UNIVERSITY OF CALIFORNIA SAN DIEGO

Benchmarking continuous phenotype prediction with multi-omic microbiome data

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

in

Biology

by

Patrick Imran McGrath

Committee in charge:

Professor Andrew Bartko, Chair
Professor Rachel Dutton, Co-Chair
Professor Sara Jackrel
Professor Rob Knight

2021

The thesis of Patrick Imran McGrath is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

Thesis Approval Page iii

Table of Contents iv

List of Figures v

Acknowledgements vi

Abstract of the Thesis vii

Chapter 1 Introduction 1
1.1 Microbiome -omics data 1
1.2 Predictive models on the microbiome 2
1.3 Aims 3

Chapter 2 Methods 5
2.1 Preprocessing 5
2.2 Unit Benchmark 5
2.3 Procrustes analysis 6

Chapter 3 Results 7
3.1 Cohort-specific effects biasing predictive performance 7
3.2 Performance by algorithm across hyperparameter space 8
3.3 Hyperparameter transferability 9

Chapter 4 Discussion 15
4.1 Current Limitations 16
4.2 Future Work 18

Appendix A Code availability 19

Bibliography 20

LIST OF FIGURES

Figure 3.1: Cohort-specific effects bias predictive performance 11
Figure 3.2: Performance by algorithm across hyperparameter search space, relative to the null model 12
Figure 3.3: Transferability of hyperparameters between datasets and between 16S and metagenomics data 13
Figure 3.4: Time savings of hyperparameter transfer 14

ACKNOWLEDGEMENTS

I would like to acknowledge the ML Benchmarking team of Yoshiki Vázquez-Baeza, Shi Huang, and George Armstrong for their immense collaboration in this work. I would also like to acknowledge the Center for Microbiome Innovation and IBM for continuing to support my studies and research, and Andrew Bartko for being my thesis advisor.

ABSTRACT OF THE THESIS

Benchmarking continuous phenotype prediction with multi-omic microbiome data

by

Patrick Imran McGrath

Master of Science in Biology

University of California San Diego, 2021

Professor Andrew Bartko, Chair
Professor Rachel Dutton, Co-Chair

Large-scale microbiome datasets from 16S amplicon sequencing provide opportunities for building predictive models with supervised machine learning to answer questions of biological significance. Prior regression analyses have used supervised learning to predict variables of the sampled microbial environment, such as pH, age, or other host phenotypes and disease states, however little justification has been made for the use of specific algorithms on microbiome data. We performed a large-scale comprehensive benchmark for 12 regression algorithms across an extensive grid search for tuning algorithm hyperparameters. We compared these algorithms in predicting age and BMI with three large human datasets: The National FINRISK Study, the Study of Latinos, and the International Multiple Sclerosis Microbiome Study. Each dataset is comprised of matched samples with both 16S amplicon and shotgun metagenomic sequencing. We found that cohort effects bias predictive performance, and propose a standardized metric for comparing models across cohorts and predictive targets. After standardization, boosting and ensemble-type algorithms are both highly accurate and robust relative to other algorithms in all our cohorts, and require less thorough hyperparameter tuning to produce a good model. Hyperparameters may be transferred from models trained on 16S data to train a model on similarly-prepared metagenomics data, or vice-versa. In the aforementioned reliable ensemble and boosting algorithms, this comes at no loss of relative performance and can save time when used as a warm start to hyperparameter tuning. We recommend these practices when producing predictive models for the microbiome with intent to compare across different cohorts.

Chapter 1

Introduction

1.1 Microbiome -omics data

The volume of biological data has increased rapidly in the past decade driven by increased availability and affordability of genome sequencing [1]. The revolution in high throughput sequencing, DNA sequencing which is capable of sequencing very many DNA sequences at once, has given biologists a wealth of information [2]. Similar increases in data volume, be it in image collection and annotation, user data, or other areas of biological research using high-throughput technologies, have led to an increase in the relevance and utilization of machine learning analysis [3]. In the microbiome field, public repositories for online analysis such as Qiita have increased the availability and access to microbiome datasets [4]. This has made the data available to biologists, machine learning researchers, and students or the general public.

This wealth of datasets has largely been driven by the novel and often unexpected ways in which the human microbiome has been implicated in a variety of environmental and clinical phenomena in human health. For example, studies have linked the microbiome to areas such as celiac's disease and diet changes, as well as more unexpected areas, such as gastric cancer, circadian rhythms, and PTSD [5]. Prior research has discovered links between the gut microbiome and human diseases such as obesity and inflammatory bowel disease [6][7]. There is also much excitement about the gut-brain axis, a reported association between the gut microbiome and mental health conditions [8].

To study these relationships, a variety of data can be generated from microbiome samples depending on sequencing technology [9]. One such data preparation, 16S amplicon sequencing, is a method for identifying taxa present in a sample. While this method is considered a ‘classical’ analysis and has been in use for decades with older qPCR sequencing, it has enjoyed a renaissance in use thanks to improved analytical methods and sequencing methods which reduce cost and increase throughput [9]. However, other methods are gaining in popularity, such as shotgun metagenomics, which sequences all available DNA material in a sample and can thus not only identify taxa present in a sample, but also what genes are present [10]. Both these preparations produce tables of samples by the relative abundances of microbial taxa found in each sample. These relative abundance counts are often very sparse, and the number of unique taxa identified is very large, resulting in uniquely difficult datasets for traditional statistical analyses [11].

