Open access • Posted Content • DOI:10.1101/2021.05.03.442488

Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease — Source link

Kubinski R, Djamen-Kepaou J, Zhanabaev T, Hernandez-Garcia A ...+7 more authors

Institutions: Université de Montréal, Max Planck Society, Norwich Research Park, Centre Hospitalier Universitaire Sainte-Justine

Published on: 04 May 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Generalizability theory

Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease

Ryszard Kubinski[1*], Jean-Yves Djamen-Kepaou[1], Timur Zhanabaev[1], Alex Hernandez-Garcia[2], Stefan Bauer[3], Falk Hildebrand[4,5], Tamas Korcsmaros[4,5], Sani Karam[1], Prévost Jantchou[6], Kamran Kafi[1], Ryan D. Martin[1*]

[1]Phyla Technologies Inc, Montreal, Canada

[2]Mila, Quebec Artificial Intelligence Institute, University of Montreal, Montreal, Canada

[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany

[4]Gut Microbes & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK.

[5]Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK.

[6]Centre Hospitalier Universitaire Sainte-Justine, Montréal, Canada.

*Co-corresponding authors - send correspondence to richard@phyla.ai or ryan.martin@phyla.ai

Keywords

Inflammatory bowel disease, machine learning, gut microbiome, batch effect reduction, data normalization

Abstract

Background

Inflammatory bowel disease (IBD) patients wait months and undergo numerous invasive procedures between the initial appearance of symptoms and receiving a diagnosis. In order to reduce time until diagnosis and improve patient wellbeing, machine learning algorithms capable of diagnosing IBD from the gut microbiome's composition are currently being explored. To date, these models have had limited clinical application due to decreased performance when applied to a new cohort of patient samples. Various methods have been developed to analyze microbiome data which may improve the generalizability of machine learning IBD diagnostic tests. With an abundance of methods, there is a need to benchmark the performance and generalizability of various machine learning pipelines (from data processing to training a machine learning model) for microbiome-based IBD diagnostic tools.

Results

We collected fifteen 16S rRNA microbiome datasets (7707 samples) from North America to benchmark combinations of gut microbiome features, data normalization methods, batch effect reduction methods, and machine learning models. Pipeline generalizability to new cohorts of patients was evaluated with four binary classification metrics following leave-one dataset-out cross validation, where all samples from one study were left out of the training set and tested upon. We demonstrate that taxonomic features obtained from QIIME2 lead to better classification of samples from IBD patients than inferred functional features obtained from PICRUSt2. In addition, machine learning models that identify non-linear decision boundaries between labels are more generalizable than those that are linearly constrained. Prior to training a non-linear machine learning model on taxonomic features, it is important to apply a compositional normalization method and remove batch effects with the naive zero-centering method. Lastly, we illustrate the

50 importance of generating a curated training dataset to ensure similar performance across patient
51 demographics.

52 **Conclusions**

53 These findings will help improve the generalizability of machine learning models as we move
54 towards non-invasive diagnostic and disease management tools for patients with IBD.

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

Introduction

The human gut microbiome is a collection of microbes, viruses, and fungi residing throughout the digestive tract. The gut microbiota plays an important role in human health, influencing food digestion, the immune system, mental health, and numerous other functions (reviewed in [1]). In line with the functional role in human health, alterations in the gut microbiome have been linked to illnesses such as multiple sclerosis, type II diabetes, and inflammatory bowel disease (IBD) [2, 3]. IBD comprises two main subtypes: Crohn's disease (CD) and ulcerative colitis (UC), characterized by periodic inflammation throughout the gastrointestinal tract or localized to the colon, respectively [4]. The prevalence of IBD is increasing globally over the last several decades, from 79.5 to 84.3 per 100 000 people between 1990 and 2017, with Canada having among the highest IBD rates at 700 per 100 000 people in 2018 [5, 6]. Although the disease etiology is currently undetermined, the increasing rates of IBD have been linked to lifestyle factors, such as a Western diet [7].

Currently, IBD diagnosis and monitoring is primarily performed via blood tests, fecal calprotectin, and endoscopies. These methods can be costly, invasive, and display variable accuracy, all of which leads to delayed diagnosis and infrequent disease monitoring [8]. Therefore, there is an unmet need for the development of further non-invasive, low-cost, and rapid methods for screening, diagnosis, and disease management for the growing number of IBD patients [9, 10]. One potential diagnostic test within these constraints involves using the gut microbiome composition to identify patients with IBD.

Over the past decade, several studies have compared the gut microbiome profiles of healthy individuals and those with CD or UC [2, 11–21]. Common characteristics of the gut microbiome identified in patients with IBD are the reduction in bacterial diversity and development of a dysbiotic state, referring to alterations in the structure and function of the gut microbiome compared to healthy individuals [12, 14, 19]. Principal coordinate analysis with UniFrac [16] or

Bray-Curtis [17] distance of the gut microbiome's composition has identified differential clustering of healthy and IBD samples. Although the dysbiotic state is commonly identified in IBD patients, it remains unknown whether the microbiome initiates IBD or is only a reflection of the patient's current health status. Larger meta-analyses have aimed to identify differentially abundant taxa between IBD patients and healthy controls in order to generate potential diagnostic biomarkers, although with limited success to date [18].

Due to difficulties identifying biomarkers with standard statistical methods for disease diagnosis, the field has moved to applying predictive machine learning (ML) models for classification of patient phenotypes. Several studies have demonstrated accurate classification of patients with IBD from their gut microbiome profile with ML models [2, 12, 13, 15, 18, 22–24]. Common ML models employed for IBD classification include random forest (collection of decision trees for classification) [2, 15], logistic regression (binary linear classifier) [13], and neural networks (layers of differently weighted nodes contributing to a classification) [23, 24].

Features commonly used for IBD classification with ML models can be categorized into three groups: clinical, bacterial, and functional. Clinical features encapsulate those regarding the patient (i.e. age, sex, body mass index (BMI)) and results from other clinical tests (i.e. calprotectin, colonoscopy), which are independent of a patient's microbiome profile [25]. Taxonomy and functional features are usually determined via sequencing-based microbiome profiling, such as amplicon sequencing of the 16S rRNA gene or whole genome shotgun (WGS) sequencing of all DNA in a sample [26]. Bioinformatic tools, such as QIIME2 [27] or LotuS2 [28], provide pipelines for clustering 16S rRNA-amplicon sequences into operational taxonomic units (OTUs) which can then be compared to public databases to find taxonomy assignments [29]. WGS reads are frequently used to infer potential functions represented in the genomes of microbial community members (reviewed in [30]). Similarly, we can use known genomes in public databases to derive functional predictions in a community based solely on amplicon sequencing based taxonomy profiles, implemented in tools such as PICRUSt2 [31]. Although WGS provides greater taxonomic

resolution and estimates of microbiome functions, 16S rRNA amplicon sequencing is currently more applicable to a diagnostic test due to its speed, affordability, and standardization of analysis tools.

A critical, and often under-explored, consideration for generating ML models for disease classification is their generalizability to previously unseen cohorts of patients. A ML model that underperforms when presented with data from a new patient cohort is not reliable enough to be applied in a clinical setting [32]. Despite this, models currently used in the context of microbiome data are often only trained and cross-validated with different splits of data from the same cohort. In studies where cross-validation with an unseen sample cohort is performed, the model's performance is often lower, indicative of the model overfitting to the training set [22, 23]. A proposed explanation for the reduced performance is the potential for introduction of non-biological variability to the data by wet-lab protocols and sequencing instruments during the processing of these samples, typically observed in meta-analysis of microbiome data [12].

In order to improve model performance on unseen data, it is necessary to apply normalization and batch effect reduction techniques prior to model training. Normalization is a critical step to remove biases to feature abundance estimates, such as the data's compositional nature, heteroskedasticity, or skewness. For example, microbiome data's compositional nature prevents the direct application of standard statistical methods as they may lead to erroneous results, and requires prior application of compositional normalization methods [33, 34]. In addition, methods have been developed to remove the technical "batch effects" commonly identified in collections of samples from different studies, such as naive zero-centering methods and the recently developed empirical Bayes' method, Meta-analysis Methods with a Uniform Pipeline for Heterogeneity (MMUPHin) [35–38]. To date, the effect of various combinations of normalization and batch effect reduction techniques on ML model generalizability remains to be benchmarked.

In this article, we propose a standardized approach for evaluating the performance and generalizability of data processing pipelines and ML models with microbiome data to classify

patients with IBD. Previous microbiome ML benchmarking studies focused on performance of various combinations of model type, normalization, and microbiome compositional features using variations of fivefold cross validation [24, 39]. Fivefold cross validation fails to assess the generalizability to new, unseen sample batches as each split potentially contains samples from all batches present in the dataset. Therefore, we implemented a leave-one-dataset-out (LODO) [40] cross-validation method to directly assess cross-batch generalizability. In this approach, the model is iteratively trained on samples of all but one dataset and then tested on the left-out dataset. Different combinations of data types, normalization methods, batch effect reduction methods, and ML models were assessed in order to establish a comprehensive performance benchmark of microbiome-based disease classification in the context of IBD.

Results

Overview of samples and methods

In order to assess the cross-batch performance of each pipeline, we implemented a LODO cross validation approach. We collected 16S rRNA gene next generation sequencing data from 15 studies in North America for a total of 7707 samples, comprising 55% healthy and 45% IBD samples, of which 56% are CD and 44% are UC **(Table 1)**. We completed 15 cross-validation iterations with a single dataset removed from the training set for generation of the classification model which was then used to assess model performance (**Figure 1)**.

We evaluated the ability to classify samples from patients with IBD or non-IBD controls using different combinations of three taxonomic feature sets or six functional feature sets, eight normalization methods, four batch effect reduction methods, and nine machine learning models (**Figure 1**). The binary classification performance of each combination of feature set, normalization, batch effect reduction, and machine learning model was assessed with four classification metrics: F1 score, Matthews Correlation Coefficient (MCC), binary accuracy, and Area Under the receiver operating characteristics Curve (ROC-AUC, abbr. AUC) [60, 61]. We assessed generalizability through two methods. First, we sorted the pipeline components of interest (e.g. types of machine learning models) by the mean and standard deviation of their performance assessed by each metric. Second, in order to determine if the performance was significantly different, we performed statistical comparison of the pipelines' metrics with a Mann-Whitney U test. Therefore, the most generalizable component was defined as the top sorted method which displayed significantly better performance than baseline or other methods.

Top IBD classification was obtained using taxonomic features

Taxonomic features (species, genus or OTU) are predominantly used as input for ML models, whereas it is less common to use inferred functional features from PICRUSt2 as input.

However, previous studies have identified lower inter-individual variation of the gut microbiome's inferred functional profile than taxonomy [62, 63], suggesting that functional features may lead to better classification performance and generalizability. We processed the 16S sequencing samples with QIIME2 and PICRUSt2 to obtain taxonomy and functional feature abundance estimates, respectively.

For each ML model, we assessed the performance with taxonomy and functional abundance features in combination with normalization and batch effect reduction methods. Independent sorting of four classification performance metrics indicated that the taxonomic features classified IBD samples more effectively than functional features **(Figure 2A, Supplemental Figure 1A)**. Comparison of performance with taxonomy and functional features confirmed the significantly higher performance for classification of IBD samples with taxonomic features (**Figure 2B, Supplemental Figure 1B)**. Therefore, ML models using taxonomic features from this dataset lead to better classification of IBD samples than functional features.

Taxonomic classification with QIIME2 consists of seven hierarchical ranks, with kingdom and species at the top and bottom, respectively. Each consecutively lower taxonomy rank provides greater resolution of the gut microbiome's composition while also increasing data sparsity, which can negatively affect an ML model's performance [64]. Previous literature comparing different taxonomy ranks for disease classification indicated that lower ranks, down to genus, improved performance [65]. We assessed whether the trend for improved classification continued with the species rank and OTUs, despite their increasing sparsity. While no significant performance difference was observed between species and genus ranks, both displayed significantly higher classification performance than OTU features (**Supplemental Figure 1C)**.

Non-linear models achieve greatest classification performance

Machine learning classification models identify decision boundaries within the feature space to separate sample labels from one another. For some ML models (BNB, Linear SVC, LR), these

boundaries are linearly constrained, whereas others (RF, KNN, MLP, Radial SVC, XGBoost) can identify more complex, non-linear relationships between features and class. We assessed the generalizability of three linear and five non-linear ML models across the taxonomy and functional feature sets.

Independent sorting of ML models for each performance metric indicated that the non-linear models had greater classification performance (top five were non-linear models) than the linear models **(Figure 3A, Supplemental Figure 2A)**. Comparison of aggregate scores further confirmed non-linear models had significantly higher F1 score, MCC, AUC, and accuracy than linear models (**Figure 3B, Supplemental Figure 2B)**. In order to directly assess whether the non-linearity of a model improves classification in the context of microbiome data, we compared linear and non-linear variations of a support vector machine and logistic regression. Comparison of the two variations enables direct analysis of the impact of decision boundary constraints on performance, independent of differences in model architecture. The non-linear (radial) version of logistic regression and support vector machines (Radial) had significantly greater performance than the linear version (Linear) across all four metrics (**Figure 3C, Supplemental Figure 2C)**. Lastly, we assessed which non-linear model led to the highest classification performance. Across all four metrics, the random forest and XGBoost models were significantly better than MLP, KNN, and radial SVC models (**Supplemental Figure 2D)**. In conclusion, non-linear models provided more accurate IBD classification, likely due to the complex relationships between features and disease labels.

