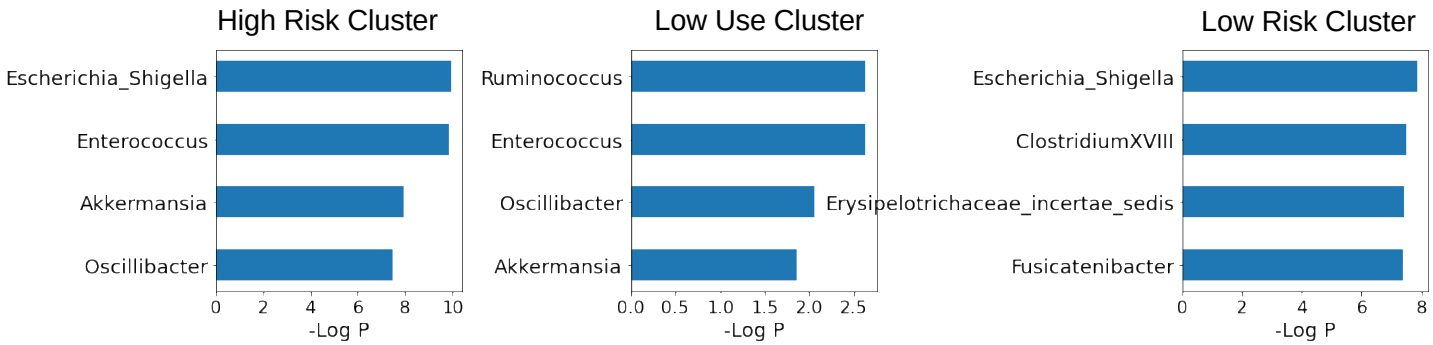
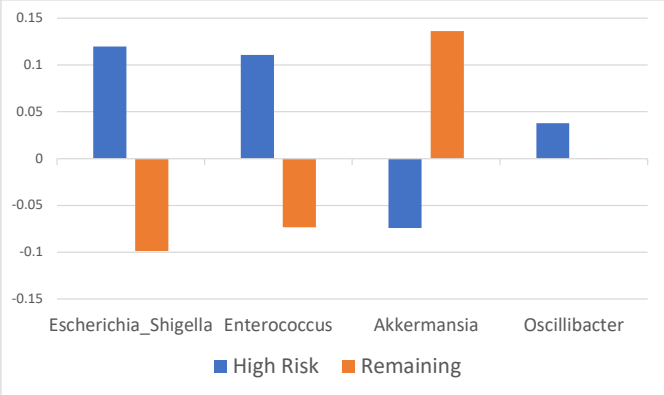
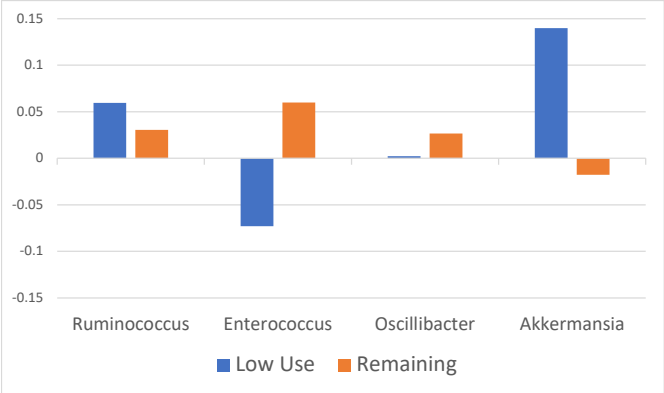


Figure 5. Top microbial features identified, by applying the Wilcoxon rank sum test to CAMs, to be useful in classifying mice in a specific cluster from the rest.

