UNIVERSITY OF CALIFORNIA SAN DIEGO

A Metabolomic and Microbial Approach to Diagnosing Advanced Fibrosis in NAFLD Patients

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Tobin Escher Groth

Committee in charge:

Professor Rob Knight, Chair
Professor Joseph Pogliano, Co-Chair
Professor Alisa Huffaker

2020

The thesis of Tobin Escher Groth is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AF	Advanced Fibrosis
AUPR	Area Under Precision-Recall
AUROC	Area Under Receiver Operating Characteristic
CA	Cholic Acid
CV	Cross-Validation
GBA	Glycinated Bile Acid
HCC	Hepatocellular Carcinoma
LR	Logistic Regression
LCFA	Long-Chain Fatty Acid
ML	Machine Learning
NAFLD	Non-alcoholic Fatty Liver disease
NASH	Non-Alcoholic Steatohepatitis
PC	Phosphocholine
PD	Phylogenetic Diversity
PE	Phosphatidylethanolamine
PR	Precision-Recall
RF	Random Forest
ROC	Receiver Operating Characteristic

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis is coauthored with Groth, Tobin; Knight, Rob; Dorrestein, Pieter; Tripathi, Anupriya; Fogelson, Kelly; Rahman, Gibraan. The thesis author was the primary author of this material.

ABSTRACT OF THE THESIS

A Metabolomic and Microbial Approach to Diagnosing Advanced Fibrosis in NAFLD Patients

by

Tobin Escher Groth

Master of Science in Biology

University of California San Diego, 2020

Professor Rob Knight, Chair
Professor Joseph Pogliano, Co-Chair

Non-alcoholic fatty liver disease (NAFLD) has emerged as a health crisis not only in the US but throughout the world. Current methods of identifying indicators of liver disease progression, such as advanced fibrosis (AF), are accurate but costly, slow, and impossible to perform on population-wide studies. To fill in this diagnostic gap, we look to build machine learning (ML) models using serum and fecal metabolomics data from NAFLD patients that can accurately classify AF. These patients encompass the entire spectrum of NAFLD which allows us to dig into the diversity and composition of their metabolomes to understand the changing metabolites between different liver disease states. Using these metabolites, we built baseline ML models and improved performance using a multitude of feature selection methods. These feature selection methods allowed us to develop a fecal metabolome ML model that could classify AF with an area under the receiver operating characteristic (AUROC) of 0.82. To further develop our findings, we combined our strongest fecal metabolome results with fecal microbial

features from a previous study to build a multi-omic machine learning model. Building this multi-omic ML model expanded on our results and led us to a stronger performing model that could accurately classify AF patients with an AUROC of 0.86.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is the most common liver disease in westernized countries, affecting roughly 1 in 3 adults in the US, and approximately 25% of adults globally (1). NAFLD is characterized by an increased accumulation of intracellular fat in hepatocytes and is often referred to as the hepatic manifestation of metabolic syndrome (2). Although the majority of NAFLD cases persist without complication, roughly 20% of individuals progress to more advanced stages of NAFLD, such as non-alcoholic steatohepatitis (NASH) (3). NASH results in inflammation and pathological wound healing, therefore increasing the risk of developing advanced fibrosis (AF), irreversible liver cirrhosis, and hepatocellular carcinoma (HCC), which has a 5-year survival rate of only 10% in the United States (4).

Despite its prevalence, limited options exist for diagnosing NAFLD, NASH, and AF, thus increasing the rate of late diagnosis and mortality from cirrhosis and HCC. Currently, the gold standard for diagnosing liver disease is through biopsy. Although effective, biopsies are highly invasive, expensive, and pose the risk of complication (5). Further, biopsies are not a feasible approach for routine screening or when completing population-wide studies (1). Thus, there is a critical need to develop robust, minimally invasive diagnostic tools capable of identifying the reversible stages of NAFLD: NASH and AF.

The gut microbiome-liver-axis is closely linked via the portal vein, biliary tract, and systemic circulation. Previous research supports that machine learning algorithms trained on key gut microbial features are capable of diagnosing advanced fibrosis with an AUROC of 0.87 (6). Through their analysis, the previous research was also able to identify key microbes that were in higher abundance for NAFLD-cirrhosis patients compared to patients at earlier stages of liver disease. Metabolites are the tools for communication between the liver and microbes and previous studies support the notion that these metabolites provide enough information to diagnose AF in NAFLD patients (7,8). In this study, we propose a method for identifying key metabolites from both serum and fecal metabolomic data to build a machine learning (ML) model that can diagnose AF with an accuracy comparable to ML models built on microbial data. Following our identification of these key metabolites, we then looked to combine the previous microbial data with our metabolomic features to build a multi-omic ML model to diagnose AF.

Methods

Diversity Analysis

We used diversity scores from different alpha and beta diversity metrics as a way to identify initial differences between our disease groups. For the serum metabolome, we compared Faith's PD and unweighted UniFrac between 55 non-NAFLD controls, 19 NAFLD without AF probands, and 25 NAFLD-cirrhosis probands, Table 1. For the fecal metabolome, we compared Faith's PD and unweighted UniFrac between 56 non-NAFLD controls, 18 NAFLD with AF probands, and 24 NAFLD-cirrhosis probands, Table 2. By identifying differences in our disease states we would know whether or not we could use ML models to try and differentiate between these groups. Before performing the analysis, we normalized our data using quantile normalization (20). Once the data was normalized we looked into both alpha and beta diversity. For alpha diversity, we looked at Faith's Phylogenetic Diversity. For beta diversity, we looked at unweighted UniFrac (9). These distance metrics were calculated using Qiime2 and SciPy and were visualized with Matplotlib in a jupyter notebooks python environment (21,22,23).

Composition Analysis

To understand the specific metabolite differences between disease states, we used composition analysis for both serum and fecal metabolome data. We separated the samples into disease states: Non-NAFLD probands and relatives, NAFLD without AF probands and relatives, and NAFLD-cirrhosis probands and relatives. The counts for each group follows the same counts used in the diversity analysis, and can be found in Tables 1 and 2. After separating the samples into groups, we then counted the relative abundance of our metabolomic features based on taxonomic groups. The hierarchy of the groups is as follows: kingdom; superclass; class; subclass; direct parent. By splitting up the metabolomic features this way we could compare the composition of the disease states at whichever taxonomic level we saw fit. Taxonomy was determined using qemistree fingerprints (24). We plotted these relative abundances using Matplotlib in a jupyter notebooks python environment (22).

Machine Learning Model Building

To perform ML analysis we first had to standardize both the fecal and serum feature tables. We chose to use sklearn's StandardScaler as we would be using the sklearn python library for our ML analysis (13). After the feature tables were processed, we split the data into cross-validation(CV)/train and test groups. For CV/training we used the non-NAFLD probands and the NAFLD-cirrhosis probands to expose the models to AF positive (NAFLD-cirrhosis) and AF negative (non-NAFLD) samples. For CV, we performed a 5-fold CV using sklearn's StratifiedShuffleSplit for each model. When testing the models, we tested on NAFLD-cirrhosis first degree relatives to expose the model to a mix of AF positive and negative samples. We measured the performance of these models using ROC and PR curves. The models that we used for our analysis were the RandomForestClassifier and the LogisticRegression from sklearn. Our results focus on the RF model as this specific ML model has shown strong results in previous analyses (6,8). When building the models we used mostly default parameters but for RF, n_estimators=500, and for logistic regression max_iter was set to the lowest threshold such that the model converged, usually set to 200 but never higher than 10,000. All ML model performance metrics can be found in Tables 4,5,6.