1.2 Predictive models on the microbiome

Given these links between health and the microbiome, many microbiome studies aim to predict clinical outcomes or classify samples based on these relative abundance tables from the 16S sequencing preparation [12]. These studies, which aim to predict outcomes such as human aging, post-mortem intervals (the time since death of a decaying corpse), or IDB sensitivity, through employing machine learning algorithms with varying levels of success [13]. Machine learning advances in optimizing algorithms, and in the development of open source libraries, have made robust and reliable machine learning tools available to researchers, students, and even citizen-scientists [14]. Machine learning algorithms commonly used in these microbiome studies include classes of algorithms such as decision tree-based ensembles (Random Forests), support vector machines, and k-nearest neighbor algorithms with distance metrics relevant to biodiversity [15]. These studies have been typically classification problems, where the model attempts to predict a condition, which may be "healthy" or "symptomatic" in, for example, a study of cirrhosis [16]. However, many questions of biological interest such as predicting age are regression problems, where the outcome is a continuous variable [17].

To aid researchers in building accurate predictive models, prior studies have conducted benchmarking that compares machine learning algorithms' performance on microbiome data. Two necessary steps in developing a robust and accurate machine learning model are choosing the right algorithm for your problem, and tuning other input variables of the algorithm [15]. These input variables, known as hyperparameters, dictate the architecture of the model and how it learns. Optimal values for hyperparameters cannot be inferred, so they are typically tuned by training multiple models on different values for each hyperparameter. This tuning is a necessary step in applying machine learning algorithms as it impacts the accuracy and generalizability of the model. It is notable that a full characterization of these models during hyperparameter tuning is absent from prior studies. There are few studies which have large variety of machine learning algorithms that may be applied to microbiome data, and their performance across the hyperparameter tuning process [18] [17].

1.3 Aims

This study aims to benchmark supervised machine learning algorithms, available in public toolkits (including but not limited to Scikit-Learn) on three large, highly controlled microbiome datasets of human samples. The first of the three datasets is from The National FINRISK study, a population survey of risk factors in the Finnish population, and it totals over 6400 microbiome samples [19]. The second dataset, the Hispanic Community Health Study / Study of Latinos [20] is a similar epidemiologic study in Hispanic and Latino populations in the United States, with over

1,100 microbiome samples. Our third dataset comes from the International Multiple Sclerosis Microbiome Study (iMSMS) and currently has over 1,100 microbiome samples [21]. The large sample sizes of these studies reduces the risk of underfitting machine learning models, as more training data is available. Each dataset also offers some common sample metadata, including continuous variables such as the host's age and BMI. Most critically, in each of these studies, samples are sequenced with both 16S amplicon and shotgun metagenomics sequencing giving true biological replicates between the two sequencing preparations. We assess twelve algorithms from three different families of algorithm which share a common architecture or learning principle: Ensemble algorithms, which are made up of multiple weaker models, Boosting algorithms, which iteratively add weaker models to their ensemble with a gradient boosting technique, and Linear algorithms, which are variants on linear regression. We evaluate the performance of each algorithm relative to each other and to a null model in terms of prediction error. Our scope is unique in that we explicitly compare the performance behavior of these algorithms for predicting continuous variables across both 16S and metagenomics data preparations, with the same predictive targets in each of three population-scale datasets. We aim to resolve the population effects impacting comparisons of models between different studies. We also seek to characterize the performance of regression algorithms in the hyperparameter tuning process, identifying robust algorithms and those which require more careful consideration. Lastly we provide recommendations for training models between multiple cohorts and the possibility of transferring hyperparameters with minimal relative performance impact.

Chapter 2

Methods

2.1 Preprocessing

For each target phenotype and data preparation (16S and metagenomics), identical quality control and preprocessing steps are carried out. We first filter samples containing missing values in the target phenotype, and then filter low-abundance features with fewer than ten counts. Feature tables are rarefied to a uniform sampling depth of 1000.

2.2 Unit Benchmark

To carry out this benchmarking at scale, we developed a source-controlled and unit-tested plugin “Q2-MLAB” in the QIIME2 framework [22]. Each algorithm is either available in Scikit-Learn 0.24.1 or implements a compatible python API [14]. For each algorithm, multiple values for each of its hyperparameters were defined. We iteratively search through hyperparameter value combinations, where one model is trained per hyperparameter combination. For algorithms where the total number of valid combinations exceeds 1000, we limited our search space to a random 1000 hyperparameter combinations. For algorithms with fewer than 1000 combinations, we created models for the full hyperparameter search space. One model is trained on one hyperparameter set in each cohort: dataset, predictive target, and data preparation. To measure how well models generalize to unseen data, we train each model using cross validation - a technique for training and evaluating a model on different subsets or "folds" of the input data. Each model (one hyperparameter set) undergoes 5-fold stratified cross validation, repeated three times for an effective 15 folds of the data. Mean Absolute Error (MAE) is recorded and averaged across these fifteen folds, and variance in MAE across all 15 folds is also recorded. For each cross validation fold we also record a standardized MAE, calculated as MAE divided by the target phenotype's standard deviation in the training set.

2.3 Procrustes analysis

A Procrustes analysis of disparity allows for comparisons between the shapes of two matrices. For this we conducted our analysis using the Procrustes method in SciPy 1.6.1 [23]. We computed Procrustes disparity on a per-algorithm basis for between-preparation disparity where the rows in the two matrices are matched hyperparameter sets in 16S data and metagenomics data. Each row contains the coordinates of that hyperparameter's performance as mean MAE and variance in MAE. Between-dataset disparity was computed similarly but with matched hyperparameter sets between pairs of our three datasets. A null distribution is calculated by randomly shuffling the row order of the matrices 100 times, computing the disparity on the shuffled matrices, and taking the average disparity.