Other ML model architectures, such as convolutional neural networks (CNNs), are commonly used for classification problems with certain structure in the input data, such as image classification. In the context of microbiome data, the CNN MDeep adds structure to OTU features through hierarchical agglomerative clustering of the phylogeny-induced correlation between OTUs [58]. As MDeep is currently only developed for OTU features, we assessed whether this CNN architecture led to greater classification performance with OTU abundance than our MLP

247 architecture. Comparison of each performance metric across all normalization and batch effect
248 reduction methods indicated MDeep performance was not significantly different than our MLP
249 model (**Supplemental Figure 2E**).

250 Due to the significantly better performance of non-linear classification models and
251 taxonomic features, our subsequent analysis of normalization and batch effect reduction methods
252 utilized only taxonomic feature sets and non-linear models.

253

254 **Evaluation of normalization methods**

255 We assessed normalization methods which account for different biases commonly
256 observed in next-generation sequencing data: compositionality, heteroskedasticity, and
257 skewness. We selected two normalizations designed for compositional data: the isometric log
258 ratio (ILR) and centered log ratio (CLR) [66]. We selected two normalization methods which aim
259 to reduce the heteroskedasticity: the arcsine square root (ARS) transformation [67] of the total
260 sum scaling (TSS) values and the variance stabilized transformation (VST) from the R package
261 DESeq2 [68]. Next, we assessed a log transformation of the TSS values (LOG), which reduces
262 the positive skew commonly seen in the distribution of microbiome data. Lastly, we assessed
263 normalization by TSS alone to remove differences in sequencing depth between samples or no
264 normalization (NOT).

265 Independent sorting of each performance metric consistently identified the compositional
266 normalization methods (CLR and ILR) as the most generalizable across non-linear models,
267 followed by the variance/distribution modifiers (ARS, LOG, VST), and TSS as the consistently
268 lowest performing normalization (**Figure 4, Supplemental Figure 3A**). Furthermore, the
269 compositional methods led to significantly better performance than the other normalization types
270 across all four metrics (**Supplemental Figure 3B**), whereas the variance/distribution modifiers
271 and scaling method were only significantly better than no normalization. These results indicate

the importance of normalization methods which account for the compositional properties of microbiome data prior to model training.

Evaluation of batch effect reduction methods

A common issue with combining next-generation sequencing datasets for meta-analyses is the systematic differences between datasets due to differences in technical protocols. These differences add non-biological variation to the samples, decreasing the ability to ascertain biological signals [69]. Various approaches have been proposed to remove technical artifacts from dataset collections, of which we selected two relevant to microbiome data [37, 38]. First, zero-centering methods aim to reduce batch effects by centering the mean of each feature within a batch to zero. Second, Meta-analysis Methods with a Uniform Pipeline for Heterogeneity in microbiome studies (MMUPHin) [39] (microbiome specific empirical Bayes' methods), estimate and remove batch-specific parameters for each feature. Two variations of MMUPHin were implemented to simulate the scenario of obtaining a new dataset when implemented for a diagnostic test. The first (#1) applied MMUPHin to the training and test sets separately, whereas the second (#2) only applied MMUPHin to the training set (see Methods for detailed description).

The different batch effect reduction methods were sorted by the mean and standard deviation of their performance across taxonomic features, non-linear models, and all normalization methods. Our sorting method indicated that zero-centering was the most generalizable approach across the non-linear models. Whereas MMUPHin #1 and #2 were less generalizable than no batch reduction, with MMUPHin #2 the least generalizable (**Figure 5, Supplemental Figure 4A**). Additionally, zero-centering led to significantly higher binary accuracy and MCC compared to all other methods, whereas F1 score and AUC were higher only when compared to no batch effect reduction and MMUPHin #2. MMUPHin #1 had significantly better performance than MMUPHin #2, with no difference in performance observed compared to no batch effect reduction (**Supplemental Figure 4B)**.

298

299 **Evaluation of model performance on sample and patient subgroups**

300 The samples used to assess the performance of different combinations of normalizations,
301 batch effect reduction, and ML models were drawn from across sample collection methods (i.e.
302 stool and biopsy) and patient demographics (i.e. paediatric and adult samples). While we did not
303 set inclusion criteria for samples based on these differences, previous research has demonstrated
304 distinct differences in microbiome composition between sample types and demographic groups
305 [18, 70, 71]. For example, principal coordinate analysis (PCoA) with weighted UniFrac distance
306 [72] and principal component analysis (PCA) of CLR-transformed taxonomic features indicated
307 paired biopsy and stool samples from the same individual cluster separately [73].

308 We compared the model performance for the sample and patient demographics for which
309 we were able to acquire sufficient metadata and have been associated with microbiome
310 alterations: sample type (biopsy vs. stool), IBD subtype (CD vs. UC), sex (Female vs. Male), BMI
311 (BMI < 30 vs. BMI > 30), and age (Adult vs. Pediatric). To assess the performance within each
312 demographic, we included the predictions from taxonomic features (species, genus, OTU) with a
313 compositional normalization method, zero-centering batch effect reduction, and a non-linear ML
314 model. Our analysis focused on the MCC performance metric as it is more robust to imbalanced
315 label distribution [61], which occurred when the samples were grouped by the five metadata
316 categories mentioned. A logistic regression function was used to assess changes in performance
317 corresponding to each demographic while controlling for the other metadata (**Table 2)**.

318 The models displayed reduced performance for biopsy samples compared to stool samples,
319 increased performance for samples from adult patients compared to paediatric patients, and
320 decreased performance of samples from patients with BMI less than 30 compared to patients with
321 BMI greater than 30. On the other hand, there was no difference in classification performance for
322 females compared to males or for samples from patients with CD compared to patients with UC
323 (**Table 2**). Similar results were reproduced with F1 scores, AUC, and accuracy (**Supplemental**

Table 1). The metadata groups with different performance between the two categories coincided with those that are not equally represented in our dataset, highlighting the importance of accounting for different demographic groups in a microbiome based diagnostic test.

Evaluation of top performing pipeline combinations for IBD classification

Our analysis identified the features, ML models, normalization methods, and batch effect reduction methods which led to the most generalizable performance. In order to determine the best overall combination of features, data processing, and ML model we assessed the top three performing models (**Table 3**). The top three models consisted of the most generalizable individual components: taxonomic features (genus), non-linear model (XGBoost or RF), compositional normalization (ILR or CLR), and zero-centering to remove batch effects. Therefore, the combination of the most generalizable methods led to the best classification performance.

Identification of important features for classification with a XGBoost model

In addition to predicting disease diagnoses, machine learning models can be used to identify biomarkers for disease by identifying features important for disease classification. We characterized the feature importance from the second-best overall data processing and ML model pipeline (**Table 3**). We did not analyze the feature importance of the best-performing model because the ILR normalized values no longer correspond to the starting features thereby preventing interpretation of feature importance. For an XGBoost model, the importance corresponds to a feature's contribution to the model's decision during training, referred to as the gain value [57]. We extracted the features' gain values from each of the 15 LODO iterations, sorted by the mean of all iterations, and plotted the top fifteen features (**Figure 6)**. In addition, we determined the change in abundance for each taxonomy to assess whether our dataset aligned with previous findings on changes of the microbiome in IBD.

Amongst the top features are many taxa in the short chain fatty acid (SCFA) producing Clostridium XIVa/IV clusters, including bacteria from the *Eubacterium*, *Coprococcus*, *Lachnospira*, and *Ruminiclostridium* genera (**Figure 6A)**. Aligning with previous studies, these bacteria were decreased, with the exception of *Coprococcus 3*, in IBD samples vs control samples in our dataset (**Figure 6B)** [2, 74]. *Fusobacterium* and *Veillonellaceae* genera, commonly increased in the gut microbiome of IBD patients, were also top contributors to the XGBoost classifier (**Figure 6A/B)** [2, 75]. In addition, the *Prevotellaceae* genus was the second most important feature, with the decreased abundance in IBD samples agreeing with previous studies showing decreased abundance in the gut microbiome of patients with CD and UC (**Figure 6A/B)** [76]. XGBoost classifiers have the best potential for use as a diagnostic test due to their performance as well as their interpretability and utility in identifying disease biomarkers.

Discussion

We assessed how different feature sets, ML models, normalization methods, and batch effect reduction methods affect predictive performance across patient cohorts in a LODO cross validation approach. The limited applicability of a PCR-based diagnostic test with a handful of microbiomes for IBD diagnosis [77] has led the field to explore the use of ML models for disease diagnosis. Our benchmark provides practical suggestions for ways to improve the performance of an IBD diagnostic test using the gut microbiome composition. First, genus abundance estimates from 16S rRNA sequencing need to be normalized by a compositional normalization method, with CLR normalization being the most appropriate as it allows for each features importance to the ML models decision to be assessed. Second, zero-centering batch effect reduction should be applied to each batch of samples collected, sequenced, and processed together to reduce systematic batch differences. Following normalization and batch effect reduction, an XGBoost or random forest classification model should be trained and optimal hyperparameters determined for implementation as a diagnostic test. With respect to the training dataset, it is important to account for patient demographics or technical differences between samples that have been associated with gut microbiome alterations. We suggest several options for optimal performance: (1) ensure balanced representation in the training dataset, (2) include the metadata labels as a feature for the model, or (3) deploy diagnostic ML models built specifically for one demographic group. In addition, the LODO cross-validation methodology is an important tool for the selection of new data preprocessing and model building methods.

Previous studies have demonstrated greater consistency of functional feature abundances than taxonomic feature abundance in both healthy individuals [78–80] and those with IBD [63, 81]. In fact, some studies were unable to identify a single bacterium present in every IBD patient from their cohort [62]. The reduced variation and sparsity of functional features led us to hypothesize that functional abundance profiles would lead to better classification of IBD samples.

However, through our LODO cross validation, we found that classification performance with functional features was significantly worse than with taxonomic features **(Figure 2B, Supplemental Figure 1B)**. We postulate the reason for the reduced classification performance with functional profiles is due to the limited recapitulation of functional profiles with PICRUSt2 [31, 82] and the inability of 16S rRNA sequencing to identify strain-level functional differences of the present bacteria [83]. To overcome these limitations in future studies, measurement of the microbiome's gene content by WGS, transcriptomes by RNA-seq, or metabolites by metabolomics need to be explored. In fact, functional profiles from whole genome sequencing led to better predictions of patients with IBD who achieved remission with vedolizumab than taxonomy abundance [23]. While whole genome sequencing may improve disease classification, its much higher cost than 16S rRNA sequencing substantially hinders the technology's adoption as a diagnostic test.

A major hurdle in the implementation of sequencing based diagnostic tests in the clinic is the observed systematic differences between sample preparations. In a previous study, removal of these batch effects with an empirical Bayes' or zero centering approach led to improved classification [84]. However, our work only identified improved cross-batch classification performance with zero-centering and not the empirical Bayes' method MMUPHin (**Figure 5)**. Current empirical Bayes' approaches are designed and optimized for disease mechanism and biomarker discovery where the disease covariate is known and incorporated in the method. The inclusion of a disease covariate is not applicable to a diagnostic scenario though, where the diagnosis label is to be determined. The lack of improvement in classification performance with MMUPHin #1 compared to no batch reduction is potentially due to its implementation in a scenario the method was not optimized for.

Similar to batches of samples collected for a diagnostic test, the batches in our dataset were not balanced, with some containing only a single diagnosis class (e.g. all samples coming from IBD patients). In cases where the batch and diagnosis label are confounded, batch correction

methods tend to reduce the disease associated differences in the process of removing the batch differences [37]. Therefore, the more advanced removal of batch effects by MMUPHin likely led to an over-adjustment within the unbalanced batches and removal of the disease differences. Whereas, the less sophisticated removal of batch effects with the covariate naive zero-centering approach retained sufficient biological signal between disease labels for non-linear ML models to correctly classify samples across batches. Batch correction methods that do not require input of a covariate have been developed, such as frozen surrogate variable analysis or reference principal component integration (RPCI) [85, 86], although their applicability to microbiome data has not been assessed.

The sparse availability of metadata for the samples led to several limitations in our analysis. First, the identification of CD and UC patients relied on the accuracy of the diagnosis coding in the public databases. However, there were no studies explicitly validating the registration of CD and UC diagnosis codes. Second, although our study demonstrated reliable results, gaps in the publicly available data prevented us from several critical analyses. For instance, we lacked information on how the patients were diagnosed in every study, the timing of sample collection in relationship to their diagnosis and disease progression, current disease activity quantification, DNA extraction and sample storage information. Furthermore, there was limited information on environmental factors such as medication usage, alcohol usage, smoking, diet, and other factors known to alter the gut microbiome which could affect our analysis [81, 87]. Of the sample information and patient demographic data we obtained, clear differences in performance of our top pipelines were observed (**Table 2**). Therefore, future studies with improved lifestyle and clinical metadata are needed to systematically address how these factors affect performance of a gut microbiome diagnostic test.

