Multi-view integration of microbiome data for predicting host disease and identifying disease-associated features

Efrat Muller[1], Itamar Shiryan[1], Elhanan Borenstein[1,2,3,*]

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel;

[2] Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel;

[3] Santa Fe Institute, Santa Fe, NM, USA;

* Corresponding author

Abstract

Machine learning (ML) has become a widespread strategy for studying complex microbiome signatures associated with disease. To this end, metagenomics data are often processed into a single "view" of the microbiome, such as its taxonomic (species) or functional (gene) composition, which in turn serves as input to such ML models. When further omics are available, such as metabolomics, these can be analyzed as additional complementary views. Following training and evaluation, the resulting model can be explored to identify informative features, generating hypotheses regarding underlying mechanisms. Importantly, however, using a single view generally offers relatively limited hypotheses, failing to capture simultaneous shifts or dependencies across multiple microbiome layers that likely play a role in microbiome-host interactions.

In this work, inspired by the broad domain of *multi-view learning*, we constructed an integrated ML analysis pipeline using multiple microbiome views. We specifically aimed to investigate the impact of various integration approaches on the ability to predict disease state based on multiple microbiome-related views, and to generate biological insights. Applying this pipeline to data from 25 case-control metagenomics studies, we found that multi-view models typically result in performances that are comparable to the best-performing single view, yet, provide a mixed set of informative features from different views, while accounting for dependencies and links between them. To further enhance such models, we developed a new framework termed MintTea, based on sparse generalized canonical correlation analysis, to identify *multi-view modules* of features, highlighting shared trends in the data expressed by the different views. We showed that this framework identified multiple modules that were both highly predictive of the disease state, and exhibited strong within-module associations between features from different views. We accordingly advocate for using multi-view models to capture multifaceted microbiome signatures that likely better reflect the complex mechanisms underlying microbiome-disease associations.

Introduction

The use of machine learning (ML) in human microbiome studies, for predicting disease state or other host phenotype based on microbiome features, has been rapidly increasing in popularity and scale[1–5]. The motivation for such studies typically involves evaluating the overall association strength of the microbiome with the host phenotype[1,4,6,7], using model explainability methods to identify a subset of important features and generate new biological hypotheses[8–10], or developing new diagnostic tools, risk score calculators[9,11–14], microbiome-based recommendation systems[15,16], and even forensic tools[17]. Indeed, a recent study showed that ML can distinguish group differences even when samples cannot be well-separated based on common ecological distance metrics, further demonstrating the utility of ML approaches in this domain[1].

Such ML modelling efforts generally start with the processing of raw metagenomic data (as well as other omics, when available) into structured feature tables, followed by the application of specifically tailored ML pipelines that typically include a feature selection step, model construction, and a careful performance evaluation[3]. Clearly, the entire processes, therefore, involves multiple choices the researcher must make, which, in turn, could dramatically impact the study outcome[1,4,5,18].

One specifically dramatic choice in developing such ML models is the microbiome representation that will be used, i.e., the type of features that will serve as input for the ML pipeline. Different representations capture different aspects of the microbiome, and can be seen as complementary "views" through which the microbiome can be studied. Raw shotgun metagenomics, for example, can be processed into several different such views: Researchers can choose to focus on taxonomic profiling (and at varying taxonomic levels) using tools such as Kraken2[19], mOTUs2[20] or MetaPhlAn3[21] for taxonomy assignment and abundance estimation. They can also use functional profiling, estimating the abundances of genes or pathways in the metagenome, as generated by tools such as HUMAnN3[21] or MetaLAFFA[22]. Additional profiles that could potentially be inferred based on shotgun data include strain markers[23], single nucleotide variation profiles[20], microbial growth rates[24], microbial genomic structural variants[25], and others. When additional omics are available, further aspects of the human microbiome and its interaction with the host can be used for modelling. Focusing on the gut microbiome, the addition of fecal metabolomics, for example, can provide valuable insight into the metabolic environment in the gut, thus incorporating complementary

information into the model. Below, we refer to these different types of data as different *views*, although in the literature, they may be interchangeably referred to as different *omics*, *modalities*, *representations*, or *feature sets*.