Chapter 3

Results

3.1 Cohort-specific effects biasing predictive performance

We conducted a benchmarking of 12 machine learning algorithms for predicting host age and BMI in three large datasets: FINRISK, SOL, and iMSMS. We found that differences in predictive performance were due to dataset variability and the scale of the target phenotype. Datasets with higher standard deviation in the target phenotype had higher MAE in predicting that target (Figure 3.1B). To account for this, for each dataset and target phenotype we normalized MAE by the standard deviation of that target phenotype for that dataset. Therefore a standardized MAE of 1 denotes that the average error of the model is one standard deviation. This brought our results into a frame of reference where we could make meaningful comparisons between algorithms and between 16S and metagenomics sequencing. We compare our results to a baseline null model, which when "trained" on the target's mean and standard deviation will return a prediction randomly drawn from a normal distribution around the target phenotype's mean (Figure 3.1C). After correcting for the dataset effects, the average MAE (across all datasets and target phenotypes) was 0.733 standard deviations and the average MAE of the null model was 1.127 standard deviations. While comparisons between studies may lead to certain conclusions, e.g. that the FINRISK dataset produces the most accurate model for BMI prediction, these comparisons need to account for population effects. Raw predictive performance may be limited by the sampled population.

3.2 Performance by algorithm across hyperparameter space

A major computational limitation to building accurate predictive models is the time required to tune hyperparameters, the inputs to a model set by the practitioner. The values for each hyperparameter will vary based on the predictive modeling problem and the dataset used to train the model. While it is impossible to determine "best" or optimal hyperparameters, a grid search technique is often used to approximate this by evaluating all combinations of a range of hyperparameter values. This is known as hyperparameter tuning. A more exhaustive grid search is more likely to find hyperparameters which result in the most accurate predictions, however this computation is expensive and can take from a few hours up to a few weeks.

From benchmarking on our three microbiome datasets, we developed a knowledge base of hyperparameter combinations and their performance on several machine learning algorithms. For each algorithm, we conducted a randomized grid search for hyperparameter tuning, a step that ensures the parameters which define the model's architecture are well-suited for the task and data. In Figure 3.2, we combine results for each algorithm (across all three datasets and two target phenotypes) using our standardized MAE and plot the difference in standardized MAE from the null model. We find that Boosting and Ensemble algorithms LGBMRegressor, AdaBoost, RandomForestRegressor, ExtraTreesRegressor are reliable for all hyperparameter combinations, whereas other Boosting algorithms GradientBoostingRegressor and XGBoost can produce very poor results. Most models from Linear algorithms such as RidgeRegressor and LinearSVR perform worse than null, suggesting careful hyperparameter tuning is required to get meaningful models in these cases.

3.3 Hyperparameter transferability

In each dataset, every sample was sequenced with both 16S marker gene sequencing and shotgun metagenomics (MG) sequencing, giving true biological replicates between these two data preparations. Since each algorithm underwent the same hyperparameter search with both the 16S and MG data, we can compare the relative performance of a hyperparameter combination in one data preparation to another (or likewise from one dataset to another). We model transferability of hyperparameters as the Procrustes disparity between two matrices of hyperparameters by performance metrics. Each matrix is row-indexed by hyperparameter set and for each hyperparameter set we measure the mean MAE and MAE variance. For each dataset and algorithm we use the same set of hyperparameters so these matrices are overlapping. When the disparity is low between two matrices, this indicates that hyperparameters are transferable between data subsets. With each comparison, we also compute a null distribution by shuffling the row-order of the matrices and computing a new disparity. Shuffling the row order breaks the relationship between matched hyperparameter sets in the two matrices. Null model performance (from randomly shuffling labels within each algorithm, breaking the pairings of hyperparameter sets between groups) is approximately 1, or the highest possible disparity.

We find that between-dataset disparity is not consistent across cohorts and target phenotypes for the reliable Ensemble /Boosting algorithms (Figure 3.3A). When looking at between-preparation transferability (16S to metagenomics, and vice versa), we find that Linear algorithms have generally bad transferability whereas Boosting models have generally good transferability (Figure 3.3B). Models that are reliable according to the analysis in Figure 3.2 also have good transferability between preparations. Random Forests, LightGBM, and XGBoost hyperparameters perform similarly well between data preparations, suggesting tuned hyperparameters tuned from one data preparation may be used in another with minimal performance difference (Figure 3.3B).

The main advantages of transferring hyperparameters from 16S to MG data are in hyperparameter tuning time - rather than the full grid search for the MG data, we can transfer the hyperparameters of the best 16S model (in one algorithm) to train a model with the MG data. When done with the same algorithm and assuming perfect transferability, i.e. that the best hyperparameters in one data preparation produces is also the best in the other, this is preferable to redoing a full grid search with the MG data. In Figure 3.4, a full grid search that takes anywhere from a day to over a week, depending on the dataset and target phenotype. However, training a single model takes from a minute to an hour to train. In practice this means half the total hyperparameter tuning runtime, since a full grid search only needs to be done once. In these situations, transferred hyperparameters can provide a “warm-start” for further hyperparameter tuning.