Feature Selection

To improve ML model accuracy, we employed a number of feature selection methods. We used the RF built-in feature importance scoring to iteratively reduce the number of features by only retaining features that met a certain feature importance threshold. We also used the biological relevance of features to create a subset of features. For biological relevance, we used the results from the composition analysis as well as literature review to guide our selection. From the serum features, we selected bile acids and two types of phosphocholine (179 total features). From the fecal features, we selected bile acids and fatty acids (91 total features). The final method of feature selection that we employed was the use of Songbird to identify key features that are highly associated with AF status (25). When building a Songbird model for the serum data we used default parameters as well as --p-formula "ATTRIBUTE_adv_fibrosis", --p-differential-prior 0.4, --p-epochs 5000 (Q^2=0.349344). When building the Songbird model for the fecal data we used default parameters as well as --p-formula

"ATTRIBUTE_adv_fibrosis+ATTRIBUTE_groups", --p-differential-prior 0.4, --epochs 2500 (Q^2=0.028746). Using these models we built 3 different feature sets: top 50 features associated with AF, bottom 50 features associated with AF, and top 25+bot 25 features associated with AF. Once we created a feature set for each of these unique criteria we followed our normal method of CV, training, and testing the model.

Results

Study Population Characteristics

This analysis was based on a study that recruited patients encompassing the full spectrum of NAFLD (non-NAFLD controls, NAFLD without AF, NAFLD-cirrhosis) as well as their first degree relatives. Our analysis made use of 192 total patients, made up of 25 NAFLD-cirrhosis probands and 34 first degree relatives, 18 NAFLD without AF probands and 17 first degree relatives, as well as 56 non-NAFLD controls and 42 of their first degree relatives. Patient samples were included or excluded from the subsequent analysis depending on fecal and serum metabolomics data availability. Patient information and specific demographic counts for the serum, fecal, and multi-omic analysis can be found in Tables 1,2, and 3.

Serum and Fecal Metabolomic Profiling

We first assessed the overall metabolomic diversity for our serum and fecal metabolome. We compared the diversity metrics between probands of our different disease states (non-NAFLD, NAFLD without AF, and NAFLD-cirrhosis). We looked at both alpha diversity (intra-sample diversity) as well as beta diversity (inter-sample diversity) to identify changes as disease state increased. To investigate alpha diversity we used Faith's Phylogenetic Diversity (PD). In the serum metabolome, we identified an overall decrease in diversity for NAFLD-cirrhosis samples compared to our non-NAFLD controls ($p<0.001$) and NAFLD without AF probands ($p<0.001$), Fig. 1A. Within the fecal metabolome, we did not match this trend and saw an increase in Faith's PD for NAFLD without AF probands compared to the other disease groups (no significance), Fig 1C. For serum beta diversity, we used the unweighted UniFrac metric (9) and found

that within the serum metabolome, our NAFLD without AF probands had the lowest diversity score compared to non-NAFLD controls ($p<0.001$) and NAFLD-cirrhosis probands, Fig 1B. These results match a trend identified in a previous analysis investigating microbial diversity in NAFLD patients, which suggests an overall decrease in metabolome diversity at early disease stages that develop into a dysbiotic, more diverse metabolome as patients progress to late-stage liver disease(6). The unweighted UniFrac results for the fecal metabolome show an overall decrease in diversity as disease state increased where non-NAFLD controls exhibited higher diversity scores than both NAFLD without AF probands ($p<0.001$) and NAFLD-cirrhosis probands ($p<0.001$), Fig. 1D. The fecal metabolome beta diversity trend suggests that as liver disease progresses, the metabolome becomes less diverse and more similar between late-stage disease patients. All of our diversity results, barring alpha diversity in the fecal metabolome, show that metabolites in fecal and serum samples are changing as disease state increases in severity, supporting our hypothesis that we can use these metabolite changes to diagnose AF in our NAFLD patients.

The results of our diversity analysis show that there are changes in the fecal and serum metabolome as disease state increases in our samples. To understand the specifics of the changing metabolome, we assessed the relative abundance of metabolites at the direct parent level and compared these abundances between our disease states. For the serum metabolome, Glycinated bile acids (GBAs) and some phosphocholine (PC) were enriched in NAFLD-cirrhosis probands, while another form of phosphocholine was seen at decreased levels, Fig. 2A. A potential explanation for the shift in GBA relative abundance could be related to the effects of liver disease on the microbiome. As liver disease increases in severity, a previous study identified that gram-positive bacteria decrease in abundance while gram-negative bacteria increase in abundance (6). Gram-positive bacteria have also been recognized to play a role in deconjugating conjugated bile acids, like GBAs (10). The decrease in abundance of gram-positive bacteria in late stage liver disease could be a cause for the overall increase in abundance of GBAs in NAFLD-cirrhosis probands. Changes in PC abundance within liver disease is recognized by a review discussing the importance of maintaining a consistent ratio of PC to phosphatidylethanolamine

(PE) to avoid disease progression (11). In the fecal metabolome, NAFLD-cirrhosis probands are enriched with a different kind of bile acid, trihydroxy bile acids (also known as cholic acid (CA)), while long-chain fatty acids (LCFA) are decreased in cirrhosis probands, Fig. 2B. CA has been linked with increases in inflammation, a characteristic of liver disease progression, suggesting that CA enrichment could be contributing to the progression to later stages of liver disease (10). The decrease in LCFAs is harder to characterize but it could be due to an overall increase in uptake of LCFAs commonly seen in the steatotic livers of rats (12). By investigating the metabolomic composition of patients in different disease states, we identified specific metabolomic changes that our diversity analysis results hinted at. The combination of these results supports our overall goal of using fecal and serum metabolites to build machine learning models for AF classification.

Build a Metabolome-Derived ML Model for AF Classification

Following the results of our composition and diversity analysis of the fecal and serum metabolome, we began to build ML models to diagnose AF in our NAFLD patients. We used Random Forest (RF) Classifier and Logistic Regression (LR) models from sklearn to classify samples (13). Our results focused on the RF model but we used the LR model as a linear control. In our model building process, we used a wide variety of feature selection methods to identify the strongest subset of features for the serum and fecal metabolome. Our feature selection process is detailed in the Methods section.

The best performing RF serum model was built off 23 features (20 serum metabolites, and 3 metadata features: BMI, age, sex), and achieved an AUROC of 0.64, Fig 3C. The performance of the serum model was hindered greatly by overfitting, which was identified by comparing CV scores to test scores, Table 4. Due to the consistent overfitting seen with the serum model, we did not include serum features in any downstream analyses.