Notably, not all views necessarily offer the same informative value, and several studies have shown that different views may be more or less suitable for different classification tasks[4,5,18,26,27]. Moreover, different views are commonly analyzed independently, or in a cross-view univariate manner using, for example, correlation networks[28–31]. In the context of a classification problem, however, it is reasonable to assume that more sophisticated multivariate integration of different views may improve both model performance and downstream biological hypothesis generation. Indeed, several recent reviews have highlighted the benefits of multi-view/multi-omic ML modelling across different biological and clinical domains[32–34], and successful multi-omic classification models have been introduced, for example, in breast cancer therapy prediction[35], early diagnosis of Parkinson's disease[36], and COVID-19 severity prediction[37]. In the microbiome domain, one recent study integrated taxonomic profiles and host transcriptomics data to identify multi-omic signatures that are either shared or specific across different gastrointestinal diseases[38].

Here, we focus on multi-view learning – a ML approach for prediction tasks where the input data is composed of several different views, often with different numbers of features and characteristics. Multi-view learning methods can roughly be categorized as either "early integration", where views are first concatenated into a single (wide) table before training, "intermediate integration" where some prior transformation is applied to all views to uncover intra- and inter-view structures before modelling, or "late integration" where models are trained for each view independently and a final ensemble model is designed to combine specific predictions to a final prediction. As expected, each approach carries different benefits and different chalanges[39]. In this study, we accordingly aimed to explore and promote the use of multi-view learning for microbiome research, presenting an integrative analysis of multi-view data for human disease classification. We hypothesized that mechanisms underlying a certain microbiome-disease association may often involve various taxa, functions, and metabolites, and that these features will, in turn, exhibit coordinated patterns that such advanced integration methods could potentially identify. To comprehensively assess the use of multi-view learning in microbiome research, we collected data from 25 diverse case-control cohorts with available shotgun metagenomics data, 7 of which further included metabolomics data, and for which several different "views" can be obtained. We compared prediction accuracy of models trained

on single-view data to concatenated multi-view data (the "early integration" approach), and examined different feature selection approaches to appropriately address the high dimensionality of the data. We then developed an "intermediate integration" approach to identify modules of features from all views that jointly shift in disease, thus reflecting both interactions between views and association with disease. We showed that some of the identified modules are indeed in agreement with previous observations concerning the role of the microbiome in disease, and provided a complete catalog of identified multi-view modules across all datasets analyzed. This work thus serves as a proof of concept for the benefit of integrating multiple microbiome data types in a single ML pipeline for disease classification, and for going beyond naïve, early-integration, approaches.

Results

Overview

To examine different multi-view integration approaches across diverse microbiome datasets, we obtained 25 different datasets of shotgun metagenomics sequencing. All datasets are from case-control studies, spanning multiple health conditions and diseases (Supplementary Table S1). For each dataset, we generated 3 different feature sets based on the metagenomic data, namely: a taxonomic profile ("T"), a gene family profile ("G"), and a pathway-level profile ("P") (see Methods, Figure 1A). For 7 of the 25 datasets, metabolomics data was also available, and we hence additionally included a metabolite profile ("M") for these datasets. We specifically focused initially on 4 datasets, describing cases-control studies of ulcerative colitis (UC) and Crohn's disease (CD)[40], late-stage colorectal cancer (CRC)[41], and liver cirrhosis[42]. We then also describe result highlights for all of the additional datasets.