The identified features were further examined with their cluster means to study how they help separate a cluster of mice from the remaining mice. Figure 6 shows three groups of bar plots. Microbial features were also normalized to have zero mean. The bacterial genus *Escherichia Shigella* was identified to be the most significant feature to separate the *high-risk* mice from the remaining mice as well as to separate the low-risk mice from the rest. This feature exhibited clearly increased abundance for the *high-risk* mice on average and decreased abundance for the *low-risk* mice, so it was the most useful feature to discriminate between these two groups of mice. The genus *Enterococcus* was also selected twice to be the second significant feature to separate *high risk* mice from the rest and *low use* mice from the rest and showed an increased abundance among *high-risk* mice. Overall, an increased abundance of Escherichia *Shigella*, *Enterococcus*, and decreased abundance of *Akkermansia* helped differentiate high risk mice from other mice. A decreased abundance of Enterococcus and increased abundance of *Akkermansia* were associated with low cocaine use mice. The decreased abundance of *Escherichia Shigella* and *Clostridium XVIII* but increased abundance of *Erysipelotrichaceae incertae sedis* appeared to be protective of mice from getting addicted.





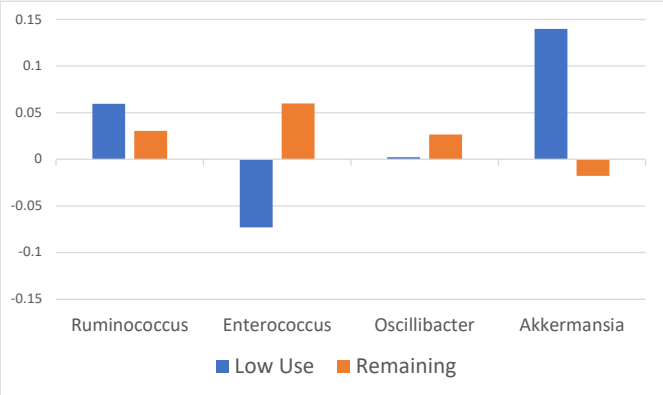


Figure 6. Grouped bar plots to compare the cluster means of the identified microbial features where Erysipelotricha… indicates the genus Erysipelotrichaceae incertae sedis. Features differentiating high risk mice, low use mice, and low risk mice respectively from the remaining mice are shown on the top, middle, and bottom plots.

Conclusion

In this study, a useful analytic pipeline, consisting of three major components: dimension reduction, cluster analysis, and classification, has been proposed to analyze the relationship between microbiome and cocaine use behavior. Linear and non-linear dimension reduction methods were tested so that the k-means method could find clusters of mice based on their behavioral patterns. The non-linear dimension reduction method, UMAP, was shown to be better in this example for creating homogenous groups, with a Silhouette cluster validity score much higher than those based on principal components. The k-means algorithm found three clusters, characterizing mice with high risk for addiction, low cocaine use, and low addiction risk, respectively. While it is possible to create a neural network based on microbiome to directly predict the behavioral variables, because of the high heterogeneity in cocaine use behavior, it would be difficult to interpret which microbe is related to the risk for addiction. Instead, the proposed pipeline allows a refined classification of cocaine use patterns to identify the gut bacteria that potentially relate to the classification. A neural network with fully connected layers was tuned and trained to classify mice into one of the three categories based on the abundance of 68 bacterial genera and produced high AUC values, which justifies evaluating the critical features used in this neural network. Using a model interpretation strategy, it was found that an increased abundance of *Escherichia Shigella*, *Enterococcus*, *Clostridium XVIII* and decreased abundance of *Akkemensia* and *Erysipelotrichaceae Incertae Sedis* were risk factors. Future research directions may expand the analyses to human samples, and/or include more microbial species in the analysis. It may be useful to handle the zero inflation commonly encountered in microbial features rather than removing sparse features as done in this work.

Acknowledgements

Anthony B Song would like to thank Drs. George Weinstock and Dong-Binh Tran for their guidance during his research, and for providing him the mouse data. Anthony also acknowledges the Academic Year Fellowship awarded by the Jackson Laboratory for Genomic Medicine during which this research was done.