The strongest RF fecal ML model was built from the top 25 and bottom 25 features associated with AF, as determined by Songbird, as well as 3 metadata features (BMI, age and sex). Using these features the model achieved an AUROC of 0.82 and classified AF positive samples with an area under

the precision-recall (AUPR) of 0.76, Figure 4C. Prior to including the metadata categories, we investigated the specific features that made up this feature set to identify what our potential biomarkers could be. When looking at the features at the subclass taxonomic level we see a majority of our feature importance is reliant on unclassified features, with amino acids, fatty acids and bile acids making up the next largest majority of importance, Fig 5. It is unfortunate that a majority of the features are unclassified, but it is reassuring to see that both bile acids and fatty acids appear again, similar to the results of our fecal composition analysis, Fig 2B. By validating our composition results during our fecal model classification, we support that we are using the trend we previously identified to classify our samples for AF.

Building a Multi-Omic ML Model for AF Classification

After identifying our strongest metabolome feature set, we combined our fecal metabolome results with results from a previous analysis to build a multi-omic model for classifying AF. The previous analysis used 16S sequencing data from the same NAFLD patients to build a model using microbial features (6). Since our features were from an overlap in samples, we filtered our samples so that we only included samples that had fecal metabolome and fecal microbiome features, Table 3. We didn't include any serum metabolome features in our multi-omic model building process due to the overfitting issues seen previously in our analysis, Table 4.

The top-performing multi-omic model was built off 27 fecal microbial features, 14 fecal metabolites, and 3 metadata categories (BMI, age, sex) and was able to achieve a AUROC of 0.86, Fig 6C. The features for this model were derived initially from the best fecal metabolome model, Fig 4C, as well as the top predictive microbial features from Caussey et al. Most of the fecal metabolite features were removed during the feature selection process. The strong results of the multi-omic model suggest that combining different types of data does not cause overfitting or poor performance for RF ML models. During our feature selection process, we removed 36 fecal metabolites which suggests that the ML model relies more heavily on fecal microbial features when classifying AF. However, the importance of fecal

metabolites cannot be overlooked as suggested by the performance of our fecal metabolome ML model, Fig 4C.

Discussion

In this study, we used features from the fecal and serum metabolome to build ML models that could accurately diagnose AF within NAFLD patients. While the serum metabolome models did not garner the strongest of results due to overfitting (Figure 3), the best performing fecal metabolome ML model (Songbird Top25+Bot25 features plus BMI, age, and sex) was able to diagnose AF in NAFLD patients with an AUROC of 0.82. The results from the fecal metabolome model provide substantial evidence in support of our hypothesis that we could use metabolites and machine learning models as a potential alternative to more invasive forms of AF diagnosis. We substantiated these results by combining our fecal metabolite features, and metadata features with fecal microbial features from a previous study, to build a fecal multi-omic model that could diagnose AF with an AUROC of 0.86.

Building a machine learning model off of metabolomics data is important not only because it shows the potential for metabolomics to diagnose AF, but also shows that these changing metabolites are shifting in accordance with microbes identified in previous studies. Our study pulled metabolomics data from the same patients as Caussy et al. who identified similar changes in diversity as well as composition for the fecal microbiome. Across both studies, we see trends of decreasing diversity as disease state increases, within patients for the serum metabolome and microbiome as well as between patients in the fecal metabolome. The serum metabolome and fecal microbiome also mimic the hourglass shaped results seen when comparing beta diversity between disease states. Though it can be hard to compare composition between microbes and metabolites, we do see consistent trends. Some of the key species of microbes that were enriched in NAFLD-cirrhosis patients were from the family Enterobacteriaceae which shows not only a shift towards gram-negative bacteria, but Enterobacteriaceae have also been identified to proliferate following changes in bile acid synthesis (6,14). In both fecal and serum composition we saw

changes in relative abundance for bile acids, a common theme in liver disease, but we can now relate this change to specific microbes that have been previously identified.

Following these results, we would like to see future studies make use of multiple types of metabolomic data and investigate the co-occurence of microbes and metabolites. In our case, the serum metabolome posed significant issues in combining metabolome data and prevented us from building a multi-omic model that incorporated serum metabolites. Serum metabolites are still very informative and can provide strong results when using them to identify NAFLD (15), but further investigation is needed to understand their relevance to AF and liver disease progression. Our results and the results from previous analyses (16) suggest associations between metabolites, microbes, and the liver in regard to liver disease. The microbiome has already been shown to have a direct role in affecting human metabolism, and we believe that this interaction cannot be overlooked, especially in regard to liver disease (7,17). Tools such as MMVEC are able to use metabolomics and microbial data to investigate the co-occurrence and therefore the interaction between specific microbes and metabolites (18). Future studies that have access to microbial and metabolomic data could use tools like these to further understand the specifics of microbe-metabolite interactions and use this information to identify co-occurring biomarkers.

We acknowledge the following limitations of this study. Firstly, the discrepancy between the samples used between the fecal, serum, and multi-omic portions of the analysis. We understand that this keeps us from being able to directly compare the performance metrics and diversity results across our study, but the standalone results still support our initial hypothesis of a changing metabolome with liver disease progression. Secondly, this study makes use of a number of metabolomic features that can notoriously hinder studies due to batch effects (19). As seen in our serum ML models, the performance of the models can be greatly affected by these batch effects, but through our comparison of CV and testing scores, we remain certain that only the serum portion of this analysis was affected by these batch effects. Lastly, the associations we have identified throughout this study do not necessarily confirm causality. Further studies using similar data types are needed to confirm the results we have identified and

substantiate our hypothesis that fecal and serum metabolites can be used alone or in conjunction with microbes to build a machine learning model capable of classifying AF. However, the strengths of this study include a multi-part validation of metabolome changes during liver disease progression as well as a thorough feature selection process ensuring peak performance for ML models.