In addition to analyzing these multi-view data without integrating the different profiles, we implemented two different integration approaches (Figure 1B), with a two-pronged objective in mind: First, optimizing disease classification based on multiple different microbiome views, and second, generating biological hypotheses that capture the interactions between the different microbiome views and between these views and the disease state. We followed common terminology in this domain and labeled the two approaches an "early integration" and "intermediate integration", where the former refers to a naïve concatenation of different views into a single joint table ("T+G+P" or "T+G+P+M"), and the latter refers to a joint analysis of

interactions between views and between each view and the phenotype, ultimately identifying multi-view modules of features with a concordant disease-associated behavior.

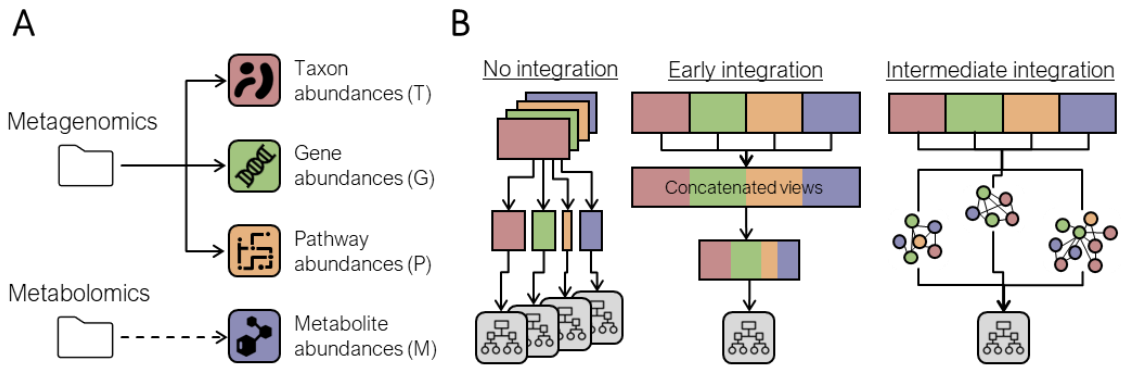

Figure 1: Multi-view microbiome data and integration approaches. (A) Data types used in this analysis. Throughout the manuscript, taxonomic, gene, pathway and metabolite abundances are labeled with the letters 'T', 'G', 'P', and 'M', respectively. (B) Different integration approaches carried out using multiple data types.

An early integration approach: Single-view vs. multi-view disease prediction

To first assess the information captured in single vs. multi-view feature sets, we implemented a ML pipeline with random forest (RF) classifiers to predict the disease state of the host based on microbiome features (Figure 2A). This pipeline included multiple commonly-used steps, such as filtration of rare features, clustering of highly correlated features, and feature selection (see Methods for full details). Feature selection, specifically, was used within the cross-validation procedure and before model training, aiming to address the substantially high number of features, simplify training, and discard non-informative features (Methods and Supplementary Table S2). Importantly, though feature selection is often focused on optimizing model performance, here we intentionally implemented and tested a few different "all-relevant" feature selection methods[43], whose goal is to also retain all informative (and potentially biologically relevant) features, even in the presence of redundancy (see Supplementary Note 1 for further details and a list of the feature selection methods tested). Notably, we found that feature selection markedly enhanced our ability to predict disease state across multiple datasets and feature sets (Supplementary Figure S1A). We also examined the stability of each feature selection method – an important property of feature selection that is

often overlooked[44] – by comparing the selected features when the method is applied to different subsets of the data. We found that different methods varied significantly in their stability (Supplementary Figure S1B-C), and that higher stability was, perhaps not surprisingly, strongly correlated with higher model AUCs (Pearson correlation > 0.7, $P < 10^{-7}$ across all feature selection methods).

We then ran this pipeline once for each feature set independently ("single-view" models), and once using a concatenated table of all feature sets (the "multi-view" model). Receiver operating characteristic (ROC) area under the curve (AUC) was averaged over cross-validation iterations for each dataset and each feature set (See Methods).