Other non-invasive diagnostic tests for IBD, such as fecal calprotectin, continue to have significant differences between the reports on the sensitivity and specificity for classifying IBD patients from non-IBD [88, 89]. While high performance levels have been reported, one recent

448 study identified a 78% accuracy for identifying patients with IBD using fecal calprotectin [90],
449 which is approximately 10% lower than our best model. Furthermore, while we focused solely on
450 IBD classification, ML models using microbiome composition have wider applicability than
451 singular biomarkers such as calprotectin. Models using microbiome data have already been
452 implemented to predict if a patient with IBD will respond to a medication [23], to predict a patient's
453 postprandial glycemic response [91], and for classification of other diseases, such as Parkinson's
454 disease [48], to name a few.
455

456 ## Conclusion

457 With sufficient data and validation, analysis of the fecal gut microbiome can indeed be
458 leveraged as a multi-purpose predictive tool. Given the significant delay [92–94] and associated
459 costs of diagnosis [95, 96], it is critical to continue exploration of approaches that increase
460 accessibility of diagnosis and decrease the cost of testing [97] in a community health or primary
461 care setting. XGBoost and random forest machine learning models with microbiome data have
462 the potential to achieve these goals. Further work to gather more well-annotated data, improve
463 performance and assess models with validation studies is required.
464
465
466
467
468
469
470
471
472

Methods

Acquisition of sample data

Sample FASTQ files were acquired from the European Nucleotide Archive (ENA) browser. The sample metadata was acquired from the corresponding publication's supplementary materials or the QIITA microbiome platform. Only samples collected from individuals in North America were used from each dataset. The dataset accessions and technical information regarding the samples in each dataset are available in **Supplemental Table 2**.

The following fifteen studies were included in our dataset:

1. The American Gut cohort is from a large, open platform which collected samples from individuals in the US to identify associations between microbiomes, the environment, and individual's phenotype [41]. We included available samples that did not contain any self-reported diseases in the metadata.
2. The CVDF study determined the effect of cardiorespiratory fitness on microbiome composition and comprises a range of fitness levels [42].
3. The GEVERSM study assessed the microbiome composition of treatment naive, newly diagnosed, paediatric patients with IBD and adult patients diagnosed with IBD for 0 to 57 years [2].
4. The GEVERSC cohort consists of additional samples from paediatric and adult patients added to the GEVERSM study [2].
5. The GLS study longitudinally sampled 19 patients with CD (Crohn's disease activity index (CDAI) between 44 and 273) and 12 healthy control individuals [20].
6. The Human Microbiome Project (HMP) study longitudinal tracked paediatric and adult patients ranging from newly diagnosed to diagnosed for 39 years. Diagnosis was confirmed by colonoscopy prior to enrollment in the study along with several other inclusion criteria listed in the corresponding publication [43].

7. The MUC study collected mucosal biopsies from 44 paediatric patients with CD and 62 non-IBD paediatric control patients [21].
8. PRJNA418765 was a longitudinal study of patients with CD that were refractory to anti-TNF initiating ustekinumab assessed at week 0, 4, 6 and 22. To be included, patients required at least three months Crohn's disease history and a CDAI between 220 and 450 [44].
9. PRJNA436359 was a longitudinal study of new onset and treatment naive paediatric patients with UC receiving a variety of medications at week 0, 4, 12, and 52. Inclusion criteria consisted of presence of disease beyond the rectum, Paediatric Ulcerative Colitis Activity Index (PUCAI) of 10 or more, and no previous therapy [45].
10. QIITA10184 was a study comparing five different fecal collection methods and their effect on the healthy participant's microbiome composition identified with 16S rRNA gene sequencing [46].
11. QIITA10342 study assessed the microbiome composition and function of healthy individuals in two American Indian communities in the United States [47].
12. QIITA10567 samples consist of the control individuals in a study linking alterations in microbiome composition to Parkinson's disease [48].
13. The QIITA1448 study compared microbiome composition of individuals in traditional agricultural societies in Peru to those in industrialized cities in the United States [49].
14. The QIITA2202 study collected longitudinal stool samples from two healthy individuals alongside detailed lifestyle characteristics to correlate with microbiome composition [48, 50].
15. The QIITA550 study collected longitudinal stool samples from two individuals to assess temporal changes in microbiome composition [51].

Taxonomy classification with QIIME2

Taxonomy abundance tables were generated from the FASTQ files using QIIME2 (v2020.2) [27]. Reads were trimmed to remove low quality reads (trimming parameters listed in Supplemental Table 1), chimeras removed, and sequences denoised using Dada2 [52] or Deblur (for GLS and AG only). The processed sequences were clustered into OTUs and the centroid sequences classified with a Naive Bayes classifier [53] at 99% identity using the Silva 132 99% reference database [29, 54, 55]. For classification, the corresponding 16S rRNA gene hypervariable region's sequences were extracted from the Silva 132 99% reference database with the QIIME2 plugin feature-classifier's extract-reads function using the primers from the respective study. The extracted reads and the corresponding taxonomy were used to train the Naive Bayes classifier with the QIIME2 plugin feature-classifier's fit-classifier-naive-bayes function. Taxonomic feature tables were collapsed to species (level 7) and genus (level 6) classification for further analysis.

Inferring Functional Abundance with PICRUSt2

Functional abundance tables were generated using PICRUSt2 (v2.3.0) from the OTU abundance table and representative OTU sequences from QIIME2. We generated abundance tables from the six different databases incorporated into PICRUSt2: Clusters of Orthologous Groups of proteins (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs (KO), Enzyme Commission (EC), Pfam protein domain (PFAM), TIGR protein family (TIGRFAM) and MetaCyc pathways. Each database is independently curated and provides information on different aspects of the functional properties present in the microbiome.

Leave-One-Dataset-Out (LODO) cross validation

The generalizability of each model, normalization, and batch effect reduction method, was determined through a cross validation strategy which assessed predictive performance on

549 previously unseen batches of samples **(Supplemental Figure 5)**. As there were 15 datasets, we
550 iterated through the full dataset 15 times, generating the training set by removing all samples from
551 a single dataset to a separate test set. The training set was used to prune features that were not
552 present in at least 10% of samples from one dataset. Following pruning, the remaining features
553 were selected from the test set and the samples were normalized and batch reduced with the
554 respective methods. Lastly, the training set was balanced to have the same number of healthy
555 and IBD samples by subsampling the label with the greater number of samples, while maintaining
556 the proportion of samples from each collection site, disease label (UC/CD/Control), and sample
557 type (stool/biopsy).

558 For our modified implementation of MMUPHin, the data processing was adjusted to ensure
559 the training and test sets were batch reduced separately. For MMUPHin #1, the test dataset's
560 samples were removed and the data from the remaining studies batch reduced with MMUPHin
561 prior to training the model. Independently, the full dataset (with training studies and test dataset)
562 was batch reduced and the test dataset's samples then used to assess the model's classification
563 performance (**Supplemental Figure 5**, MMUPHin #1). For MMUPHin #2, the training studies
564 were batched reduced with MMUPHin prior to model training and the model's classification
565 performance then assessed on non-batch reduced samples from the test dataset (**Supplemental**
566 **Figure 5**, MMUPHin #2). Lastly, feature abundance for some samples following MMUPHin batch
567 effect reduction on the training set when QIITA2202 was left out and the test set when HMP was
568 left out were all zero. Rows with all zero are not appropriate input for the compositional
569 normalization methods, therefore we replaced the feature values for these samples with equal
570 relative abundance prior to normalization.
571

572 **Feature Selection**

573 Following taxonomy classification and inference of functional abundance, features present
574 in less than 10% of the samples within each dataset were pruned from the dataset. The feature
575 pruning was performed on the training set only, with the features then selected from the test set.
576

577 **Normalization methods**

578 When possible, normalization methods were implemented using python (v3.6.12) and R
579 (v3.6.3) packages with the methods already incorporated. For CLR and ILR normalization, zero
580 values were first replaced with the multiplicative replacement function prior to normalization with
581 the clr and ilr functions, respectively, from the python package SciKit-Bio (v0.5.2). CLR performs
582 a log transformation of abundance values, which are normalized by the geometric mean of all
583 features. ILR uses a change of coordinate space projection calculation to transform proportional
584 data (or relative abundances) to a new space with an orthonormal basis.

585 For TSS normalization, the counts for each feature were divided by the sum of all feature
586 counts in the sample with a custom python function. The method constrains the sample row sum
587 to one, aiming to similarly scale all samples while maintaining biological information of microbial
588 abundances. For ARS normalization, the TSS normalized values were transformed with sqrt
589 function followed by the arcsin function from the python package numpy (v1.19.2). The LOG
590 normalization was also applied to the TSS normalized values using the log function from numpy
591 following replacement of all 0s with 1.

592 For VST normalization, we used the varianceStabilizingTransformation function in the R
593 package DESeq2 (v1.26.0). VST aims to factor out the dependence of the variance in the mean
594 abundance of a feature. The method numerically integrates the dispersion relation of the feature
595 mean fitted with a spline, evaluating the transformation for each abundance in the feature. VST
596 normalization was performed similarly to the previously described modified MMUPHin

implementation, with the training set normalized separately from the test set as the normalization is dependent on all samples present in the dataset.

Batch effect reduction methods

We explored two methods for batch effect reduction: naive zero-centering and an empirical Bayes method. The naive zero-centering batch effect reduction entails centering the mean of each feature within each batch to zero [37]. We also assessed MMUPHin, a recently developed empirical Bayes method designed specifically for zero-inflated microbial abundance data. MMUPHin estimates parameters for the additive and multiplicative batch effects, using normal and inverse gamma distributions, respectively. The estimated parameters are then used to remove the batch effects from the dataset [38, 56]. For MMUPHin, the sample type (stool/biopsy) was used as a covariate for MMUPHin #1 and the sample type and disease label (UC/CD/Control) were covariates for MMUPHin #2. We considered a batch as the whole dataset or split a dataset into multiple batches when the metadata indicated different sample preprocessing methods or samples were processed in different locations.

Standard machine learning models

We assessed the classification performance of standard machine learning and deep learning models. The standard models were implemented using the python package SciKit-Learn (v0.23.2). Hyperparameters were not optimized and decided prior to experimentation.

Bernoulli Naive Bayes Classifier

The Bernoulli Naive Bayes Classifier (BNB) model converts the feature space to binary values and then estimates parameters of a Bernoulli distribution for classification purposes. We implemented the BNB model using the default settings in SciKit-Learn.

Random Forest

Random Forest (RF) models use an ensemble of decision trees that discriminate the feature space by a sequence of threshold conditional statements. The power of the model comes from its non-linear classification capabilities and the number of trees used to label classification. We implemented the Random Forest classifier with the following modifications to the default SciKit-learn settings: n_estimaters = 500, max_features = sqrt, and class_weight = balanced.

K-Nearest Neighbour Classifier

The K-Nearest Neighbour Classifier (KNN) classifies each sample by majority vote of the K nearest neighbours in its surrounding. We implemented the K Neighbors classifier with the following modifications to the default SciKit-learn settings: n_neighbors = 6, weights = distance, and metric = manhattan.

Support Vector Machine Classifier

The Support Vector Machine Classifier (SVC) identifies multivariate decision boundaries that separate class labels. We implemented two SVC variations, the first with a linear kernel, constraining the decision boundary to a linear hyperplane, using the SGDClassifier class from SciKit-learn with the following modifications to default settings: loss = modified_huber, tol = 10e-5, and max_iter = 10000. The second variation used the radial basis function kernel with the SVC class from SciKit-Learn, which removes the linear constraint of the decision boundary, with the following modifications to the default settings: tol = 10e-6, class_weight = balanced, and max_iter = 100000.

Logistic Regression

Logistic Regression classification estimates the probability of a certain class in a binary classification problem using a statistical fit to the logistic function. We implemented the

649 LogisticRegression class from SciKit-Learn with the following modifications to the default settings:
650 solver = sag, class_weight = balanced, and max_iter = 10000. For the non-linear variation, the
651 feature space was first transformed with the radial basis function kernel implemented with the
652 rbf_kernel function from SciKit-Learn prior to fitting a logistic regression model.
653
654 Gradient Boosted Trees (XGBoost)
655 Gradient boosted trees consist of a collection of sequential decision trees, where each tree
656 learns and reduces the error of the previous tree [57]. The gradient boosted trees model was
657 implemented with the XGBoost package's (v1.2.0) XGBoostClassifier class with the following
658 modifications to default settings: n_estimators = 500.
659
660 **Deep learning models**
661 The deep learning models were built with the python package Tensorflow (v2.2.0). The
662 models were trained for up to 100 epochs with a batch size of 16 and samples shuffled. The best
663 weights were selected using early stopping (EarlyStopping callback) by monitoring the validation
664 loss (5% split of the training set) with a min_delta = $1x10^{-3}$ and patience = 10.
665
666 *Multilayer Perceptron (MLP)*
667 A MLP is a neural network architecture composed of one or more layers of fully connected
668 neurons that take as input the weights of the previous layer and output the result of an activation
669 function to the subsequent layer. For binary classification, the final layer contains a single node
670 that predicts the class probability. We implemented an MLP architecture with three hidden layers
671 of 256 neurons using a rectified linear unit (ReLU) activation function followed by a Dropout layer
672 with a dropout rate of 50%. The final layer predicted the class label with a sigmoid activation
673 function. The model was trained using a binary cross entropy loss function and the Adam
674 optimizer with a learning rate of 0.001.