FIGURES

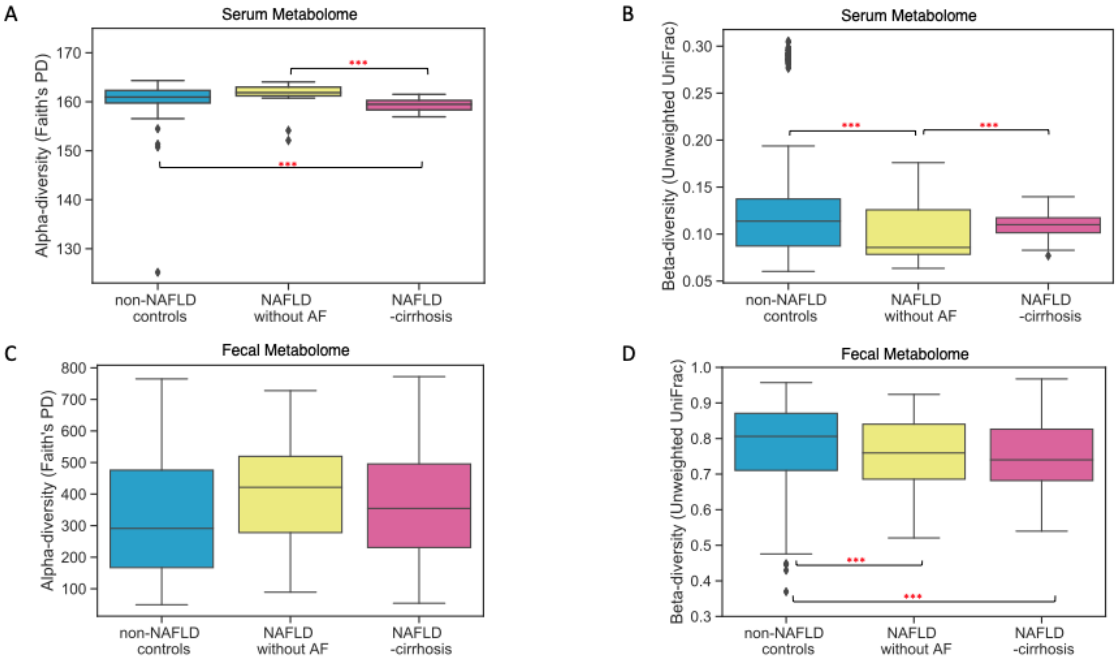

Figure 1: Serum and fecal metabolite diversity shifting between disease states. For serum metabolite diversity, **A.** Faith's PD comparison shows an overall decrease in diversity in the NAFLD-cirrhosis groups compared to non-NAFLD controls ($p<0.001$) and NAFLD without AF probands ($p<0.001$). **B.** Unweighted UniFrac for the serum metabolites shows a drop in diversity for NAFLD without AF probands compared to non-NAFLD controls ($p<0.001$) as well as NAFLD-cirrhosis probands ($p<0.01$). **C.** Faith's PD for fecal metabolites does not show the same trend, rather an increase in diversity for moderate and late-stage disease states (no significance). **D.** Unweighted UniFrac for fecal metabolites shows non-NAFLD controls having significantly higher diversity compared to NAFLD without AF probands ($p<0.001$) and NAFLD-cirrhosis probands ($p<0.001$). Group counts for each comparison can be found in Table 1 for serum metabolites and Table 2 for fecal metabolites. The box plots show the quartiles and whiskers show the rest of the distribution (1.5 inter-quartile range) the center liner corresponds to the median. All p-values were determined using a two-sided Kruskal-Wallis test.

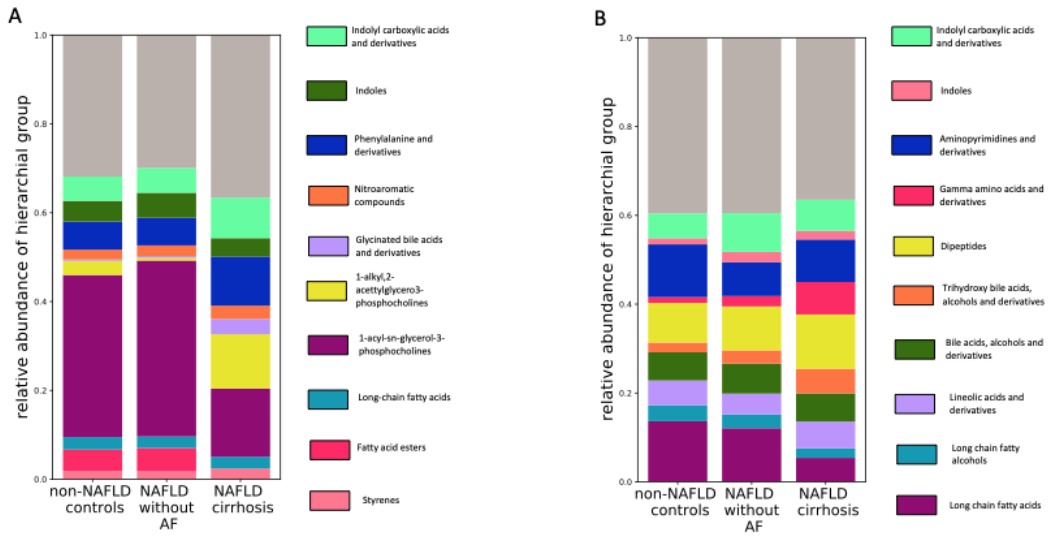

Figure 2: Serum and fecal metabolite composition differences between disease groups.
A. Serum composition differences between groups, viewed at the direct parent taxonomic level. **B.** Fecal composition differences between disease groups, viewed at the direct parent taxonomic level. Metabolites are compared using relative abundance. Taxonomy based on qemistree fingerprint classification (24). Group counts for each composition plot can be found in Table 1 for serum metabolites and Table 2 for fecal metabolites.

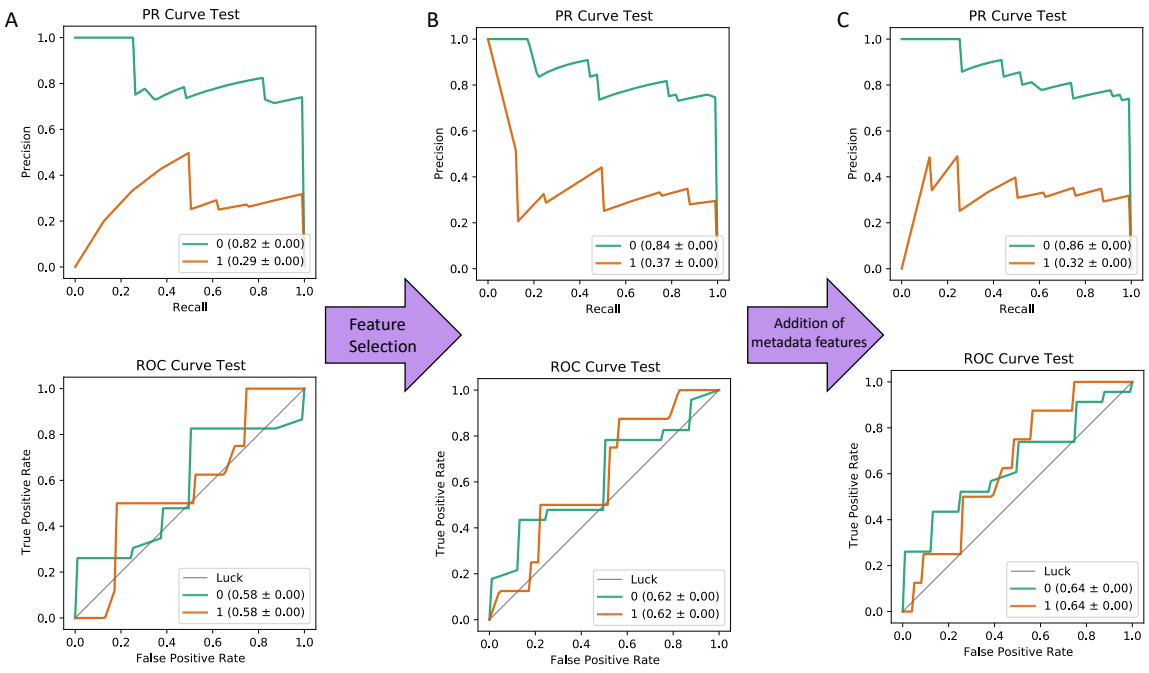

Figure 3: Serum metabolome machine learning model improvement. A. The PR and ROC test scores for the base model built off all serum metabolites (1847 total features). **B.** A similar plot showing the PR and ROC test scores following feature selection. The model shown is the Further Selected model that was built from features with a feature importance >0.02 (20 total features). **C.** The final best performing serum model that was built off the Further Selected features as well as the metadata features BMI, age and sex. All serum machine learning models underwent 5-fold cross validation on Non-NAFLD and NAFLD-cirrhosis probands before being tested on NAFLD-cirrhosis first degree relatives.