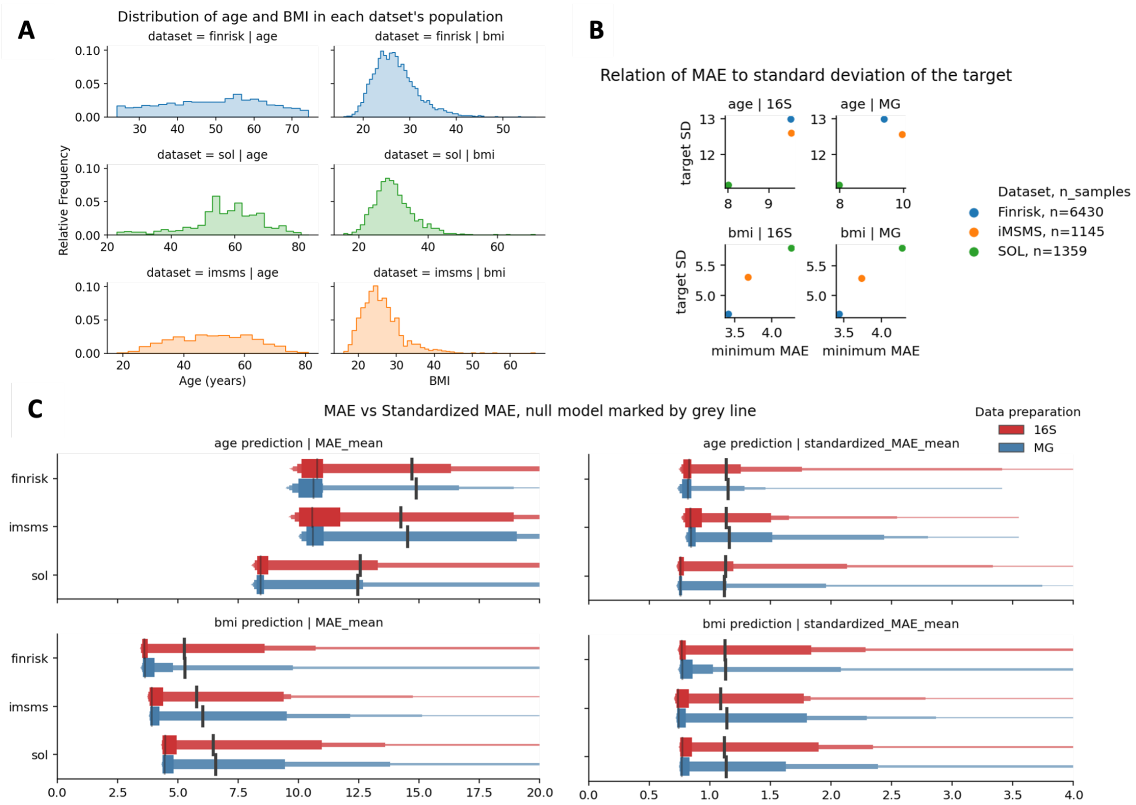

Figure 3.1: **Cohort-specific effects bias predictive performance.** In **3.1A**, we plot a histogram of target phenotype values, namely age and BMI, for our three datasets. Observations for each dataset are normalized independently such that the area of each dataset's bars sum to 1. In **3.1B**, we show the minimum MAE from all models produced with each dataset by target phenotype and data preparation. In the y-axis, standard deviation of the target phenotype is over all observations in that dataset, with the number of observations given in the legend. **3.1C** shows the application of our standardization to account for cohort effects. On the left, predictive performance (in MAE) is dependent on both the dataset and target phenotype. On the right, standardized MAE brings these results into a common frame of reference. These letter-value plots represent the middle 50% of the data in the large center box, the next 25% in the next largest box, 12.5% in the next, and so on, with the height of boxes corresponding to these depths. The median is marked by the thin grey line within the center box, and the performance of a null model is marked by the thick black line.