675

676 *Convolutional Neural Network*

677 We implemented MDeep, a CNN architecture recently designed for microbiome data [58].
678 CNNs require an inherent structure to present in the data, which is added to the OTU dataset by
679 hierarchical agglomerative clustering of the phylogeny-induced correlation between OTUs. We
680 built a phylogenetic tree with the align_to_tree_mafft_fasttree function in the QIIME2 phylogeny
681 python plugin using the OTU representative sequences obtained from clustering 16S rRNA
682 sequences with QIIME2. The phylogenetic tree was imported into R using the phyloseq package
683 and the cophenetic distance between OTUs determined with the R package ape. The cophenetic
684 distance was then used to calculate the phylogeny-induced correlation as described in the original
685 study and OTUs clustered using the HAC function from the MDeep GitHub repository
686 (https://github.com/lichen-lab/MDeep).

687

688 **Performance metrics**

689 To measure the performance of the various normalization, batch effect reduction, and model
690 combinations we used four commonly used metrics for binary classification: F1 score, Area Under
691 the receiver operating characteristic Curve (AUC), binary accuracy and Matthews Correlation
692 Coefficient (MCC). Since the number of samples in each dataset ranged from 23 to 1279, we first
693 balanced the number of samples from each dataset by up sampling each (with replacement) to
694 100 000 samples while maintaining the confusion matrix proportions for each individual dataset.
695 Balancing the number of samples ensured that altered performance with a single, large dataset
696 did not control the overall score and changes in performance for small studies was still observed.
697 The up sampled dataset was then used to calculate the respective metrics using the functions
698 implemented in SciKit-Learn.

699

700 **Sample subgroup performance analysis**

701 We assessed the performance of our algorithm for five different metadata variables, each
702 with two categorical labels. The samples were grouped by the five variables, with the two
703 categories for each variable coded as 0 or 1. The performance metric was calculated within each
704 grouping for classification of control samples and either UC or CD (depending on the specific
705 grouping). For the logistic regression analysis, the metric was input as the dependent variable
706 and the five metadata groups as the independent variables. The MCC score was scaled with the
707 MinMaxScaler from SciKit-Learn to scale the range from 0 to 1 as required for the logistic function.
708
709 **Feature Importance from XGBoost Classifier**

710 In order to determine the importance of each taxonomy, we collected the features' gain
711 value from our second-best pipeline composed of CLR normalized, zero-centered, genus
712 abundance features with an XGBoost Classifier. The gain values were collected from the trained
713 XGBoost classifier in each LODO iteration separately.
714
715 **Taxonomy Differential Abundance**

716 Differential taxonomy abundance was performed with Analysis of Compositions of
717 Microbiomes with Bias Correction (ANCOM-BC) (v1.0.5) [59]. The fold change between control
718 samples and IBD samples (UC and CD) was determined with a Bonferroni multiple comparison
719 correction applied to the p-values.
720
721
722
723

Abbreviations

ANCOM-BC	Analysis of Compositions of Microbiomes with Bias Correction
ARS	Arcsine square root transformation
AUC	Area Under the receiver operating characteristics Curve
BMI	Body mass index
BNB	Bernoulli Naive Bayes Classifier
CD	Crohn's disease
CDAI	Crohn's disease activity index
CLR	Centered-log ratio
CNN	Convolutional neural network
COG	Clusters of Orthologous Groups of proteins
EC	Enzyme Commission
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
ILR	Isometric-log ratio
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbour Classifier
KO	KEGG orthologs
LODO	Leave-one-dataset-out
LOG	Log transformation
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
ML	Machine learning
MLP	Multilayer Perceptron
MMUPHin	Meta-analysis Methods with a Uniform Pipeline for Heterogeneity

NOT	No normalization
OTU	Operational taxonomic unit
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PFAM	Pfam protein domain
PUCAI	Paediatric Ulcerative Colitis Activity Index
RF	Random Forest
RPCI	Reference principal component integration
SCFA	Short chain fatty acid
SVC	Support Vector Machine Classifier
TSS	Total sum scaling
TIGRFAM	TIGR protein family
UC	Ulcerative Colitis
VST	Variance stabilized transformation
XGBoost	eXtreme Gradient Boosting
WGS	Whole genome shotgun

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Publicly available datasets were analyzed in this study. The raw sequencing data for the following 16S rRNA datasets were downloaded from European Nucleotide Archive at the following accession numbers: American Gut (PRJEB11419), CVDF (PRJNA308319), GEVERSC (PRJEB13680), GEVERSM (PRJEB13679), GLS (PRJEB23009), MUC (PRJNA317429), PRJNA418765, PRJNA436359, QIITA10184 (PRJEB13895), QIITA10342 (PRJEB13619), QIITA10567 (PRJEB14674), QIITA1448 (PRJEB13051), QIITA2202 (PRJEB6518), QIITA550 (PRJEB19825). The raw sequencing data for the HMP 16S rRNA dataset was downloaded from ibdmdb.org.

Competing interests

RK is a founder of Phyla Technologies Inc and is currently the Chief Scientific Officer. RM, JD, and TZ were employed by Phyla Technologies Inc at the time of the manuscript.

Funding

The work in this manuscript was funded by Investissement Québec Programme innovation – volet 1 and Quebec Ministry of Economy and Innovation's Entrepreneurship Assistance Program (PAEN) - component 3a. The work of FH and TK were supported by the Earlham Institute

(Norwich, UK) in partnership with the Quadram Institute Bioscience (Norwich, UK) and strategically supported by a UKRI BBSRC UK grant (BB/CSP17270/1). FH and TK were also supported by a BBSRC ISP grant for Gut Microbes and Health BB/R012490/1 and its constituent projects, BBS/E/F/000PR10353 and BBS/E/F/000PR10355. FH received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 948219)

Authors' contributions

RK, JD, TZ, and RM designed the data processing pipeline, performed the experiments and analyzed the pipelines' performance. RK and RM wrote the manuscript. AHG and SB contributed to the experimental design. AHG, SB, FH, TK, SK, PJ, and KK contributed to interpretation of the results and editing and revising the manuscript. All authors reviewed, revised, and approved the final manuscript.

Acknowledgements

We would like to thank Luca Cuccia, Laura Minkova, Houman Farzin, Michael Golfi, Paul Godin, and Yasmine Mouley (Phyla Technologies Inc.) for their feedback and support as the manuscript was completed. We would also like to thank Sébastien Giguère (Valence Discovery) for his guidance during our methodology development.

Author Information

Phyla Technologies Inc, Montréal, Canada

Ryszard Kubinski, Jean-Yves Kepaou Djamen, Timur Zhanabaev, Sani Karam, Kamran Kafi, Ryan D. Martin

Mila (Québec Artificial Intelligence Institute), University of Montreal, Montreal, Canada

Alex Hernandez-Garcia

Max Planck Institute for Intelligent Systems, Tübingen, Germany

Stefan Bauer

Gut Microbes & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK.

Falk Hildebrand, Tamas Korcsmaros

Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK.

Falk Hildebrand, Tamas Korcsmaros

Centre Hospitalier Universitaire Sainte-Justine, Montréal, Canada.

Prévost Jantchou

References

1. Mohajeri MH, Brummer RJM, Rastall RA, Weersma RK, Harmsen HJM, Faas M, et al. The role of the microbiome for human health: from basic science to clinical applications. Eur J Nutr. 2018;57 Suppl 1:1–14.

2. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15:382–92.

3. Opazo MC, Ortega-Rocha EM, Coronado-Arrázola I, Bonifaz LC, Boudin H, Neunlist M, et al. Intestinal Microbiota Influences Non-intestinal Related Autoimmune Diseases. Front Microbiol. 2018;9:432.

4. Caruso R, Lo BC, Núñez G. Host-microbiota interactions in inflammatory bowel disease. Nat Rev Immunol. 2020;20:411–26.

5. Benchimol EI, Bernstein CN, Bitton A, Murthy SK, Nguyen GC, Lee K, et al. The Impact of Inflammatory Bowel Disease in Canada 2018: A Scientific Report from the Canadian Gastro-Intestinal Epidemiology Consortium to Crohn's and Colitis Canada. J Can Assoc Gastroenterol. 2019;2 Suppl 1:S1–5.

6. GBD 2017 Inflammatory Bowel Disease Collaborators. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol. 2020;5:17–30.

7. Rizzello F, Spisni E, Giovanardi E, Imbesi V, Salice M, Alvisi P, et al. Implications of the Westernized Diet in the Onset and Progression of IBD. Nutrients. 2019;11. doi:10.3390/nu11051033.

8. Ricciuto A, Mack DR, Huynh HQ, Jacobson K, Otley AR, deBruyn J, et al. Diagnostic Delay Is Associated With Complicated Disease and Growth Impairment in Paediatric Crohn's Disease. J Crohns Colitis. 2021;15. doi:10.1093/ecco-jcc/jjaa197.

9. Noiseux I, Veilleux S, Bitton A, Kohen R, Vachon L, Guay BW, et al. Inflammatory bowel disease patient perceptions of diagnostic and monitoring tests and procedures. BMC Gastroenterol. 2019;19:1–11.

10. Armstrong D, Barkun AN, Chen Y, Daniels S, Hollingworth R, Hunt RH, et al. Access to specialist gastroenterology care in Canada: the Practice Audit in Gastroenterology (PAGE) Wait Times Program. Can J Gastroenterol. 2008;22:155–60.

11. Pittayanon R, Lau JT, Leontiadis GI, Tse F, Yuan Y, Surette M, et al. Differences in Gut Microbiota in Patients With vs Without Inflammatory Bowel Diseases: A Systematic Review. Gastroenterology. 2020;158:930–46.e1.

12. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784.

13. de Meij TGJ, de Groot EFJ, Peeters CFW, de Boer NKH, Kneepkens CMF, Eck A, et al. Variability of core microbiota in newly diagnosed treatment-naïve paediatric inflammatory bowel disease patients. PLoS One. 2018;13:e0197649.

885 14. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, et al. A microbial
886 signature for Crohn's disease. Gut. 2017;66:813–22.

887 15. Tedjo DI, Smolinska A, Savelkoul PH, Masclee AA, van Schooten FJ, Pierik MJ, et al. The
888 fecal microbiota as a biomarker for disease activity in Crohn's disease. Sci Rep. 2016;6:35216.

889 16. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al.
890 Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol.
891 2017;2:17004.

892 17. Clooney AG, Eckenberger J, Laserna-Mendieta E, Sexton KA, Bernstein MT, Vagianos K, et
893 al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal
894 intercontinental study. Gut. 2020. doi:10.1136/gutjnl-2020-321106.

895 18. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity
896 and IBD. FEBS Lett. 2014;588:4223–33.

897 19. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, et al. Integrative
898 analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals
899 exquisite inter-relationships. Microbiome. 2013;1:17.

900 20. Vázquez-Baeza Y, Gonzalez A, Xu ZZ, Washburne A, Herfarth HH, Sartor RB, et al. Guiding
901 longitudinal sampling in IBD cohorts. Gut. 2018;67:1743–5.

902 21. Liu T-C, Gurram B, Baldridge MT, Head R, Lam V, Luo C, et al. Paneth cell defects in Crohn's
903 disease patients promote dysbiosis. JCI Insight. 2016;1:e86907.

904 22. Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics
905 differentially classify disease state and treatment outcome in pediatric Crohn's disease.
906 Microbiome. 2018;6:13.

907 23. Ananthakrishnan AN, Luo C, Yajnik V, Khalili H, Garber JJ, Stevens BW, et al. Gut Microbiome
908 Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases.
909 Cell Host Microbe. 2017;21:603–10.e3.

910 24. Topçuoğlu BD, Lesniak NA, Ruffin MT 4th, Wiens J, Schloss PD. A Framework for Effective
911 Application of Machine Learning to Microbiome-Based Classification Problems. MBio. 2020;11.
912 doi:10.1128/mBio.00434-20.

913 25. Waljee AK, Lipson R, Wiitala WL, Zhang Y, Liu B, Zhu J, et al. Predicting Hospitalization and
914 Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning.
915 Inflamm Bowel Dis. 2017;24:45–53.

916 26. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome
917 definition re-visited: old concepts and new challenges. Microbiome. 2020;8:103.

918 27. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible,
919 interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol.
920 2019;37:852–7.

921 28. Hildebrand F, Tadeo R, Voigt AY, Bork P, Raes J. LotuS: an efficient and user-friendly OTU
922 processing pipeline. Microbiome. 2014;2:30.

923 29. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal
924 RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res.
925 2013;41 Database issue:D590–6.

926 30. Frioux C, Singh D, Korcsmaros T, Hildebrand F. From bag-of-genes to bag-of-genomes:
927 metabolic modelling of communities in the era of metagenome-assembled genomes. Comput
928 Struct Biotechnol J. 2020;18:1722–34.

929 31. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for
930 prediction of metagenome functions. Nature Biotechnology. 2020;38:685–8. doi:10.1038/s41587-
931 020-0548-6.

932 32. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine Learning SNP Based
933 Prediction for Precision Medicine. Front Genet. 2019;10:267.

934 33. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and
935 microbial differential abundance strategies depend upon data characteristics. Microbiome.
936 2017;5:27.