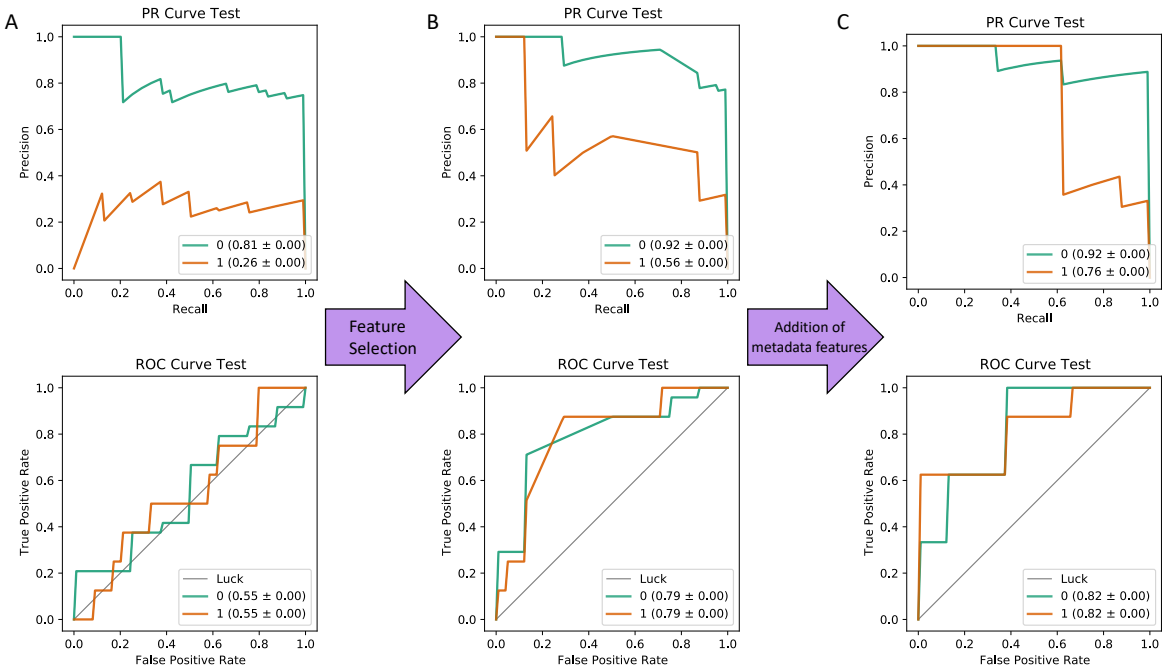

Figure 4: Fecal metabolome machine learning model improvement. A. The PR and ROC test scores for the base model built off all fecal metabolites (985 total features). **B.** A similar plot showing the PR and ROC test scores following feature selection. The model shown was built off the Songbird Top25+Bot 25 AF associated features (50 total features). **C.** The final best performing fecal model that was built off the Songbird Top25+Bot25 features as well as the metadata features BMI, age and sex. All fecal machine learning models underwent 5-fold cross validation on Non-NAFLD and NAFLD-cirrhosis probands before being tested on NAFLD-cirrhosis first degree relatives.

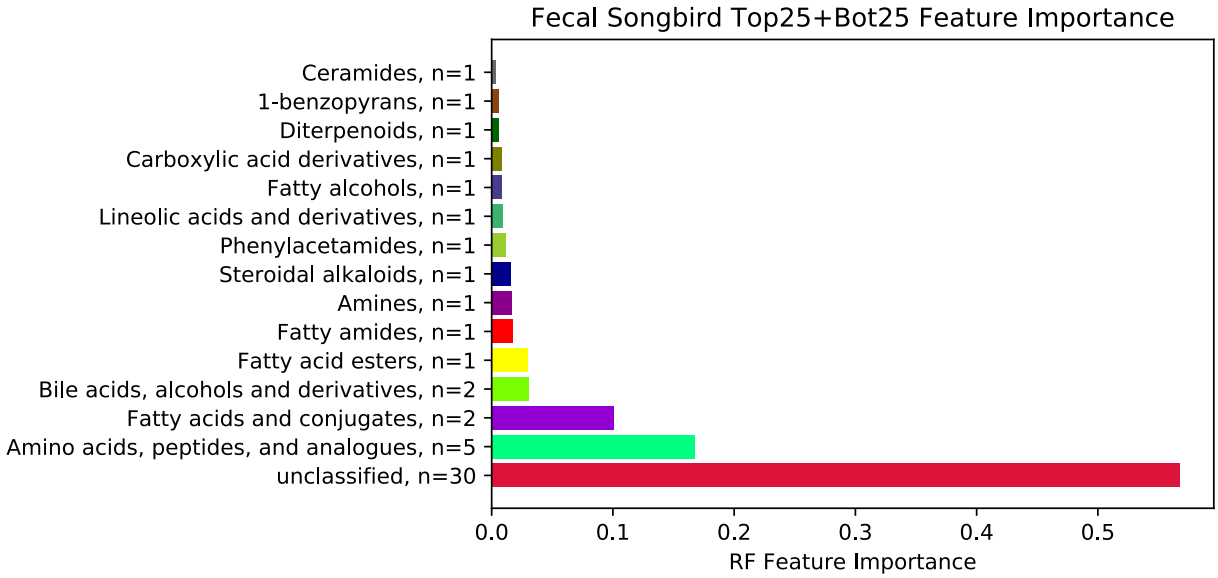

Figure 5: Feature importance of fecal Songbird Top25+Bot25 features. Feature importance values were taken from the RF model built off the fecal Songbird Top25+Bot25 features. Metabolites are viewed at the subclass level. Metabolite classification provided by qemistree fingerprints (24).