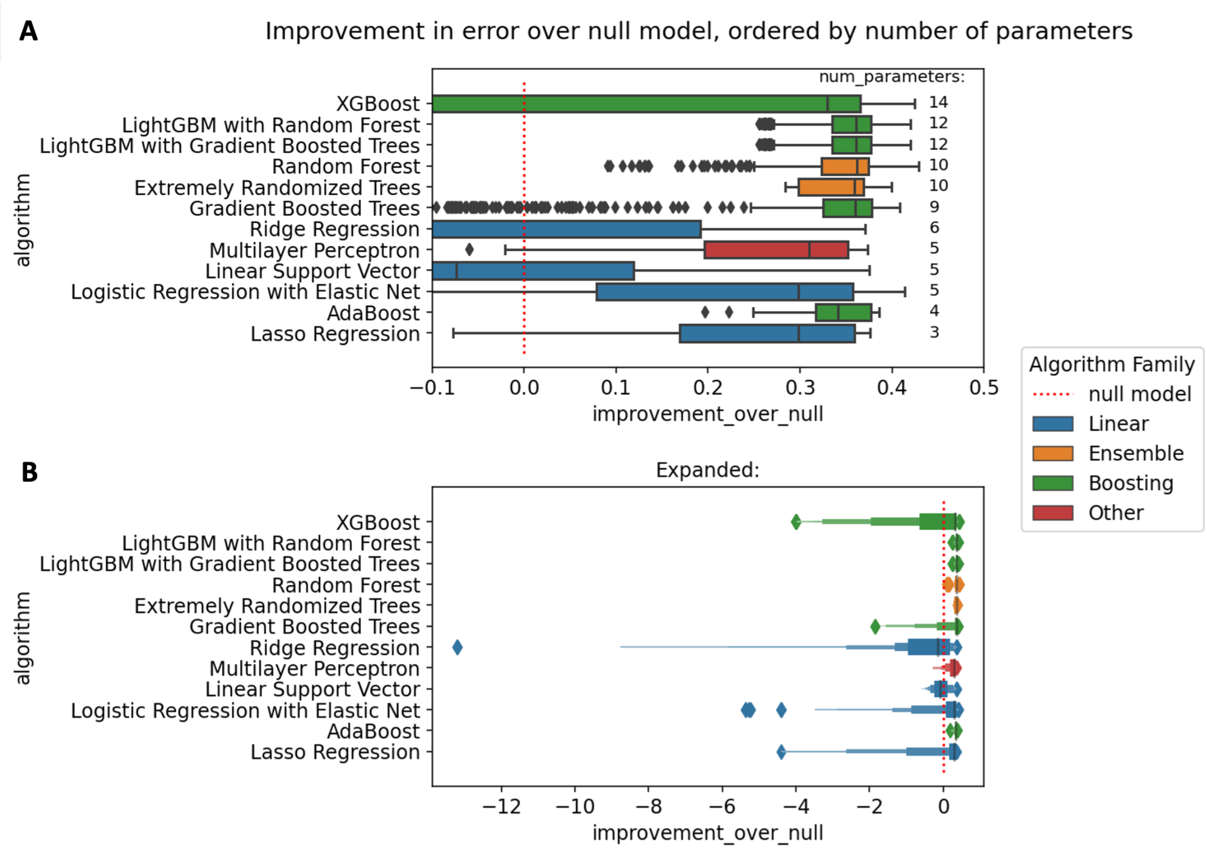

Figure 3.2: **Performance by algorithm across hyperparameter search space, relative to the null model.** In **3.2A**, box plots show the improvement in performance relative to the null model in detail, limiting the x-axis to performance better than the null model (positive values). In **3.2B**, an expanded view of the same results as a letter-value plot displays the long tail of performance worse than the null model exhibited by certain algorithms. Each point in both **3.2A** and **3.2B** represents one model/one hyperparameter combination, and each model's MAE is evaluated 5 cross-validation folds, repeated three times. For each algorithm, models are included from all datasets and target phenotypes. Improvement over the null model is measured as the standardized MAE of the null model minus the standardized MAE of a given model, thus greater is better. Algorithms are ordered by number of tunable hyperparameters (num_parameters) and broken down by algorithm families that share similar architecture.