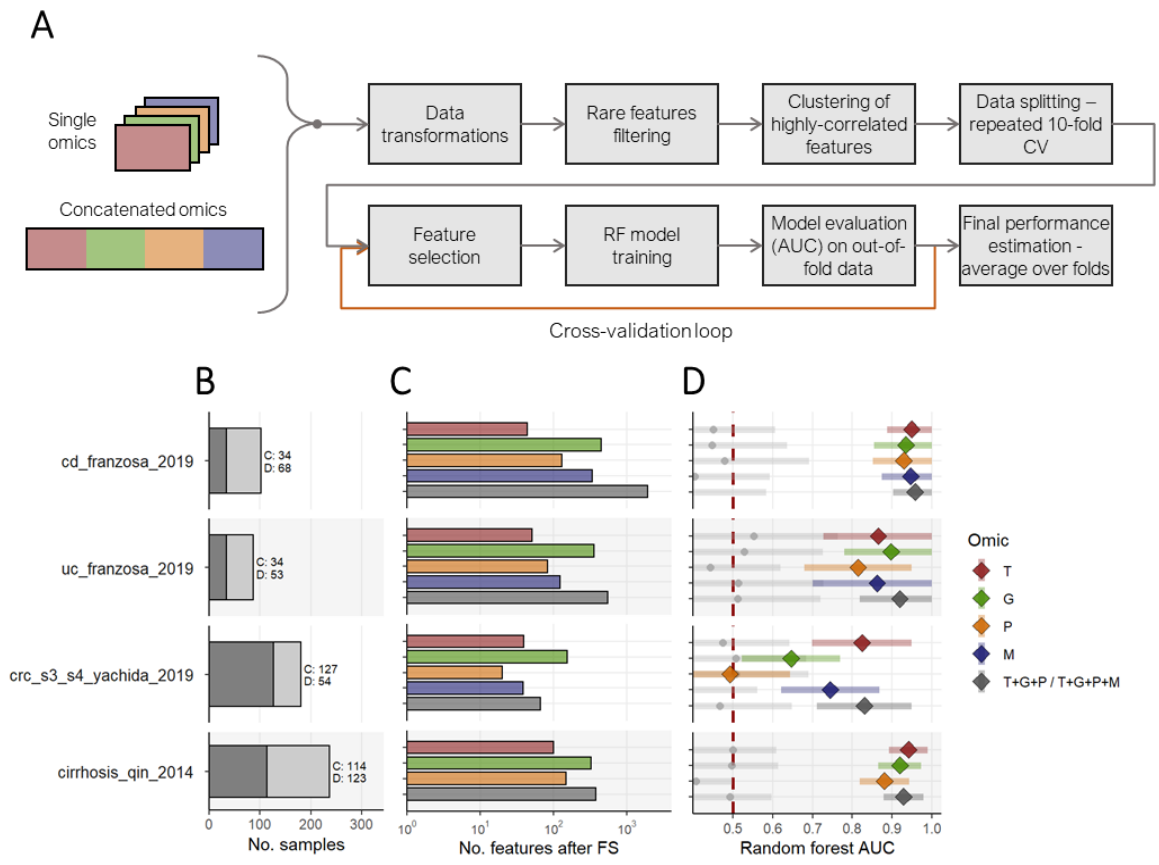

Figure 2: Single- and multi-view machine learning modelling of host diseases. (A) A diagram of a machine learning pipeline for microbiome data. Model evaluation is carried out through a repeated cross-validation procedure. Final estimations of model performance are averaged over repeats and folds. (B) Sample sizes for 4 of the datasets used in this study. Dark and light grey indicate the number of subjects in the "control" study group (noted: 'C'), and the "disease" group (noted 'D'), respectively. See Figure S1 for equivalent details about additional metagenomic datasets analyzed. (C) Number of features selected by the feature selection step, per dataset and per feature set ('view') used. (D) Mean AUC (and standard deviations) of Random Forest classifiers,

per dataset and per feature set used. Grey dots indicate performance of null models, where label values were shuffled prior to training. FS: Feature selection.