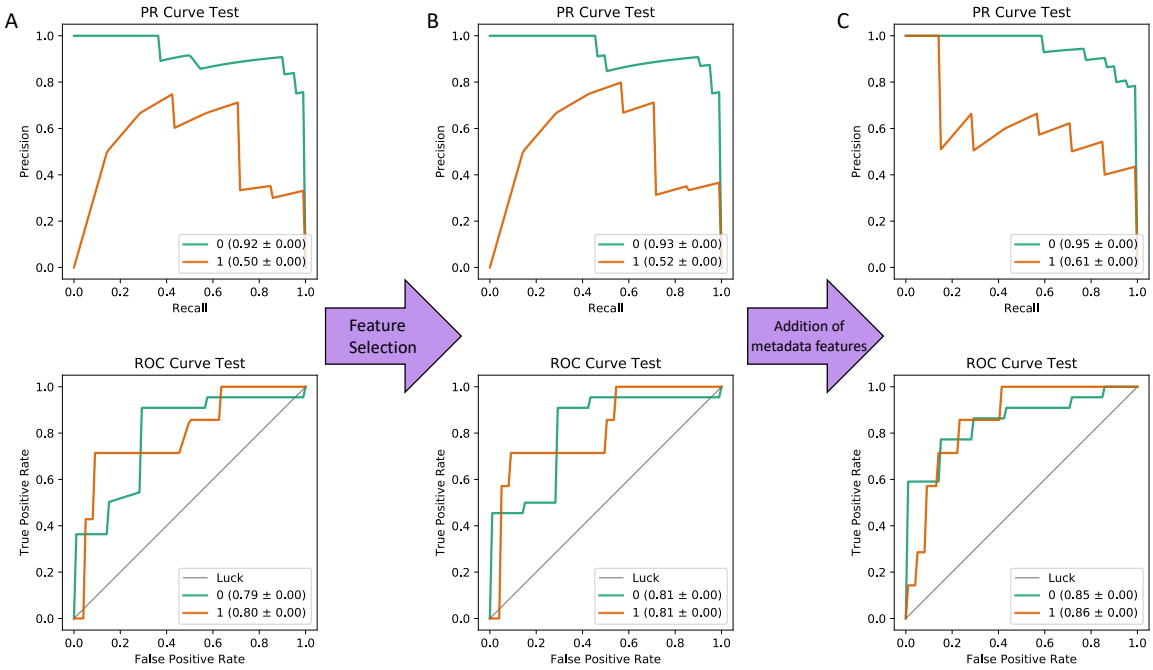

Figure 6: Multi-omic (metabolome and microbiome) machine learning model improvement. A. The PR and ROC test scores for the base model built off the top performing fecal metabolome model (Songbird Top25+Bot25) and the top 27 predictive microbial features (77 total features). **B.** A similar plot showing the PR and ROC test scores following feature selection. The model shown was built off the selected features that had a feature importance >0.005 (27 microbial features, 14 fecal metabolites). **C.** The final best performing multi-omic model that was built off the selected features as well as the metadata features BMI, age and sex. All multi-omic machine learning models underwent 5-fold cross validation on Non-NAFLD and NAFLD-cirrhosis probands before being tested on NAFLD-cirrhosis first degree relatives.

TABLES

Table 1: Demographics for the serum metabolome data set. This table provides the overall counts for the samples used in the serum diversity, composition and machine learning analysis. Samples are split into groups based on disease state. We include sex, BMI and age counts as these are the metadata features we used in our machine learning models.

Serum Metabolome Data Demographic				
Disease Stage	**Group Counts**	**Sex**	**BMI**	**Age**
Non-NAFLD	Probands = 55	Female = 70	Underweight = 2	18-29 = 36
			Normal = 51	30s = 7
			Overweight = 28	40s = 1
	1° Relatives = 42	Male = 27	Obese = 16	50s = 19
			Not Provided = 0	>60s = 34
NAFLD without AF	Probands = 19	Female = 21	Underweight = 0	18-29 = 2
			Normal = 7	30s = 3
			Overweight = 12	40s = 9
	1° Relatives = 15	Male = 13	Obese = 15	50s = 8
			Not Provided = 0	>60s = 12
NAFLD-cirrhosis	Probands = 25	Female = 45	Underweight = 0	18-29 = 3
			Normal = 5	30s = 10
			Overweight = 18	40s = 10
	1° Relatives = 34	Male = 14	Obese = 35	50s = 7
			Not Provided = 1	>60s = 29

Table 2: Demographics for the fecal metabolome data set. This table provides the overall counts for the samples used in the fecal diversity, composition and machine learning analysis. Samples are split into groups based on disease state. We include sex, BMI and age counts as these are the metadata features we used in our machine learning models.

Fecal Metabolome Data Demographic				
Disease Stage	**Group Counts**	**Sex**	**BMI**	**Age**
Non-NAFLD	Probands = 56	Female = 71	Underweight = 2	18-29 = 36
			Normal = 51	30s = 7
	1° Relatives = 42	Male = 27	Overweight = 27	40s = 1
			Obese = 18	50s = 19
			Not Provided = 0	>60s = 35
NAFLD without AF	Probands = 18	Female = 20	Underweight = 0	18-29 = 4
			Normal = 7	30s = 3
	1° Relatives = 17	Male = 15	Overweight = 13	40s = 9
			Obese = 15	50s = 8
			Not Provided = 0	>60s = 11
NAFLD-cirrhosis	Probands = 24	Female = 44	Underweight = 0	18-29 = 3
			Normal = 5	30s = 10
	1° Relatives = 34	Male = 14	Overweight = 17	40s = 10
			Obese = 35	50s = 7
			Not Provided = 1	>60s = 28

Table 3: Demographics for the multi-omic data set. This table provides the overall counts for the samples included in the multi-omic machine learning analysis. We include sex, BMI and age because we used these metadata features in our machine learning models.

Multi-Omic Data Demographic				
Disease Stage	**Group Counts**	**Sex**	**BMI**	**Age**
Non-NAFLD	Probands = 54	Female = 69	Underweight = 2	18-29 = 36
			Normal = 51	30s = 7
	1° Relatives = 42	Male = 27	Overweight = 27	40s = 1
			Obese = 16	50s = 18
			Not Provided = 0	>60s = 34
NAFLD without AF	Probands = 18	Female = 20	Underweight = 0	18-29 = 2
			Normal = 7	30s = 3
	1° Relatives = 15	Male = 13	Overweight = 11	40s = 9
			Obese = 15	50s = 8
			Not Provided = 0	>60s = 11
NAFLD-cirrhosis	Probands = 25	Female = 45	Underweight = 0	18-29 = 3
			Normal = 5	30s = 10
	1° Relatives = 34	Male = 14	Overweight = 18	40s = 10
			Obese = 35	50s = 7
			Not Provided = 1	>60s = 29

Table 4: Serum machine learning model performances. All serum models were cross-validated and tested and the PR and ROC scores were recorded for both. Cross-validation was performed on non-NAFLD (AF negative) and NAFLD-cirrhosis (AF positive) probands. Testing was performed on NAFLD-cirrhosis first degree relatives (mix of AF positive and negative). PR and ROC scores are the AUC values for each metric. In the case where the AF positive (1) and negative (0) classification AUCs were different, we included both. Feature selection models used a feature importance cutoff of >0.005, while further selection models used a feature importance cutoff of >0.02.