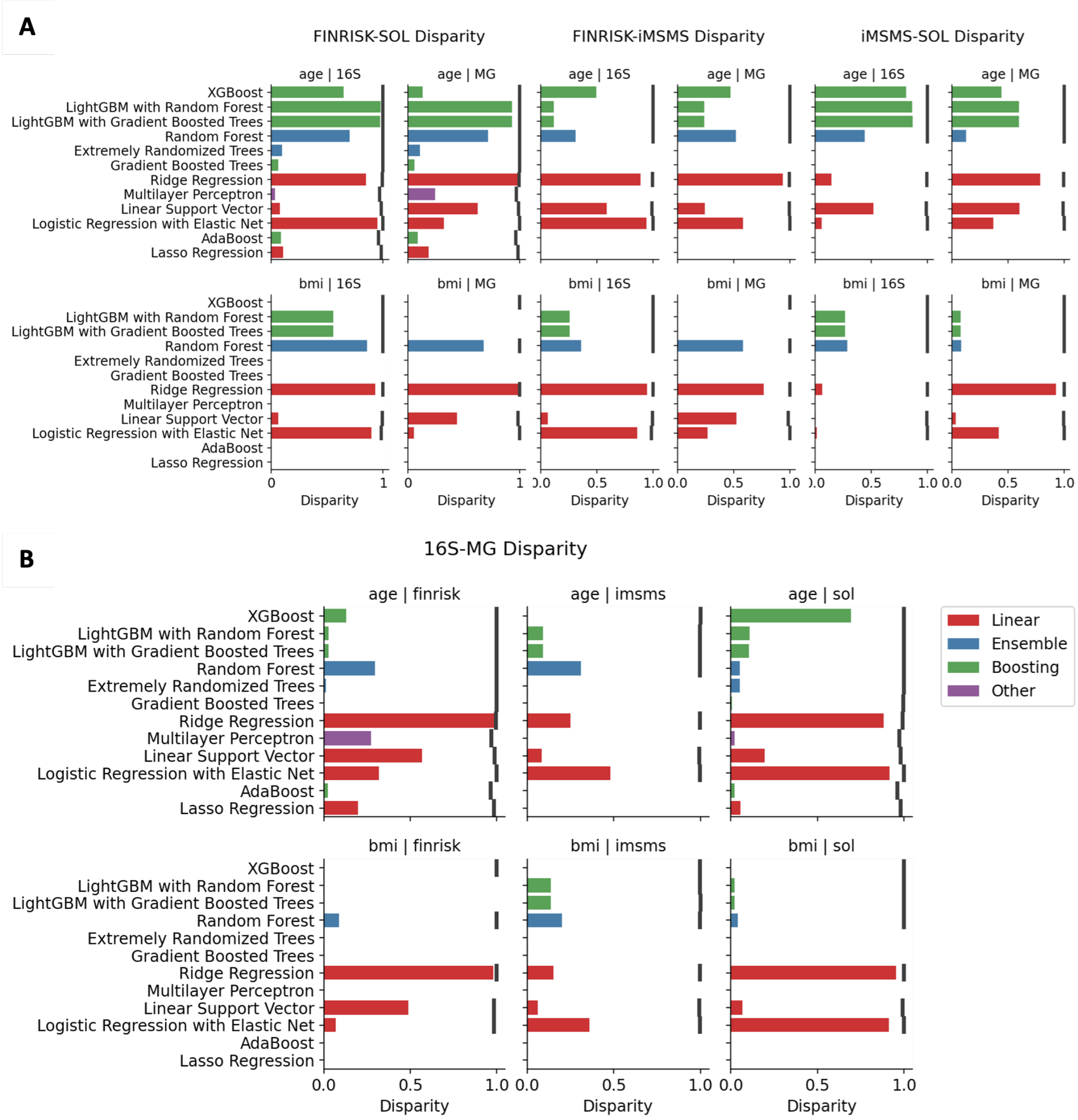

Figure 3.3: **Transferability of hyperparameters between datasets and between 16S and metagenomics.** In **3.3A**, for each algorithm we measure the disparity between dataset pairs, an indicator of how transferable hyperparameters are between those two datasets. Lower disparity corresponds to greater transferability. In **3.3B**, for each algorithm we measure the disparity between 16S and metagenomics data preparations. The black line at the end of each bar is the null distribution. Missing bars indicate one side of the comparison (either a dataset or data preparation) containing missing results due to ongoing computation at time of writing.

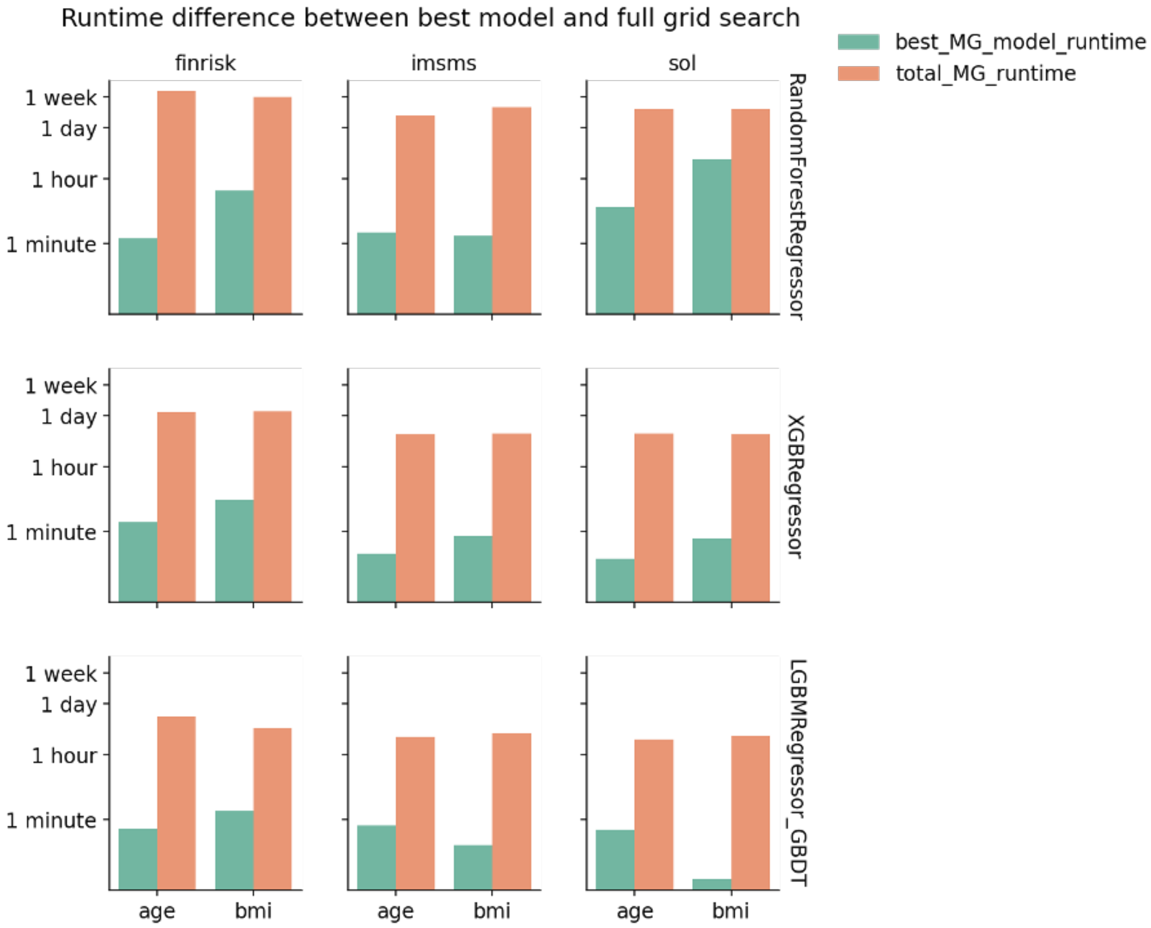

Figure 3.4: **Time savings of hyperparameter transfer.** For algorithms with the most transferable hyperparameters, we plot the runtime of the individual best model (green) with MG data against the total runtime of hyperparameter tuning with metagenomics data (orange) for the same task and cohort. Note the logarithmic scale on the y-axis. The best model was selected as the model with the lowest MAE in all algorithms for that cohort.