References

1. Consortium, H. M. P., Structure, function and diversity of the healthy human microbiome. *nature* **2012,** *486* (7402), 207-214.
2. SAMHSA, S. A. a. M. H. S. A., Behavioral health trends in the United States: Results from the 2014 national survey on drug use and health. **2014**, http://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf. 3. Kampman, K. M., The treatment of cocaine use disorder. *Science advances* **2019,** *5* (10), eaax1532.
4. Goldstein, R. Z.; Volkow, N. D., Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *American Journal of Psychiatry* **2002,** *159* (10), 1642-1652.
5. Li, C.-s. R.; Sinha, R., Inhibitory control and emotional stress regulation: Neuroimaging evidence for frontal–limbic dysfunction in psycho-stimulant addiction. *Neuroscience & Biobehavioral Reviews* **2008,** *32* (3), 581-597.
6. Meckel, K. R.; Kiraly, D. D., A potential role for the gut microbiome in substance use disorders. *Psychopharmacology* **2019,** *236* (5), 1513-1530.
7. Russell, J. T.; Zhou, Y.; Weinstock, G. M.; Bubier, J. A., The Gut Microbiome and Substance Use Disorder. *Frontiers in Neuroscience* **2021**, 1134.
8. Mayer, E. A.; Tillisch, K.; Gupta, A., Gut/brain axis and the microbiota. *The Journal of clinical investigation* **2015,** *125* (3), 926-938.
9. Babor, T. F.; Caetano, R., Subtypes of substance dependence and abuse: implications for diagnostic classification and empirical research. *Addiction* **2006,** *101* (Suppl. 1), 104-10. 10. Moloney, R. D.; Desbonnet, L.; Clarke, G.; Dinan, T. G.; Cryan, J. F., The microbiome: stress, health and disease. *Mammalian genome* **2014,** *25* (1), 49-74.
11. Nestler, E. J., The neurobiology of cocaine addiction. *Science & practice perspectives* **2005,** *3* (1), 4.
12. Volpe, G. E.; Ward, H.; Mwamburi, M.; Dinh, D.; Bhalchandra, S.; Wanke, C.; Kane, A. V., Associations of cocaine use and HIV infection with the intestinal microbiota, microbial translocation, and inflammation. *Journal of studies on alcohol and drugs* **2014,** *75* (2), 347-357.
13. Xu, Y.; Xie, Z.; Wang, H.; Shen, Z.; Guo, Y.; Gao, Y.; Chen, X.; Wu, Q.; Li, X.; Wang, K., Bacterial diversity of intestinal microbiota in patients with substance use disorders revealed by 16S rRNA gene deep sequencing. *Scientific reports* **2017,** *7* (1), 1-9.
14. Kiraly, D. D.; Walker, D. M.; Calipari, E. S.; Labonte, B.; Issler, O.; Pena, C. J.; Ribeiro, E. A.; Russo, S. J.; Nestler, E. J., Alterations of the host microbiome affect behavioral responses to cocaine. *Scientific reports* **2016,** *6* (1), 1-12.
15. Sun, J.; Kranzler, H. R.; Gelernter, J.; Bi, J., A genome-wide association study of cocaine use disorder accounting for phenotypic heterogeneity and gene–environment interaction. *Journal of Psychiatry and Neuroscience* **2020,** *45* (1), 34-44.
16. Morris, A.; Beck, J. M.; Schloss, P. D.; Campbell, T. B.; Crothers, K.; Curtis, J. L.; Flores, S. C.; Fontenot, A. P.; Ghedin, E.; Huang, L., Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American journal of respiratory and critical care medicine* **2013,** *187* (10), 1067-1075.
17. Lin, S. W.; Freedman, N. D.; Shi, J.; Gail, M. H.; Vogtmann, E.; Yu, G.; Klepac-Ceraj, V.; Paster, B. J.; Dye, B. A.; Wang, G. Q., Beta-diversity metrics of the upper digestive tract microbiome are associated with body mass index. *Obesity* **2015,** *23* (4), 862-869.
18. Sharma, D.; Xu, W., phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. *Bioinformatics* **2021,** *37* (21), 3707-3714.
19. Namkung, J., Machine learning methods for microbiome studies. *Journal of Microbiology* **2020,** *58* (3), 206-216.
20. Oh, M.; Zhang, L., DeepMicro: deep representation learning for disease prediction based on