937 34. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are
938 Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017;8.
939 doi:10.3389/fmicb.2017.02224.

940 35. Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome
941 studies. PLoS Comput Biol. 2018;14:e1006102.

942 36. Wang Y, LêCao K-A. Managing batch effects in microbiome data. Briefings in Bioinformatics.
943 2019. doi:10.1093/bib/bbz105.

944 37. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group
945 differences may lead to exaggerated confidence in downstream analyses. Biostatistics.
946 2015;17:29–39.

947 38. Ma S, Shungin D, Mallick H, Schirmer M, Nguyen LH, Kolde R, et al. Population Structure
948 Discovery in Meta-Analyzed Microbial Communities and Inflammatory Bowel Disease.
949 doi:10.1101/2020.08.31.261214.

950 39. Song K, Wright FA, Zhou Y-H. Systematic Comparisons for Composition Profiles, Taxonomic
951 Levels, and Machine Learning Methods for Microbiome-Based Disease Prediction. Front Mol
952 Biosci. 2020;7:610845.

953 40. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis
954 of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link
955 with choline degradation. Nat Med. 2019;25:667–78.

956 41. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American
957 Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3.
958 doi:10.1128/mSystems.00031-18.

959 42. Estaki M, Pither J, Baumeister P, Little JP, Gill SK, Ghosh S, et al. Cardiorespiratory fitness
960 as a predictor of intestinal microbial diversity and distinct metagenomic functions. Microbiome.
961 2016;4:42.

962 43. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al.
963 Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature.
964 2019;569:655–62.

965 44. Doherty MK, Ding T, Koumpouras C, Telesco SE, Monast C, Das A, et al. Fecal Microbiota
966 Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease
967 Patients. MBio. 2018;9. doi:10.1128/mBio.02120-17.

968 45. Schirmer M, Denson L, Vlamakis H, Franzosa EA, Thomas S, Gotman NM, et al.
969 Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis
970 Patients Are Linked to Disease Course. Cell Host Microbe. 2018;24:600–10.e4.

971 46. Vogtmann E, Chen J, Amir A, Shi J, Abnet CC, Nelson H, et al. Comparison of Collection
972 Methods for Fecal Samples in Microbiome Studies. Am J Epidemiol. 2017;185:115–23.

973 47. Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-Tito AJ, Xu J, et al. Gut
974 Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. Curr
975 Biol. 2015;25:3161–9.

976 48. Hill-Burns EM, Debelius JW, Morton JT, Wissemann WT, Lewis MR, Wallen ZD, et al.
977 Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut
978 microbiome. Mov Disord. 2017;32:739–49.

979 49. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al.
980 Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun.
981 2015;6:6505.

982 50. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al.
983 Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15:R89.

984 51. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al.
985 Moving pictures of the human microbiome. Genome Biol. 2011;12:R50.

986 52. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High
987 resolution sample inference from amplicon data. doi:10.1101/024034.

988 53. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of
989 rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73:5261–7.

990 54. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-
991 species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research.
992 2014;42:D643–8. doi:10.1093/nar/gkt1209.

993 55. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, et al. 25 years of serving
994 the community with ribosomal RNA gene reference databases and tools. J Biotechnol.
995 2017;261:169–76.

996 56. Ma S. MMUPHin: Meta-analysis Methods with Uniform Pipeline for Heterogeneity in
997 Microbiome Studies. R package version 0.99.3. 2019.

998 57. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd
999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,
1000 NY, USA: ACM; 2016. doi:10.1145/2939672.2939785.

1001 58. Wang Y, Bhattacharya T, Jiang Y, Qin X, Wang Y, Liu Y, et al. A novel deep learning method
1002 for predictive modeling of microbiome data. Brief Bioinform. 2020. doi:10.1093/bib/bbaa073.

1003 59. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nat
1004 Commun. 2020;11:3514.

1005 60. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data - IEEE
1006 Conference Publication.
1007 https://ieeexplore.ieee.org/abstract/document/8949568?casa_token=GzthpwK9bOkAAAAA:vV-
1008 LF2CYOeiUi4xgtw_R1B0aAPaQWkUkgBpEYqac4bsB6OceUWdp2kgTuBRLMLAUdS6idoYz0H
1009 s. Accessed 27 Nov 2020.

1010 61. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1
1011 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6.

1012 62. Moustafa A, Li W, Anderson EL, Wong EHM, Dulai PS, Sandborn WJ, et al. Genetic risk,
1013 dysbiosis, and treatment stratification using host genome and gut microbiome in inflammatory
1014 bowel disease. Clin Transl Gastroenterol. 2018;9:e132.

1015 63. Davenport M, Poles J, Leung JM, Wolff MJ, Abidi WM, Ullman T, et al. Metabolic alterations
1016 to the mucosal microbiota in inflammatory bowel disease. Inflamm Bowel Dis. 2014;20:723–31.

1017 64. Karlsson I, Bostrom H. Handling Sparsity with Random Forests When Predicting Adverse
1018 Drug Events from Electronic Health Records. 2014 IEEE International Conference on Healthcare
1019 Informatics. 2014. doi:10.1109/ichi.2014.10.

1020 65. Bang S, Yoo D, Kim S-J, Jhang S, Cho S, Kim H. Establishment and evaluation of prediction
1021 model for multiple disease classification based on gut microbial data. Sci Rep. 2019;9:1–9.

1022 66. Pawlowsky-Glahn V, Egozcue JJ. Compositional data and their analysis: an introduction.
1023 Geological Society, London, Special Publications. 2006;264:1–10.

1024 67. Bailey K, Sokal RR, James Rohlf F. Biometry: The Principles and Practice of Statistics in
1025 Biological Research (2nd ed.). Journal of the American Statistical Association. 1982;77:946.
1026 doi:10.2307/2287349.

1027 68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
1028 seq data with DESeq2. Genome Biol. 2014;15:550.

1029 69. Taub MA, Bravo HC, Irizarry RA. Overcoming bias and systematic errors in next generation
1030 sequencing data. Genome Med. 2010;2:1–5.

1031 70. Kim YS, Unno T, Kim BY, Park MS. Sex Differences in Gut Microbiota. World J Mens Health.
1032 2020;38:48–60.

1033 71. Radjabzadeh D, Boer CG, Beth SA, van der Wal P, Kiefte-De Jong JC, Jansen MAE, et al.
1034 Diversity, compositional and functional differences between gut microbiota of children and adults.
1035 Sci Rep. 2020;10:1040.

1036 72. Durbán A, Abellán JJ, Jiménez-Hernández N, Ponce M, Ponce J, Sala T, et al. Assessing gut
1037 microbial diversity from feces and rectal mucosa. Microb Ecol. 2011;61:123–33.

1038 73. Mas-Lloret J, Obón-Santacana M, Ibáñez-Sanz G, Guinó E, Pato ML, Rodriguez-Moranta F,

et al. Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample. Sci Data. 2020;7:92.

74. Nagao-Kitamoto H, Kamada N. Host-microbial Cross-talk in Inflammatory Bowel Disease. Immune Netw. 2017;17:1–12.

75. Glassner KL, Abraham BP, Quigley EMM. The microbiome and inflammatory bowel disease. J Allergy Clin Immunol. 2020;145:16–27.

76. Chen L, Wang W, Zhou R, Ng SC, Li J, Huang M, et al. Characteristics of fecal and mucosa-associated microbiota in Chinese patients with inflammatory bowel disease. Medicine . 2014;93:e51.

77. Wyatt A, Kellermayer R. PCR Based Fecal Pathogen Panel Testing Should Be Interpreted with Caution at Diagnosis of Pediatric Inflammatory Bowel Diseases. Ann Clin Lab Sci. 2018;48:674–6.

78. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

79. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 2016;352:565–9.

80. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457:480–4.

81. Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, et al. Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. mSystems. 2018;3. doi:10.1128/mSystems.00188-17.

82. Sun S, Jones RB, Fodor AA. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. Microbiome. 2020;8:46.

83. De Filippis F, Pasolli E, Ercolini D. Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease. Curr Biol. 2020. doi:10.1016/j.cub.2020.09.063.

84. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010;10:278–91.

85. Parker HS, Corrada Bravo H, Leek JT. Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ. 2014;2:e561.

86. Liu Y, Wang T, Zhou B, Zheng D. Robust integration of multiple single-cell RNA sequencing datasets using a single reference space. Nat Biotechnol. 2021. doi:10.1038/s41587-021-00859-x.

87. Bryrup T, Thomsen CW, Kern T, Allin KH, Brandslund I, Jørgensen NR, et al. Metformin-induced changes of the gut microbiota in healthy young men: results of a non-blinded, one-armed intervention study. Diabetologia. 2019;62:1024–35.

88. Lewis JD. The utility of biomarkers in the diagnosis and therapy of inflammatory bowel

disease. Gastroenterology. 2011;140:1817–26.e2.

89. Ma C, Battat R, Parker CE, Khanna R, Jairath V, Feagan BG. Update on C-reactive protein and fecal calprotectin: are they accurate measures of disease activity in Crohn's disease? Expert Rev Gastroenterol Hepatol. 2019;13:319–30.

90. E Penna FGC, Rosa RM, da Cunha PFS, de Souza SCS, de Abreu Ferrari M de L. Faecal calprotectin is the biomarker that best distinguishes remission from different degrees of endoscopic activity in Crohn's disease. BMC Gastroenterol. 2020;20:35.

91. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163:1079–94.

92. Nguyen VQ, Jiang D, Hoffman SN, Guntaka S, Mays JL, Wang A, et al. Impact of Diagnostic Delay and Associated Factors on Clinical Outcomes in a U.S. Inflammatory Bowel Disease Cohort. Inflamm Bowel Dis. 2017;23:1825–31.

93. Zaharie R, Tantau A, Zaharie F, Tantau M, Gheorghe L, Gheorghe C, et al. Diagnostic Delay in Romanian Patients with Inflammatory Bowel Disease: Risk Factors and Impact on the Disease Course and Need for Surgery. J Crohns Colitis. 2016;10:306–14.

94. Vavricka SR, Spigaglia SM, Rogler G, Pittet V, Michetti P, Felley C, et al. Systematic evaluation of risk factors for diagnostic delay in inflammatory bowel disease. Inflamm Bowel Dis. 2012;18:496–505.

95. Vadstrup K, Alulis S, Borsi A, Gustafsson N, Nielsen A, Wennerström ECM, et al. Cost Burden of Crohn's Disease and Ulcerative Colitis in the 10-Year Period Before Diagnosis-A Danish Register-Based Study From 2003-2015. Inflamm Bowel Dis. 2020;26:1377–82.

96. Park KT, Ehrlich OG, Allen JI, Meadows P, Szigethy EM, Henrichsen K, et al. The Cost of Inflammatory Bowel Disease: An Initiative From the Crohn's & Colitis Foundation. Inflamm Bowel Dis. 2020;26:1–10.

97. Zhang W, Wong CH, Chavannes M, Mohammadi T, Rosenfeld G. Cost-effectiveness of faecal calprotectin used in primary care in the diagnosis of inflammatory bowel disease. BMJ Open. 2019;9:e027043.

Tables

Table 1. Overview of 15 datasets used to compare the effect of different features, data preprocessing methods, and machine learning models on IBD classification performance. Available metadata (age, sex, BMI), disease activity, and medication use) is provided for each dataset. Blank spaces indicate that the respective metadata was not available for the dataset's samples. The following abbreviations are used: female (F), male (M), and other (O).

Study	Accession	Disease Type	Number of Samples	Sample Type		Age		Sex			BMI		Disease Activity		Medications		
				Stool	Biopsy	Mean	SD	F	M	O	Mean	SD	Active	Rem-ission	Biologics	Immuno-suppresants	5-ASA
American Gut	PRJEB11419	Control	1279	1279	0	46.5	12.2	600	595	1	23.3	2.7					
CVDF	PRJNA308319	Control	39	39	0	25.4	4.2	15	24		24.0	2.9					
		CD	219	219	0	12.0	2.9	87	132								
GEVERSC	PRJEB13680	Control	28	28	0	12.3	3.5	10	18								
		UC	37	37	0	11.8	3.6	22	15								
		CD	689	166	523	19.6	14.2	312	377						15	31	51
GEVERSM	PRJEB13679	Control	320	7	313	14.0	9.8	157	163								
		UC	268	106	162	24.9	17.5	121	147						2	5	52
GLS	PRJEB23009	CD	340	340	0	30.2	9.0	215	102		25.7	7.2	43	297	145	74	15
		Control	335	335	0	48.6	14.4	152	166		32.8	8.4					
HMP	ibdmdb.org	CD	66	0	66	23.5	13.0	32	34								
		Control	43	0	43	28.7	22.0	20	23								
		UC	36	0	36	27.7	17.4	20	16								
MUC	PRJNA317429	CD	35	0	35	14.5	3.5	13	22								
		Control	47	0	47	11.9	3.4	21	25								
PRJNA418765	PRJNA418765	CD	589	589	0	40.4	13.2	332	257		26.4	6.6		589	416		
PRJNA436359	PRJNA436359	UC	1178	917	261	12.6	3.3	582	596				875	303			
QIITA10184	PRJEB13895	Control	962	962	0												
QIITA10342	PRJEB13619	Control	58	58	0	43.2	15.3				31.0	7.5					
QIITA10567	PRJEB14674	Control	133	133	0	70.3	8.6				28.3	5.7					
QIITA1448	PRJEB13051	Control	23	23	0												
QIITA2202	PRJEB6518	Control	516	516	0	29.6	4.8	516									
QIITA550	PRJEB19825	Control	467	467	0	32.8	0.5	131	336								
Total			**7707**	**6221**	**1486**			**3358**	**3048**	**1**			**918**	**1189**	**578**	**110**	**118**

Table 2. Model performance for different sample types and patient demographics.