Serum Machine Learning Model Performances				
Serum Models	**Test Stage**	**Metric**	**Random Forest**	**Logistic Regression**
Base Model (1847 Features)	Cross Validation	ROC	0.98	0.98
		PR	0.96	0.91
	Test	ROC	0.58	0.71
		PR	0.82 (0), 0.29 (1)	0.87 (0), 0.54 (1)
Feature Selection (44 Features)	Cross Validation	ROC	0.98	0.99
		PR	0.99	0.98
	Test	ROC	0.59	0.55
		PR	0.82 (0), 0.28 (1)	0.76 (0), 0.46 (1)
Further Selection (20 Features)	Cross Validation	ROC	0.98	0.98
		PR	0.98	0.97
	Test	ROC	0.62	0.53
		PR	0.84 (0), 0.37 (1)	0.74 (0), 0.37 (1)
Bio Relevant Base Model (179 Features)	Cross Validation	ROC	0.96	0.97
		PR	0.97	0.88
	Test	ROC	0.62	0.54
		PR	0.85 (0), 0.39 (1)	0.79 (0), 0.30 (1)
Bio Feature Selection (41 Features)	Cross Validation	ROC	0.98	0.98
		PR	0.99	0.98
	Test	ROC	0.66	0.62
		PR	0.86 (0), 0.32 (1)	0.82 (0), 0.49 (1)
Bio Further Selection (17 Features)	Cross Validation	ROC	0.93	0.92
		PR	0.88	0.92
	Test	ROC	0.53	0.58
		PR	0.77 (0), 0.31 (1)	0.77 (0), 0.44 (1)
SB Top 50 Features (50 Features)	Cross Validation	ROC	0.92	0.96
		PR	0.88	0.93
	Test	ROC	0.48	0.70
		PR	0.73 (0), 0.24 (1)	0.82 (0), 0.66 (1)
SB Top 25 + Bot 25 Features (50 Features)	Cross Validation	ROC	0.93	0.95
		PR	0.89	0.90
	Test	ROC	0.54	0.72
		PR	0.78 (0), 0.26 (1)	0.86 (0), 0.67 (1)
SB Bottom 50 Features (50 Features)	Cross Validation	ROC	0.95	0.92
		PR	0.86	0.79
	Test	ROC	0.3	0.54
		PR	0.70 (0), 0.18 (1)	0.74 (0), 0.30 (1)
Further Sel + Metadata (23 Features)	Cross Validation	ROC	0.99	0.98
		PR	0.99	0.90
	Test	ROC	0.64	0.61
		PR	0.86 (0), 0.32 (1)	0.75 (0), 0.43 (1)

Table 5: Fecal machine learning model performances. All fecal models were cross-validated and tested and the PR and ROC scores were recorded for both. Cross-validation was performed on non-NAFLD (AF negative) and NAFLD-cirrhosis (AF positive) probands. Testing was performed on NAFLD-cirrhosis first degree relatives (mix of AF positive and negative). PR and ROC scores are the AUC values for each metric. In the case where the AF positive (1) and negative (0) classification AUCs were different, we included both. Feature selection models used a feature importance cutoff of >0.005, while further selection models used a feature importance cutoff of >0.02.

Fecal Machine Learning Model Performances				
Fecal Models	**Test Stage**	**Metric**	**Random Forest**	**Logistic Regression**
Base Model (985 Features)	Cross Validation	ROC	0.77	0.81
		PR	0.66	0.67
	Test	ROC	0.55	0.55
		PR	0.81 (0), 0.26 (1)	0.80 (0), 0.37 (1)
Feature Selection (45 Features)	Cross Validation	ROC	0.9	0.91
		PR	0.83	0.81
	Test	ROC	0.56	0.51
		PR	0.82 (0), 0.27 (1)	0.71 (0), 0.41 (1)
Further Selection (22 Features)	Cross Validation	ROC	0.93	0.83
		PR	0.84	0.75
	Test	ROC	0.52	0.47
		PR	0.79 (0), 0.24 (1)	0.70 (0), 0.33 (1)
Bio Relevant Base Model (91 Features)	Cross Validation	ROC	0.79	0.72
		PR	0.71	0.58
	Test	ROC	0.53	0.65
		PR	0.83 (0), 0.24 (1)	0.88 (0). 0.29 (1)
Bio Feature Selection (46 Features)	Cross Validation	ROC	0.81	0.75
		PR	0.72	0.66
	Test	ROC	0.52	0.47
		PR	0.82 (0), 0.23 (1)	0.78 (0), 0.20 (1)
Bio Further Selection (23 Features)	Cross Validation	ROC	0.84	0.82
		PR	0.77	0.71
	Test	ROC	0.54	0.57
		PR	0.84 (0), 0.24 (1)	0.85 (0), 0.25 (1)
SB Top 50 Features (50 Features)	Cross Validation	ROC	0.70	0.73
		PR	0.49	0.50
	Test	ROC	0.76	0.55
		PR	0.91 (0), 0.54 (1)	0.76 (0), 0.46 (1)
SB Top + Bot 25 Features (50 Features)	Cross Validation	ROC	0.64	0.78
		PR	0.44	0.52
	Test	ROC	0.79	0.68
		PR	0.92 (0), 0.56 (1)	0.87 (0), 0.51 (1)
SB Bottom 50 Features (50 Features)	Cross Validation	ROC	0.39	0.60
		PR	0.31	0.33
	Test	ROC	0.36	0.59
		PR	0.74 (0), 0.18 (1)	0.86 (0), 0.26 (1)
SB Top25 + Bottom 25 + Metadata	Cross Validation	ROC	0.72	0.85
		PR	0.56	0.62
	Test	ROC	0.82	0.84
		PR	0.92 (0), 0.76 (1)	0.93 (0), 0.72 (1)

Table 6: Multi-Omic machine learning model performances. All multi-omic models were cross-validated and tested and the PR and ROC scores were recorded for both. Cross-validation was performed on non-NAFLD (AF negative) and NAFLD-cirrhosis (AF positive) probands. Testing was performed on NAFLD-cirrhosis first degree relatives (mix of AF positive and negative). PR and ROC scores are the AUC values for each metric. In the case where the AF positive (1) and negative (0) classification AUCs were different, we included both. Selected models used a feature importance cutoff of >0.005.

Multi-Omic Machine Learning Model Performances				
Multi-Omic Models	**Test Stage**	**Metric**	**Random Forest**	**Logistic Regression**
Full Multi-Omic (77 Features)	Cross Validation	ROC	0.76	0.66
		PR	0.87	0.58
	Test	ROC	0.8	0.62
		PR	0.92 (0), 0.50 (1)	0.83 (0), 0.42(1)
Selected Multi-Omic (41 Features)	Cross Validation	ROC	0.79	0.69
		PR	0.89	0.66
	Test	ROC	0.81	0.69
		PR	0.93 (0), 0.52 (1)	0.83 (0), 0.55 (1)
Full Multi-Omic+Metadata (80 Features)	Cross Validation	ROC	0.85	0.62
		PR	0.95	0.53
	Test	ROC	0.86	0.58
		PR	0.95 (0), 0.59 (1)	0.83 (0), 0.42(1)
Sel Multi-Omic+Metadata (44 Features)	Cross Validation	ROC	0.84	0.70
		PR	0.94	0.53
	Test	ROC	0.86	0.66
		PR	0.95 (0), 0.61 (1)	0.82 (0), 0.48 (1)

This thesis is coauthored with Groth, Tobin; Knight, Rob; Dorrestein, Pieter; Tripathi, Anupriya; Fogelson, Kelly; Rahman, Gibraan. The thesis author was the primary author of this material.