microbiome data. *Scientific reports* **2020,** *10* (1), 1-9.
21. Marcos-Zambrano, L. J.; Karaduzovic-Hadziabdic, K.; Loncar Turukalo, T.; Przymus, P.; Trajkovik, V.; Aasmets, O.; Berland, M.; Gruca, A.; Hasic, J.; Hron, K., Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in microbiology* **2021**, 313.
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. In *Learning deep features for discriminative localization*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 2921-2929.
23. Soriano-Lerma, A.; Pérez-Carrasco, V.; Sánchez-Marañón, M.; Ortiz-González, M.; Sánchez-Martín, V.; Gijón, J.; Navarro-Mari, J. M.; García-Salcedo, J. A.; Soriano, M., Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Scientific reports* **2020,** *10* (1), 1-13.
24. Dickson, P. E.; Ndukum, J.; Wilcox, T.; Clark, J.; Roy, B.; Zhang, L.; Li, Y.; Lin, D.-T.; Chesler, E. J., Association of novelty-related behaviors and intravenous cocaine self administration in Diversity Outbred mice. *Psychopharmacology* **2015,** *232* (6), 1011-1024.
25. Chakravorty, S.; Helb, D.; Burday, M.; Connell, N.; Alland, D., A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods* **2007,** *69* (2), 330-339.
26. Servick, K., Of mice and microbes. *Science* **2016,** *353* (6301), 741-743. 27. Tran, T. D. B.; Nguyen, H.; Sodergren, E.; Dickson, P. E.; Wright, S.; Philip, V. M.; Weinstock, G. M.; Chesler, E. A.; Zhou, Y.; Bubier, J. A., Microbial glutamate metabolism predicts intravenous cocaine self-administration in Diversity Outbred mice. *bioRxiv* **2022**. 28. Bro, R.; Smilde, A. K., Principal component analysis. *Analytical methods* **2014,** *6* (9), 2812- 2831.
29. Armstrong, G.; Martino, C.; Rahman, G.; Gonzalez, A.; Vázquez-Baeza, Y.; Mishne, G.; Knight, R., Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. *Msystems* **2021,** *6* (5), e00691-21.
30. McInnes, L.; Healy, J.; Melville, J., Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
31. Shahapure, K. R.; Nicholas, C. In *Cluster quality analysis using silhouette score*, 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE: 2020; pp 747-748.
32. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *nature* **2015,** *521* (7553), 436-444. 33. Refaeilzadeh, P.; Tang, L.; Liu, H., Cross-validation. *Encyclopedia of database systems* **2009,** *5*, 532-538.
34. Fawcett, T., An introduction to ROC analysis. *Pattern Recognition Letters* **2006,** *27* (8), 861-874.
35. Zhu, L.; She, Q.; Chen, Q.; Meng, X.; Geng, M.; Jin, L.; Jiang, Z.; Qiu, B.; You, Y.; Zhang, Y., Background-aware Classification Activation Map for Weakly Supervised Object Localization. *arXiv preprint arXiv:2112.14379* **2021**.
36. Hogg, R. V.; Tanis, E. A.; Zimmerman, D. L., *Probability and statistical inference*. Pearson/Prentice Hall Upper Saddle River, NJ, USA:: 2010.

Author

Anthony Song is a senior at Edwin O. Smith High School. He is passionate about computer science and bioinformatics. He is the founder of the coding club in his school. His research proposal to the Beamline for School competition was ranked top 2 in the US 2021. His work on autonomous driving was selected for presentation at an IEEE conference 2022. He will pursue college education in machine learning and its biomedical applications.