Chapter 4

Discussion

With the preponderance of large metagenomic microbiome datasets, and the number growing larger each year, the need for robust and accurate machine learning approaches for predicting clinical or environmental outcomes is ever more pertinent. We sought to recommend hyperparameters, based on empirical performance data from our large-scale benchmarking, that would improve the efficiency and robustness of training such models for microbiome data. Specifically, the machine learning problem which we focused on was regression (prediction of continuous values). We observed that the Ensemble and Boosting family of algorithms, among which are Random Forests and LightGBM, result in the best accuracy when used with microbiome datasets.

Our benchmarking results validate the popular use of Random Forests and other tree-ensemble based algorithms by prior publications, and provide the missing empirical argument supporting their use. For example, we provide empirical justification for the choice of Random Forests in a 2018 study by Belk et al., which used Random Forests to predict the time of death in human and mouse cadavers [24]. Belk et al. cite the robustness of Random Forests to overfitting the data and excellent performance, however this is nonspecific to microbiome data and the authors provide no comparison to other algorithms which can perform the same task. We found that other algorithms such as Gradient Boosted Trees, LightGBM, and XGBoost (all popular Boosting algorithms) performed comparably to Random Forests in the best MAE it achieved, and in some datasets produced greater MAE.

The accuracy benefits and runtime costs of hyperparameter tuning were unstudied in previous benchmarking efforts such as those by Statnikov et al. were. In fact, Statnikov et al. used the default hyperparameter settings, citing the robustness of Random Forests to hyperparameter values. We confirmed the robustness of Random Forests to hyperparameter choice as applied to microbiome data, as visible in the tight range of accuracy consistently above the null model (Figure 3.2A). Notably LightGBM achieves a narrower interquartile range. Intuitions about the ability of Ensemble algorithms to perform well on sparse datasets, as in the case with microbiome data, and their robustness to parameter choice are confirmed by our results. One possible explanation for why some algorithms are more reliable than others may be the number and type of tunable hyperparameters. With Linear algorithms, the hyperparameters that can be tuned have greater influence over the model's performance. For example, a linear regression will be more sensitive to the value of its regularization hyperparameter. In contrast, Random Forests will not be as sensitive to different values for its n_estimators parameter, which dictates the number of decision trees in the ensemble. XGBoost, though it is a Boosting algorithm, has some tunable hyperparameters that one would typically see in a linear regression algorithm such as regularization terms and learning rate. This may contribute to its long tail of performance worse than the null model, a characteristic shared by Linear algorithms.

4.1 Current Limitations

Transferability of hyperparameters may be most useful in settings necessitating a transition from legacy models (built on 16S data) towards newer models built on metagenomics data. As metagenomics sequencing becomes more popular and inexpensive, this need may increase. Though transferring exact hyperparameters may not be the most prudent approach, hyperparameters from one data preparation can be used as a starting point for a limited grid search with another preparation. There may be further ways we can streamline the hyperparameter tuning process with recommended starting values for a hyperparameter grid search. The tendency of hyperparameter sets to have similar performance in different cohorts, which we call transferability, may be indicative of a pattern in hyperparameter values among the best performing models. Some hyperparameter values may be more prevalent in "good" models than in worse models. Values for hyperparameters that the best performing models all share, i.e. "consensus hyperparameters", could be used as recommendations for starting a grid search when information about transferability is unknown. To fulfill the aim of recommending a hyperparameter set, tools exist which seek to automate algorithm selection and hyperparameter tuning for a given dataset and predictive task. These "autoML" tools are useful in many situations and we do not seek to replace them. However, our database of hyperparameter sets and their performance on multiple algorithms, datasets, and target phenotypes can be a valuable resource for understanding how models learn on microbiome data.

Our datasets in this study all consisted of human gut microbiome samples. Though this allowed for direct comparisons between cohorts, we could not explore performance of algorithms in with environmental microbiome data. While human microbiome and environmental microbiome data share statistical properties such as size and sparsity, they can differ greatly in microbial diversity. It would be interesting to see whether the transferability of hyperparameters, or the consistent performance of Ensemble algorithms, is innate to the algorithm of choice or to the input data's sample type. Most consumer and clinical applications of a predictive model would be applied to a human microbiome, such as predicting phenotypes. However, our interactions with our environment shape our human microbiomes, and the environmental microbiome is increasingly of interest in climate studies. As more large datasets of environmental and human studies become available, further benchmarking may be useful to characterize machine learning performance on multiple sample types of microbiome data.

4.2 Future Work

Our work compares the performance of 16S and shotgun metagenomics data under multiple conditions. Despite the added expense of shotgun metagenomics data, we do not see an overall improvement in using microbial features derived from metagenomics data compared to 16S features. Such differences are localised to specific algorithms and cohorts, and further investigation is necessary to understand the mechanisms of those differences. The lack of a global difference in performance between the two data preparations may be due to the identical preprocessing done to both data preparations. Metagenomics data allows for higher resolution features, and identifies many more low-abundance taxa than 16S sequencing. Applying the same filter to remove low-abundance features may have eliminated an advantage of the metagenomics data. Another avenue for further differentiation is feature engineering. For our benchmarking, both 16S and metagenomics features were relative abundances of microbes in each sample. Since metagenomics data sequences all available DNA in a sample, we are not limited to just identifying which microbes are present in the community. For example, we can create a functional profile of the community based on the types of genes in a sample. Compared to typical feature engineering, which consists of mathematical transformations of input data, this approach of biologically-informed feature engineering requires domain expertise in the features themselves. Biologically-informed feature engineering that exploits advantages of metagenomics sequencing is the most promising avenue for improving predictive performance over 16S data.