Samples with available metadata were categorized into groups based on the collection method or the patient's specific demographic group based on sex, age, and BMI. Predictive performance for all combinations of taxonomic features, compositional normalizations, zero-centering batch effect reduction, and non-linear models were included in the analysis. Logistic regression was performed to assess the performance differences within each sample and demographic group while adjusting for the remaining covariates. **** indicates p-value < 0.0001, and * indicates p-value < 0.05. Coefficient refers to the corresponding independent variable's coefficient for the logistic regression function and SE refers to the standard error of the coefficient.

Group	Variable	Coefficient	SE
Sample Type	Biopsy (vs. Stool)	-0.44 *	0.2
Life Stage	Adult (vs. Pediatric)	1.39 ****	0.18
BMI Stratification	BMI <30 (vs. BMI > 30)	-0.85 ****	0.19
Sex	Female (vs. Male	-0.02	0.18
IBD Type	CD (vs. UC)	0.05	0.18

Table 3. Top three data processing and model pipelines for classifying IBD samples.

Three combinations which appeared most frequently when all models were sorted by F1 score, accuracy, AUC, or MCC.

Features	Normalization	Batch Reduction	Model	F1 Score	Accuracy	AUC	MCC
Genus	ILR	Zero-Centering	XGBoost	83.67	87.92	87.9	74.3
Genus	CLR	Zero-Centering	XGBoost	82.96	87.41	87.3	73.2
Genus	ILR	Zero-Centering	Random Forest	82.66	87.4	86.9	72.9

Figures

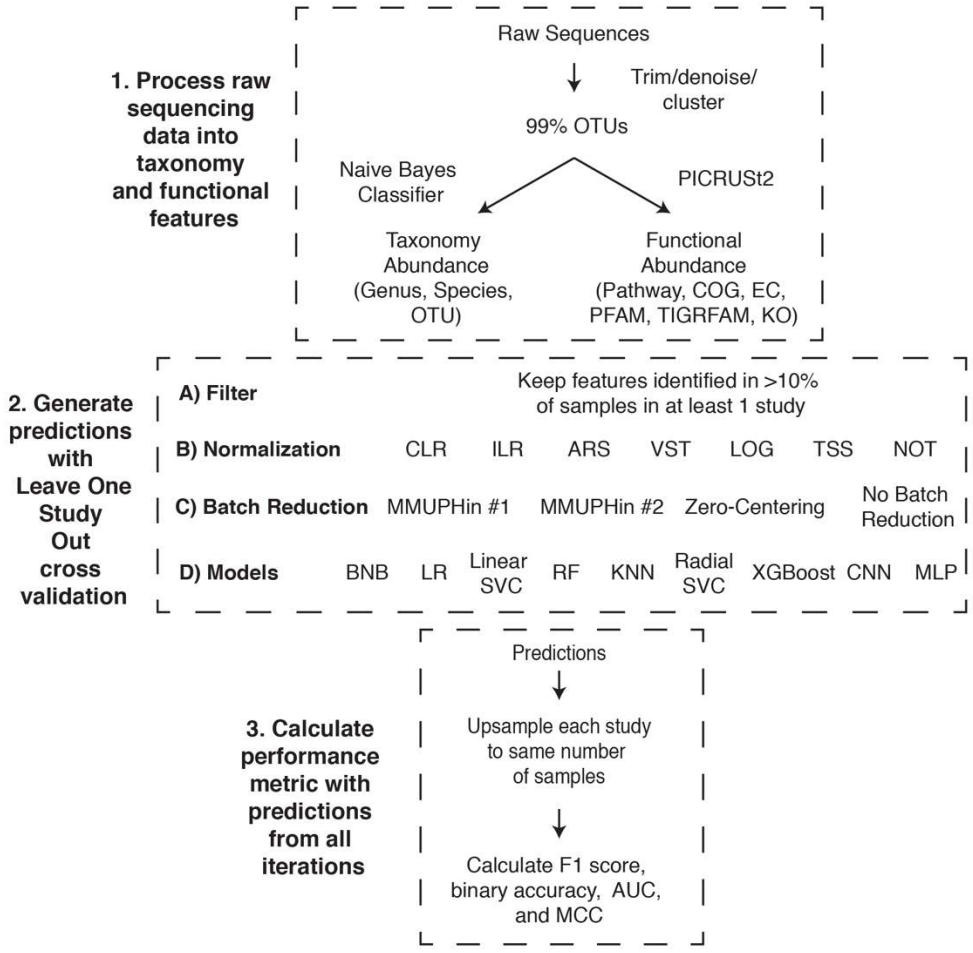


Figure 1. Leave-one-dataset-out cross-validation pipeline.

The experiments comprised three different stages to go from raw sequence files to the performance metrics. 1) Raw sequences were processed with Dada2 or Deblur and close-reference clustered into OTUs at 99% identity. The OTUs were classified to taxonomy at 99% identity with QIIME2 and used to infer functional profiles with PICRUSt2. 2) Generating predictions for the 15 iterations of our LODO cross validation consisted of all possible combinations of the listed filtering method, normalization methods, batch effect reduction methods, and models. 3) The predictions from each iteration were combined and the number of samples from each dataset up sampled to 100 000 prior to calculating the performance metrics. The descriptions of acronyms and abbreviations are the following: Clusters of Orthologous Groups of proteins (COG), Kyoto

Encyclopedia of Genes and Genomes (KEGG) orthologs (KO), Enzyme Commission (EC), Pfam protein domain (PFAM), TIGR protein family (TIGRFAM) and MetaCyc pathways (pathway), centered log-ratio (CLR), isometric log-ratio (ILR), arcsine square root transformation (ARS), variance stabilizing transformation (VST), log transformation (LOG), total sum scaling (TSS), no normalization (NOT), Bernoulli Naive Bayes (BNB), logistic regression (LR), linear support vector machine (Linear SVC), random forest (RF), K nearest neighbours (KNN), radial support vector machine (Radial SVC), eXtreme Gradient Boosting (XGBoost), convolutional neural network (CNN), multilayer perceptron (MLP).

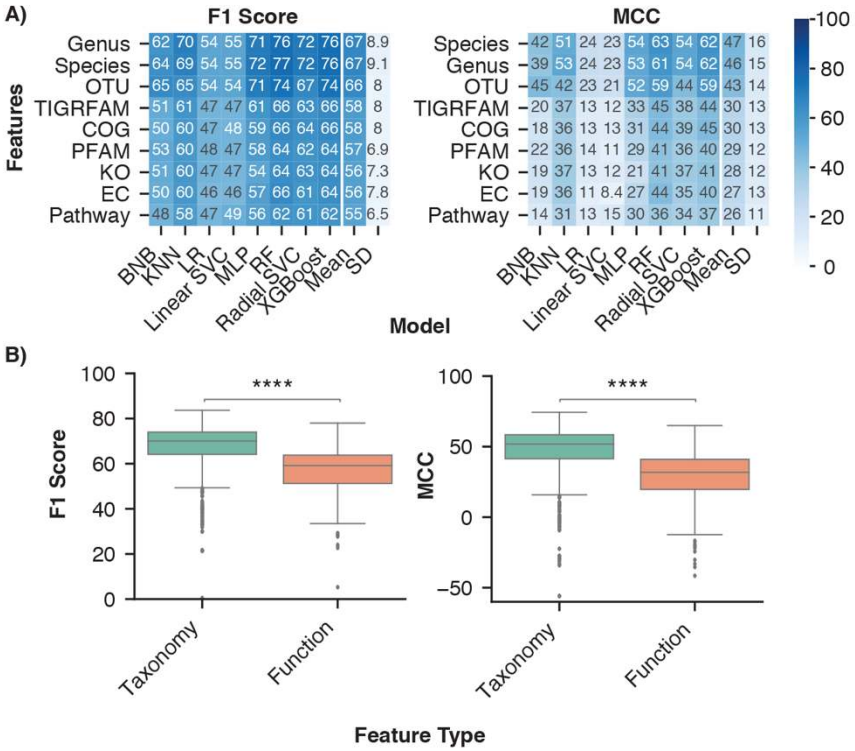


Figure 2. Optimal disease classification of microbiome samples obtained with taxonomic features.

A) Average performance of the taxonomy and functional feature sets for each ML model architecture. Rows were sorted in descending order by the mean column followed by the standard deviation (SD) column. B) Distribution of performance metrics for taxonomy and functional features across all normalization, batch effect reduction, and model combinations. Independent Mann-Whitney U tests were performed to compare aggregate performance of taxonomy and functional features. The analysis was limited to normalization (ILR, CLR, VST, ARS, LOG, TSS, NOT) and batch effect reduction (no batch reduction or Zero-Centering) methods that were performed on all feature sets. **** indicates p-value < 0.0001.

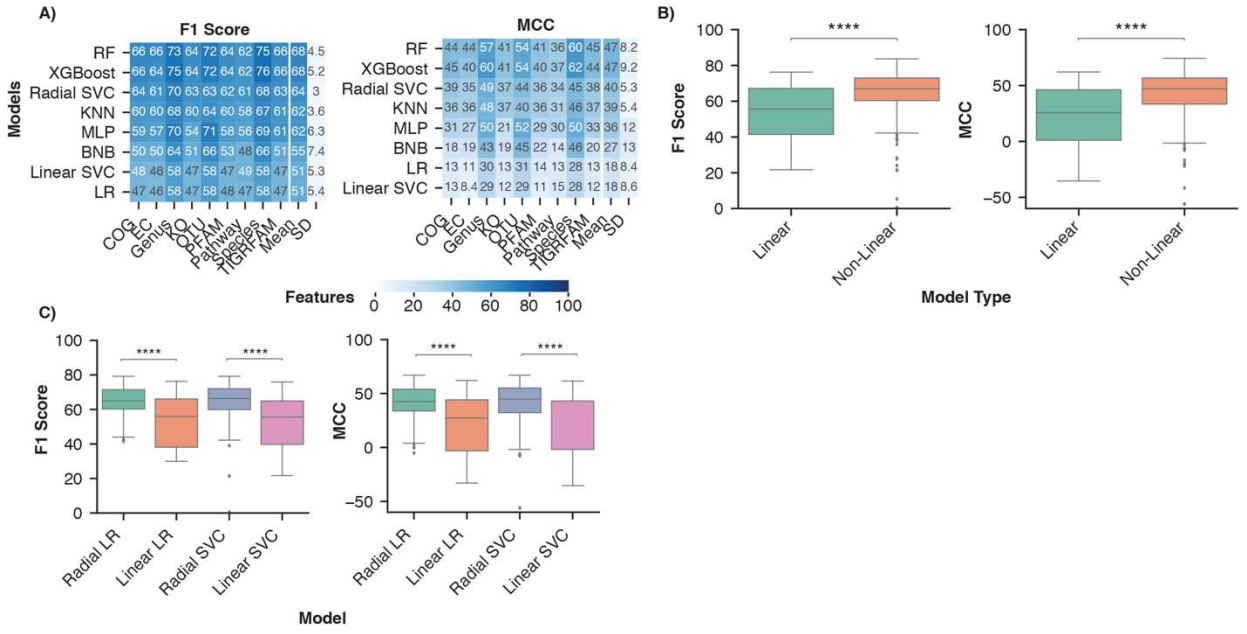


Figure 3. Non-linear models are better suited to identify decision boundaries between control and IBD samples than linear models.

A) Average model performance for each feature set across normalization and batch effect reduction methods. Rows were sorted in descending order by mean followed by the standard deviation of performance across all feature sets. B) Distribution of performance of non-linear (RF, MLP, KNN, XGBoost, radial SVC) and linear (BNB, Linear SVC, LR) models. Independent Mann-Whitney U-tests were performed to compare each performance metric. The analysis was limited to normalization (ILR, CLR, VST, ARS, LOG, TSS, NOT) and batch effect reduction (no batch effect reduction or zero centering) methods performed on all feature types. C) Distribution of classification performance with the non-linear and linear variations of logistic regression and support vector machines across all feature sets. A Mann-Whitney U test with Bonferroni correction was performed to compare the linear and non-linear variation of each model respectively. **** indicates p-value < 0.0001.

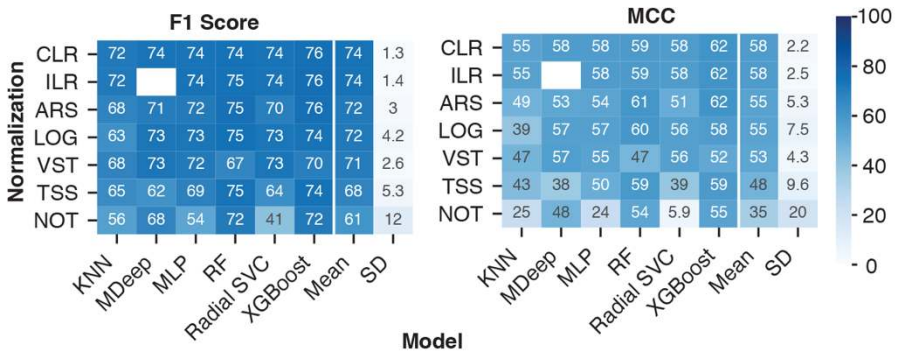


Figure 4. Compositional normalization methods lead to the highest model performance for IBD classification.