References

1. Vernon, G., Baranova, A. and Younossi, Z.M. (2011), Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Alimentary Pharmacology & Therapeutics, 34: 274-285. https://doi.org/10.1111/j.1365-2036.2011.04724.x

2. Ruissen, M., Mak, A., Beuers, U., Tushuizen, M., & Holleboom, A. (2020). Non-alcoholic fatty liver disease: a multidisciplinary approach towards a cardiometabolic liver disease, *European Journal of Endocrinology*, *183*(3), R57-R73. Retrieved Dec 15, 2020, from https://eje.bioscientifica.com/view/journals/eje/183/3/EJE-20-0065.xml

3. Brunt, E. M., Wong, V. W., Nobili, V., Day, C. P., Sookoian, S., Maher, J. J., Bugianesi, E., Sirlin, C. B., Neuschwander-Tetri, B. A., & Rinella, M. E. (2015). Nonalcoholic fatty liver disease. Nature reviews. Disease primers, 1, 15080. https://doi.org/10.1038/nrdp.2015.80

4. Golabi, P., Fazel, S., Otgonsuren, M., Sayiner, M., Locklear, C. T., & Younossi, Z. M. (2017). Mortality assessment of patients with hepatocellular carcinoma according to underlying disease and treatment modalities. Medicine, 96(9), e5904. https://doi.org/10.1097/MD.0000000000005904

5. Huang, J. F., Hsieh, M. Y., Dai, C. Y., Hou, N. J., Lee, L. P., Lin, Z. Y., Chen, S. C., Wang, L. Y., Hsieh, M. Y., Chang, W. Y., Yu, M. L., & Chuang, W. L. (2007). The incidence and risks of liver biopsy in non-cirrhotic patients: An evaluation of 3806 biopsies. Gut, 56(5), 736–737. https://doi.org/10.1136/gut.2006.115410

6. Caussy, C., Tripathi, A., Humphrey, G., Bassirian, S., Singh, S., Faulkner, C., Bettencourt, R., Rizo, E., Richards, L., Xu, Z. Z., Downes, M. R., Evans, R. M., Brenner, D. A., Sirlin, C. B., Knight, R., Loomba, R. (2019). A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. Nature Communications, 10(1), 1406. https://doi.org/10.1038/s41467-019-09455-9

7. Visconti, A., Le Roy, C. I., Rosa, F., Rossi, N., Martin, T. C., Mohney, R. P., Li, W., de Rinaldis, E., Bell, J. T., Venter, C. J., Nelson, K. E., Spector, T. D., Falchi, M. (2019). Interplay between the human gut microbiome and host metabolism. Nature Communications, 10(1), 4505. https://doi.org/10.1038/s41467-019-12476-z

8. Oh, T. G., Kim, S. M., Caussy, C., Fu, T., Guo, J., Bassirian, S., Singh, S., Madamba, E. V., Bettencourt, R., Richards, L., Yu, R. T., Atkins, A. R., Huan, T., Brenner, D. A., Sirlin, C. B., Downes, M., Evans, R. M., Loomba, R. (2020). A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. Cell Metabolism, 32(5), 878-888.e6. https://doi.org/https://doi.org/10.1016/j.cmet.2020.06.005

9. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *The ISME journal*, *5*(2), 169–172. https://doi.org/10.1038/ismej.2010.133

10. Chiang, J., & Ferrell, J. M. (2018). Bile Acid Metabolism in Liver Pathobiology. *Gene expression*, *18*(2), 71–87. https://doi.org/10.3727/105221618X15156018385515

11. van der Veen, J. N., Kennelly, J. P., Wan, S., Vance, J. E., Vance, D. E., & Jacobs, R. L. (2017). The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1859*(9, Part B), 1558–1572. https://doi.org/https://doi.org/10.1016/j.bbamem.2017.04.006

12. Berk, P. D., Zhou, S., & Bradbury, M. W. (2005). Increased hepatocellular uptake of long chain fatty acids occurs by different mechanisms in fatty livers due to obesity or excess ethanol use, contributing to development of steatohepatitis in both settings. *Transactions of the American Clinical and Climatological Association*, *116*, 335–345.

13. Pedregosa, F., and Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Retrieved December 15, 2020, from https://dl.acm.org/doi/10.5555/1953048.2078195

14. Ridlon, J. M., Alves, J. M., Hylemon, P. B., & Bajaj, J. S. (2013). Cirrhosis, bile acids and gut microbiota. *Gut Microbes*, *4*(5), 382–387. https://doi.org/10.4161/gmic.25723

15. Caussy, C., Ajmera, V. H., Puri, P., Hsu, C. L., Bassirian, S., Mgdsyan, M., Singh, S., Faulkner, C., Valasek, M. A., Rizo, E., Richards, L., Brenner, D. A., Sirlin, C. B., Sanyal, A. J., Loomba, R. (2019). Serum metabolites detect the presence of advanced fibrosis in derivation and validation cohorts of patients with non-alcoholic fatty liver disease. *Gut*, *68*(10), 1884 LP – 1892. https://doi.org/10.1136/gutjnl-2018-317584

16. Wieland, A., Frank, D. N., Harnke, B., & Bambha, K. (2015). Systematic review: microbial dysbiosis and nonalcoholic fatty liver disease. *Alimentary pharmacology & therapeutics*, *42*(9), 1051–1063. https://doi.org/10.1111/apt.13376

17. Martinez, K. B., Leone, V., & Chang, E. B. (2017). Microbial metabolites in health and disease: Navigating the unknown in search of function. *The Journal of biological chemistry*, *292*(21), 8553–8559. https://doi.org/10.1074/jbc.R116.752899

18. Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Goethem, M. W. V., Northern, T. R., Vazquez-Baeza, Y., Wang, M., Bokulich, N. A., Watters, A. Song, S. J., Bonneau, R., Dorrestein, P. C., Knight, R. (2019). Learning representations of microbe–metabolite interactions. *Nature Methods*, *16*(12), 1306–1314. https://doi.org/10.1038/s41592-019-0616-3

19. Wehrens, R., Hageman, J. A., van Eeuwijk, F., Kooke, R., Flood, P. J., Wijnker, E., Keurentjes, J. J.B., Lommen, A., van E., Henriëtte D.L.M., Hall, R. D., Mumm, R., de Vos, Ric C.H. (2016). Improved batch correction in untargeted MS-based metabolomics. *Metabolomics*, *12*(5), 88. https://doi.org/10.1007/s11306-016-1015-8

20. Hicks, S. C., Okrah, K., Paulson, J. N., Quackenbush, J., Irizarry, R. A., & Bravo, H. C. (2018). Smooth quantile normalization. *Biostatistics (Oxford, England)*, *19*(2), 185–198. https://doi.org/10.1093/biostatistics/kxx028

21. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T. Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

22. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

23. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R.J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C. Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

24. Tripathi, A., Vázquez-Baeza, Y., Gauglitz, J. M., Wang, M., Dührkop, K., Nothias-Esposito, M., Acharya, D. D., Ernst, M., van der Hooft, J. J. J., Zhu, Q., McDonald, D., Brejnrod, A. D., Gonzalez, A., Handelsman, J., Fleischauer, M., Ludwig, M., Böcker, S., Nothias, L. F., Knight, R., Dorrestein, P. C. (2020). Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nature Chemical Biology*. https://doi.org/10.1038/s41589-020-00677-3

25. Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*, *10*(1), 2719. https://doi.org/10.1038/s41467-019-10656-5