Appendix A

Code availability

Code for the plugin developed for this benchmarking, Q2-MLAB, can be found on GitHub at this link.

Bibliography

[1] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, Deright Goldasich, Pieter C Dorrestein, Robert R Dunn, Ashkaan K Fahimipour, James Gaffney, Jack A Gilbert, Grant Gogul, Jessica L Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A Jackson, Stefan Janssen, Dilip V Jeste, Lingjing Jiang, Scott T Kelley, Dan Knights, Tomasz Kosciolek, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V Melnik, Jessica L Metcalf, Hosein Mohimani, Emmanuel Montassier, Gholamali Rahnavard, Adam Robbins-pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Yoshiki Vázquez-baeza, Alison Vrbanac, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, American Gut, Morton Jt, Antonio Gonzalez, Gail Ackermann, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, L Goldasich, Dorrestein Pc, Dunn Rr, Green Jl, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Jackson Ma, Stefan Janssen, Lingjing Jiang, Kelley St, Dan Knights, and Tomasz Kosciolek. American Gut : an Open Platform for Citizen Science. *mSystems*, 3(3):1–28, 2018.

[2] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel Mcdonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group*, 7(5):335–336, 2010.

[3] Zhi Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams. Big data opportunities and challenges: Discussions from data analytics perspectives [Discussion Forum]. *IEEE Computational Intelligence Magazine*, 9(4):62–74, 2014.

[4] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen,

Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein, and Rob Knight. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*, 15(10):796–798, 2018.

[5] Yoshiki Vázquez-Baeza, Chris Callewaert, Justine Debelius, Embriette Hyde, Clarisse Marotz, James T. Morton, Austin Swafford, Alison Vrbanac, Pieter C. Dorrestein, and Rob Knight. Impacts of the Human Gut Microbiome on Therapeutics. *Annual Review of Pharmacology and Toxicology*, 58(1):253–270, 2018.

[6] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[7] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S. Dusko Ehrlich, Jun Wang, Maria Antolin, François Artiguenave, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten Friss, Maarten Van De Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem De Vos, Yohanan Winogradsky, and Erwin Zoetendal. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[8] G. Clarke, S. Grenham, P. Scully, P. Fitzgerald, R. D. Moloney, F. Shanahan, T. G. Dinan, and J. F. Cryan. The microbiome-gut-brain axis during early life regulates the hippocampal serotonergic system in a sex-dependent manner. *Molecular Psychiatry*, 18(6):666–673, 2013.

[9] Susannah G Tringe and Philip Hugenholtz. A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11(5):442–446, 2008.

[10] Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A selective review of multi-level omics data integration using variable selection. *High-Throughput*, 8(1):1–25, 2019.

[11] Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology*, 8(7), 2012.

[12] Nicholas Bokulich, Matthew Dillon, Evan Bolyen, Benjamin Kaehler, Gavin Huttley, and J Caporaso. Q2-Sample-Classifier: Machine-Learning Tools for Microbiome Classification and Regression. *Journal of Open Source Software*, 3(30):934, 2018.

[13] Jessica L Metcalf, Laura Wegener Parfrey, Antonio Gonzalez, Christian L Lauber, Dan Knights, Gail Ackermann, Gregory C Humphrey, Matthew J Gebert, Will Van Treuren, Donna Berg-Lyons, Kyle Keepers, Yan Guo, James Bullard, Noah Fierer, David O Carter, and Rob Knight. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife*, 2:1–19, 2013.

[14] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller. Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1):29–33, 2015.

[15] Yi Hui Zhou and Paul Gallins. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in Genetics*, 10(JUN):1–14, 2019.

[16] Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Computational Biology*, 12(7):1–26, 2016.

[17] Nicholas A. Bokulich, Thomas S. Collins, Chad Masarweh, Greg Allen, Hildegarde Heymann, Susan E. Ebeler, and David A. Millsa. Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *mBio*, 7(3):2011–2014, 2016.

[18] Dan Knights, Elizabeth K Costello, and Rob Knight. Supervised classification of human microbiota. *FEMS microbiology reviews*, 35(2):343–359, mar 2011.

[19] Katja Borodulin, Hanna Tolonen, Pekka Jousilahti, Antti Jula, Anne Juolevi, Seppo Koskinen, Kari Kuulasmaa, Tiina Laatikainen, Satu Ma, Markku Peltonen, Markus Perola, Pekka Puska, Veikko Salomaa, and Jouko Sundvall. Cohort Profile : The National FINRISK Study. *International Journal of Epidemiology*, (November 2017), 2018.

[20] Robert C Kaplan, Larissa M Avile, Janice Barnhart, Kiang Liu, Aida Giachello, David J Lee, John Ryan, D R Ph, Michael H Criqui, and John P Elder. Sample Design and Cohort Selection in the Hispanic Community Health Study / Study of Latinos. *Annals of Epidemiology*, 2010.

[21] Anne-katrin Pröbstel and Sergio E Baranzini. The Role of the Gut Microbiome in Multiple Sclerosis Risk and Progression : Towards Characterization of the " MS Microbiome ". *Neurotherapeutics*, pages 126–134, 2018.

[22] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G.I. Langille, Joslynn Lee, Ruth Ley, Yong Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J.J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H.D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8):852–857, 2019.

[23] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk,

Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[24] Aeriel Belk, Zhenjiang Zech Xu, David O. Carter, Aaron Lynne, Sibyl Bucheli, Rob Knight, and Jessica L. Metcalf. Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes*, 9(2), 2018.