Average model performance with each normalization method across all batch effect reduction methods. Performance following data processing with all pairwise combinations of the normalization methods (ILR, CLR, LOG, ARS, VST, TSS and NOT) and batch effect reduction methods (No batch reduction, MMUPHin #1, MMUPHin #2, and Zero-Centering) were included. Rows were sorted in descending order by the mean and standard deviation of each performance metric across the non-linear models. No analysis was performed for MDeep paired with ILR as the ILR normalized values no longer map directly to a feature, therefore removing the phylogenetic structure required for MDeep.

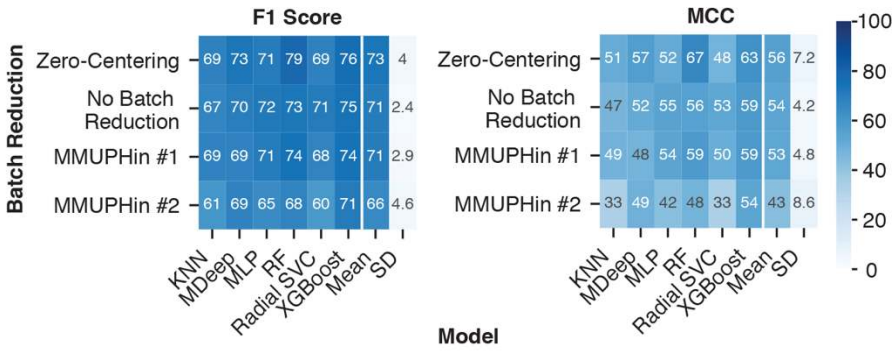


Figure 5. Batch effect reduction with the naive zero-centering method improved IBD classification.

Average performance of each batch effect reduction method across all combinations of normalization methods, taxonomic features, and non-linear ML models. Performance following data processing with all pairwise combinations of the normalization methods (ILR, CLR, LOG, ARS, VST, TSS and NOT) and batch effect reduction methods (No batch reduction, MMUPHin #1, MMUPHin #2, and Zero-Centering) were included. Rows were sorted in descending order by the mean and standard deviation of performance across all non-linear models.

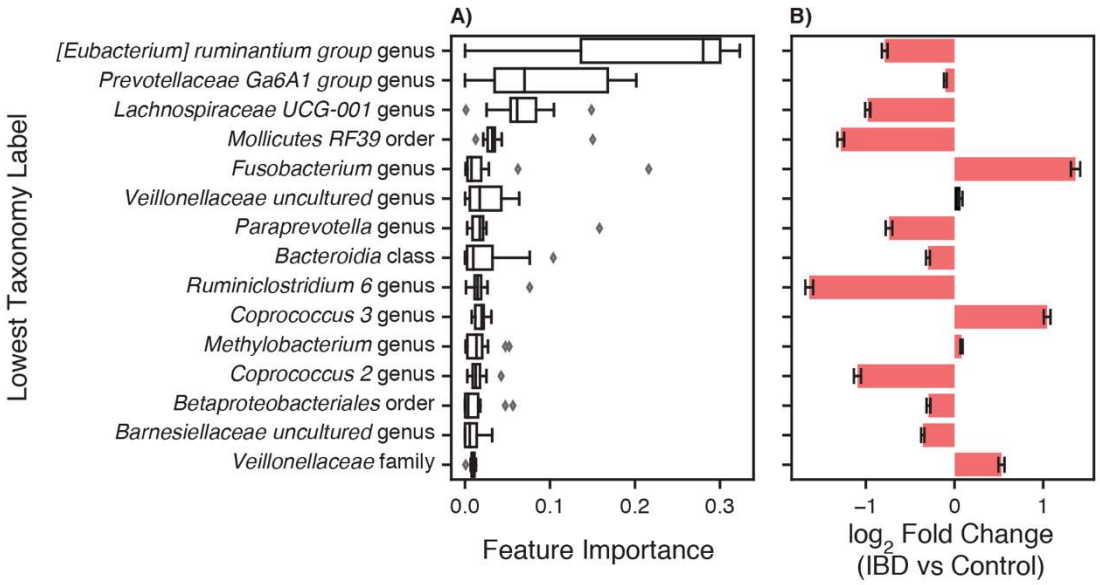


Figure 6. Features with greatest contribution to IBD classification by an XGBoost classifier. A) An XGBoost classifier was trained with CLR normalized genus abundance features and zero-centered batch effect reduction for fifteen LODO iterations. The features' gain values for each iteration were extracted and sorted by the mean gain across all iterations. The lowest classification rank for each feature was used as the label for the corresponding bar. B) Changes in taxonomy abundance between control samples and those from patients with IBD. Bars represent the fold change ± the standard error determined with Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC). Red indicates a significant fold change between IBD and control samples ($p < 0.05$) and black indicates non-significant fold change.

Supplemental Tables

Supplemental Table 1. Model performance on sample and patient demographics.

Samples were grouped by the five different metadata categories and the classification performance with the indicated metric determined for control and CD or UC (depending on the metadata group). Coefficient refers to the corresponding independent variable's coefficient for the logistic regression function and SE refers to the standard error of the coefficient. **** indicates p-value < 0.0001, *** indicates p-value < 0.001, ** indicates p-value < 0.01, and * indicates p-value < 0.05.

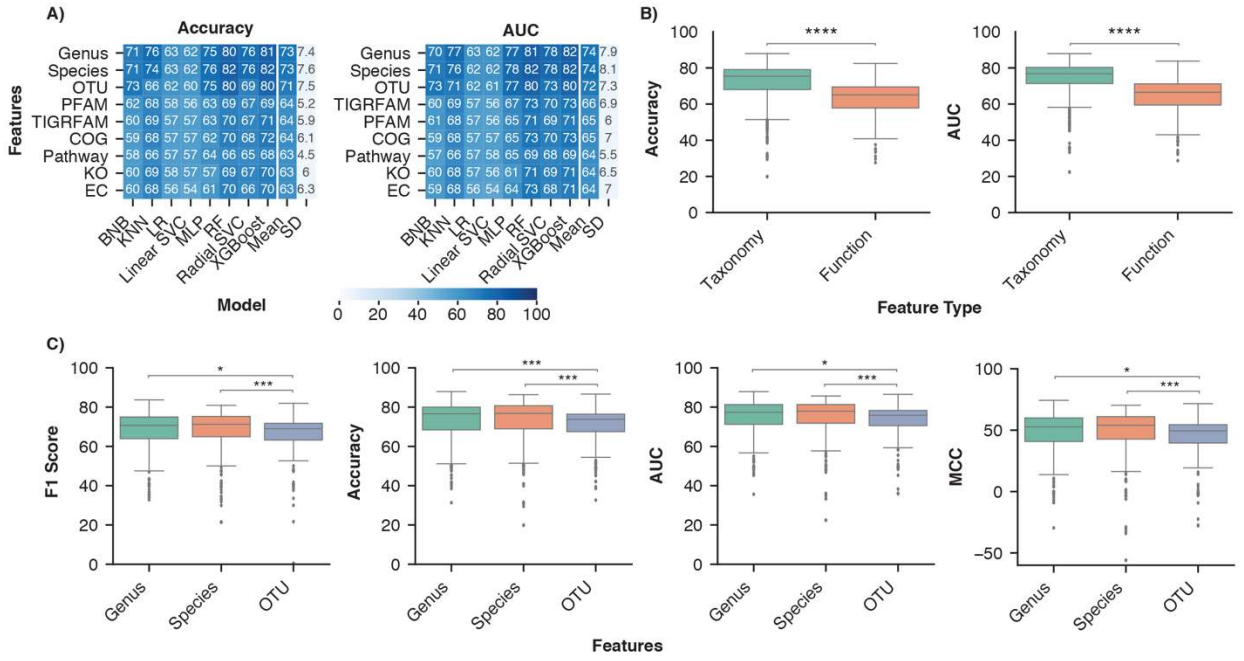
Metric	Group	Variable	Coefficient	SE
F1 Score	Sample Type	Biopsy (vs. Stool)	-0.33	0.21
F1 Score	Life Stage	Adult (vs. Pediatric)	0.47 **	0.18
F1 Score	BMI Stratification	BMI <30 (vs. BMI > 30)	0.49 **	0.19
F1 Score	Sex	Female (vs. Male	0.31	0.19
F1 Score	IBD Type	CD (vs. UC)	0.54 **	0.18
AUC	Sample Type	Biopsy (vs. Stool)	-0.18	0.2
AUC	Life Stage	Adult (vs. Pediatric)	1.25 ****	0.17
AUC	BMI Stratification	BMI <30 (vs. BMI > 30)	-0.17	0.18
AUC	Sex	Female (vs. Male	0.1	0.18
AUC	IBD Type	CD (vs. UC)	0.07	0.17
Accuracy	Sample Type	Biopsy (vs. Stool)	-0.7 ***	0.2
Accuracy	Life Stage	Adult (vs. Pediatric)	1.03 ****	0.18
Accuracy	BMI Stratification	BMI <30 (vs. BMI > 30)	0.41 *	0.19
Accuracy	Sex	Female (vs. Male	0.2	0.19
Accuracy	IBD Type	CD (vs. UC)	0.17	0.18

Supplemental Table 2. Overview of QIIME2 processing for 15 microbiome datasets.

Samples were collected from the listed ENA accession, with only samples corresponding to individuals in North America retained. Trim length was used as input for the trunc_len parameter, forward trim as the trim_left input for single end read and trim_left_f for paired-end reads, and reverse trim as the trim_left_r input for paired-end reads in the python API for QIIME2's Dada2 plugin.

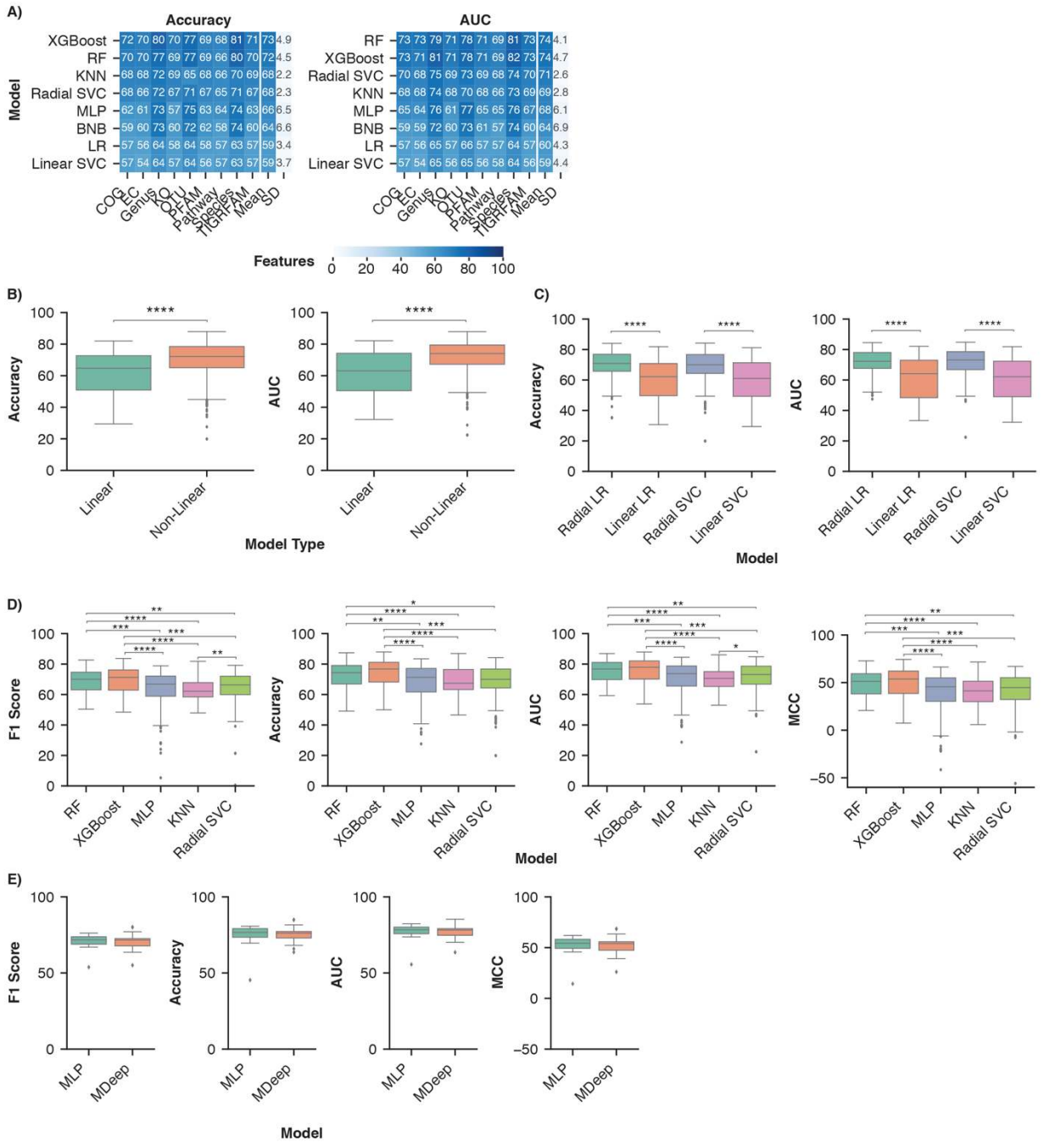
Study ID	ENA Accession	Hypervariable Region	Trim Length	Forward Trim	Reverse Trim
American Gut	ERP012803	V4	124	0	0
CVDF	PRJNA308319	V3-V4	290	40	40
GEVERSC	PRJEB13680	V4	174	0	0
GEVERSM	PRJEB13679	V4	174	0	
GLS	PRJEB23009	V4	99	0	
HMP	ibdmdb.org	V4	249	0	0
MUC	PRJNA317429	V4	174	19	21
PRJNA418765	PRJNA418765	V4	245	0	3
PRJNA436359	PRJNA436359	V4	170	0	3
QIITA10184	PRJEB13895	V4	120	0	
QIITA10342	PRJEB13619	V4	100	0	
QIITA10567	PRJEB14674	V4	99	0	
QIITA1448	PRJEB13051	V4	99	0	
QIITA2202	PRJEB6518	V4	99	0	
QIITA550	PRJEB19825	V4	149	0	

Supplemental Figures



Supplemental Figure 1. Greater classification of IBD samples with taxonomic features than functional features.

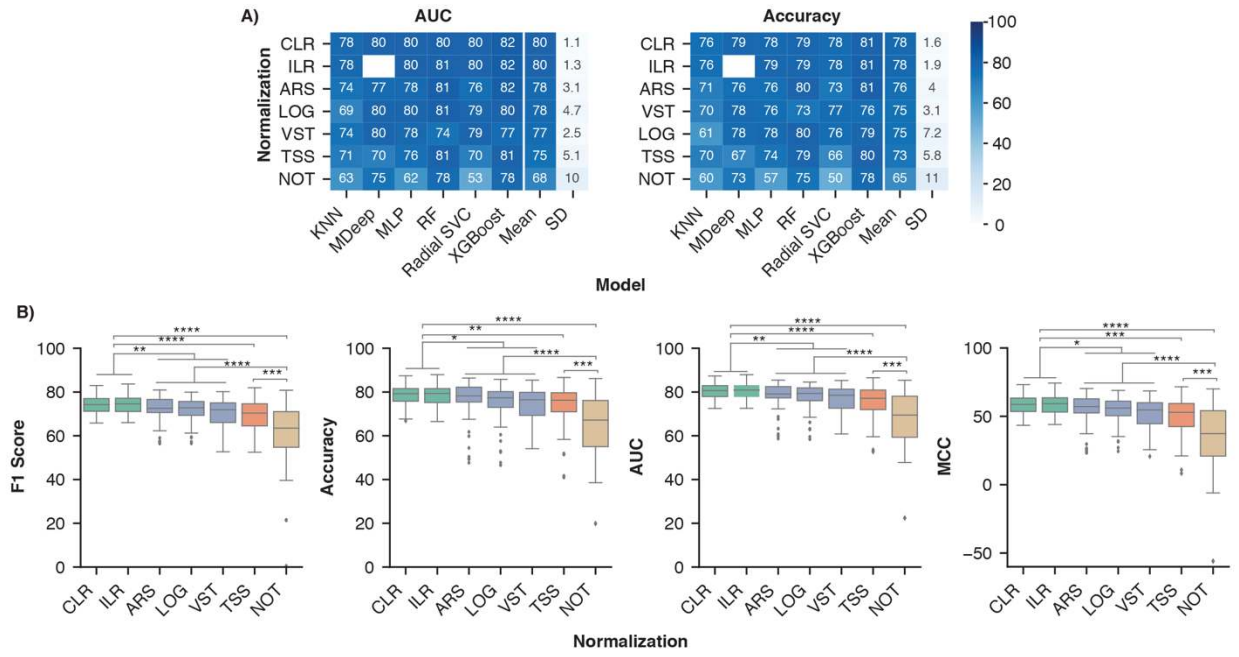
A) Average performance for each type of feature across different model architectures. Rows were sorted in descending order by the mean column followed by the standard deviation (SD) column. B) Distribution of performance metrics across all normalization, batch effect reduction, and model combinations. Independent Mann-Whitney U-tests were performed to compare aggregate performance measured by each metric of taxonomy and functional features. C) Comparison of classification performance with the three taxonomic feature sets. All pairwise comparisons were performed with a Mann-Whitney U-test followed by Bonferroni correction and the significant comparisons are indicated. The analysis was limited to normalization (ILR, CLR, VST, ARS, LOG, TSS, NOT) and batch effect reduction (No batch reduction or Zero-Centering) methods that were performed on all feature sets. **** indicates p-value < 0.0001, *** indicates p-value < 0.001, * indicates p-value < 0.05.



Supplemental Figure 2. Greater IBD classification performance with non-linear than linear ML models.

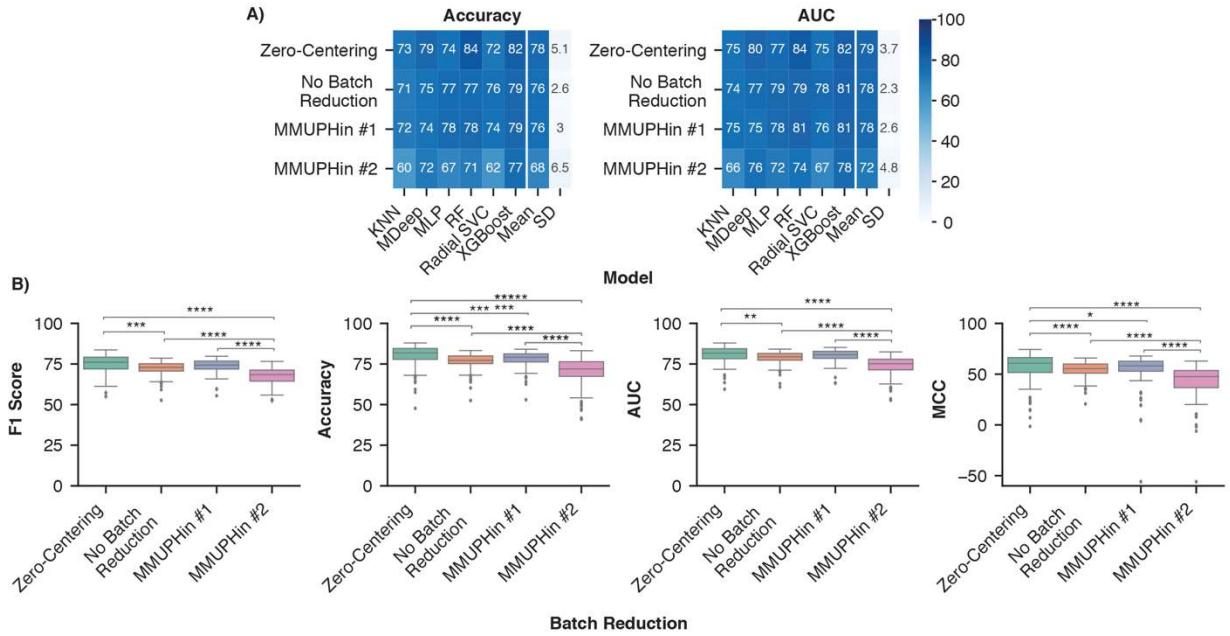
A) Average model performance for each feature set across normalization and batch effect reduction methods. B) Distribution of performance of non-linear and linear models. Independent Mann-Whitney U-tests were performed to compare each performance metric. Analysis was limited

to datasets preprocessed using normalization (ILR, CLR, VST, ARS, LOG, TSS, NOT) and batch effect reduction (No batch reduction or Zero-Centering) methods performed on all feature types. C) Distribution of classification performance with the non-linear and linear variations of logistic regression and support vector machines across all feature sets. A Mann-Whitney U test with Bonferroni correction was performed to compare the linear and non-linear variation of each model respectively. D) Comparison of IBD classification performance between the non-linear models. All pairwise comparisons were performed by Mann-Whitney U test with a Bonferroni correction and the significant comparisons were labelled. E) Comparison of two neural network architectures: the convolutional neural network MDeep or a MLP. A Mann-Whitney U test was used to compare each performance metric and the significant comparisons were labelled. **** indicates p-value < 0.0001, *** indicates p-value < 0.001, ** indicates p-value < 0.01, and * indicates p-value < 0.05.



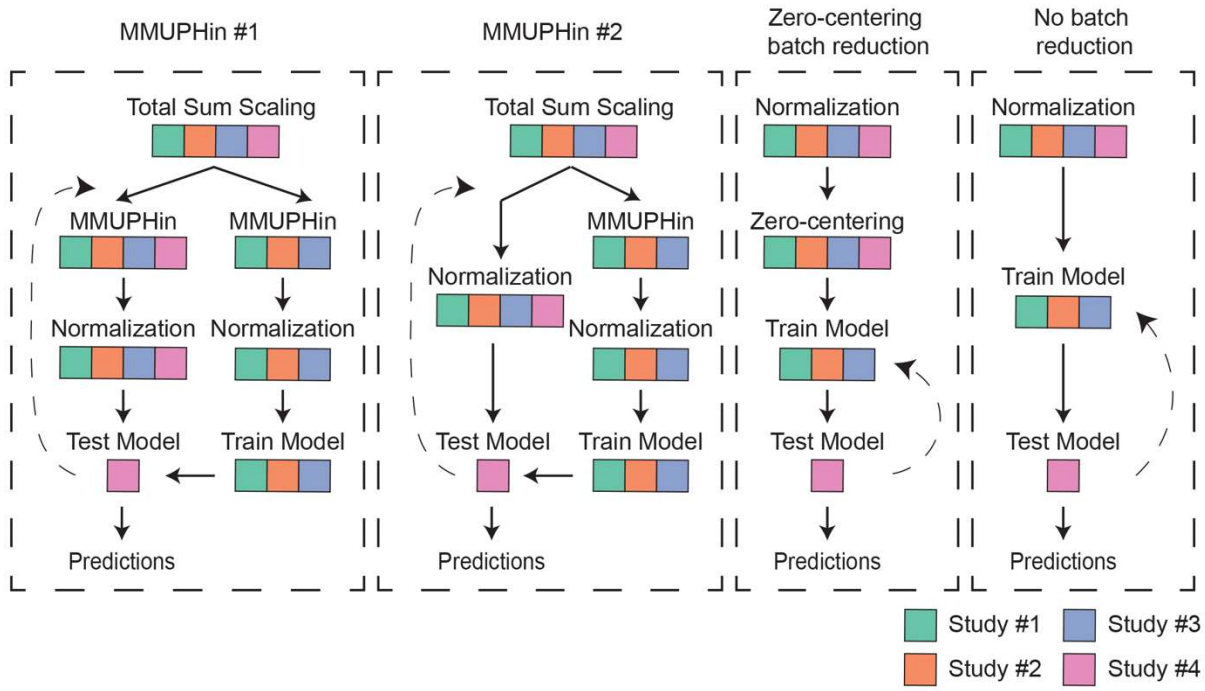
Supplemental Figure 3. Compositional normalization methods lead to the highest model performance for IBD classification.

A) Average model performance with each normalization method across all batch effect reduction methods. B) Comparing the effect of different classes of normalization methods. The compositional category consists of CLR and ILR (green), variance/distribution modifiers consists of VST, ARS, and LOG (blue), scaling consists of TSS (orange), and no normalization consists of NOT (brown). All pairwise combinations were compared with a Mann-Whitney U test with a Bonferroni correction and the significant comparisons labelled. **** indicates p-value < 0.0001, *** indicates p-value < 0.001, ** indicates p-value < 0.01, and * indicates p-value < 0.05.



Supplemental Figure 4. Removing batch effects with zero-centering improved IBD classification.

A) Batch effect reduction methods sorted by average performance across all combinations normalization methods, taxonomic features, and non-linear ML models. B) Comparing the effect of different batch effect reduction methods on classification of IBD samples. All pairwise combinations were compared with a Mann-Whitney U test and the significant comparisons were labelled. **** indicates p-value < 0.0001, *** indicates p-value < 0.001, ** indicates p-value < 0.01, and * indicates p-value < 0.05.



1362

1363 **Supplemental Figure 5. LODO cross validation pipeline for different batch effect reduction**

1364 **methods.**

1365 Pipeline performance (combinations of a normalization method, batch effect reduction method,

1366 and ML model) was determined with LODO cross validation. Four variations were used to account

1367 for different requirements in the batch effect reduction step to ensure the training and test set

1368 were independent. The coloured boxes under the text indicate which studies the corresponding

1369 step was performed with. The diagram illustrates how a single dataset (illustrated by the different

1370 colours) is removed from the dataset the model is trained with and then tested on. The dashed

1371 arrow indicates the step returned to in each iteration, with a different dataset removed in each

1372 iteration.

1373