We found that for all 4 datasets, at least one of the single-/multi-view models resulted in an AUC > 0.8, with 3 datasets reaching an AUC > 0.9, indicating a substantial disease-associated microbiome signature in IBD, advanced CRC, and liver cirrhosis (Figure 2D), in agreement with original studies and several reports on additional cohorts[45–48]. Interestingly, however, comparing in each dataset the accuracy of the single-view models to that of the multi-view model, we found that the multi-view model is generally comparable to the best single-view model, but that it does not significantly outperform it (Supplementary Figure S2; ANOVA and paired T-tests, see Methods). This finding likely indicates that the information captured in the various views often largely overlaps, introducing redundancy into the model. We also observed that among the single-view models, different views performed best for different diseases and different studies, though overall, taxonomic profiles tended to perform best (Supplementary Table S3). Of the remaining 21 datasets, 13 achieved an AUC > 0.7 in at least one model and were therefore retained for downstream analysis (Supplementary Figure S3, Supplementary Table S2).

To benchmark our entire pipeline we also compared our results to previously published studies which applied similar machine learning pipelines to the same datasets[1,4,6,41]. Notably, although several major differences exist between the pipelines in terms of modelling approaches, feature transformation, and evaluation strategy, our pipeline resulted in comparable AUC's to other pipelines across most datasets (Supplementary Figure S4). We also compared our model to null models, where study group labels were shuffled, confirming that all AUC values obtained by the true models were significantly higher than those obtained by the null models (Figure 2D, Supplementary Figure S2C).

Lastly, we also ran our pipeline on an additional set of 12 16S rRNA gene sequencing datasets from the MicrobiomeHD collection (Methods). Of these, 9 achieved an AUC > 0.7 in at least one model (single- or multi-view; Supplementary Figure S5, Supplementary Table S2). We note that taxonomic models significantly outperformed other models across all 16S datasets, which is perhaps non-surprising given that the two functional profiles (i.e. G and P) in these datasets are *inferred* based on phylogeny[49] rather than assayed directly from sequencing data.

An early integration approach: Identifying disease-associated features

Having selected relevant features and training a final model, we were also interested in identifying the top informative features, and compare these top features when using single-view vs. multi-view models (Supplementary Table S4). To measure feature importance, we used a permutation-based approach proposed by Altmann et al (2010)[50]. We further implemented and applied several additional feature importance scoring approaches (as well as simple univariate statistical test), confirming that they provide highly correlated scores, though may differ slightly in their ranking of top features (Supplementary Figure S6).

Overall, we found that the top ranked features typically coincided with previous reports on diverse cohorts. In both IBD models, for example, known butyrate-producing species such as *Roseburia hominis*, *Gemmiger formicillis*, *Faecalibacterium prausnitzii*, and *Coprococcus catus* were decreased in IBD patients compared to healthy controls, and were among the top important features in both single- and multi-view models (Figure 3, Supplementary Table S4)[51,52]. Indeed, reduced abundances of butyrate-producing species has been well-established as an IBD risk factor[53,54]. Multiple other taxonomic (e.g. *Eubacterium siraeum*, *Dorea formicigenerans*, *Oscillibacter sp. 57_20*, *Clostridium sp. CAG 242*, *Alistipes shahii*), functional (e.g. pyrimidine deoxyribonucleoside salvage, transfer RNA charging, methylglyoxal degradation), and metabolite (e.g. chenodeoxycholate / deoxycholate, urobilin, adrenic acid, nicotinic acid) features were in agreement with differential analysis results on various IBD cohorts worldwide[31,55–57]. We also note that CD is characterized by higher absolute importance scores, as illustrated by darker shades in Figure 3 (as well as generally higher AUCs), likely reflecting the more prominent shift of the microbiome and metabolome in CD compared to UC[58].

Interestingly, examining the top 10 features in the multi-view models of both CD and UC (Figure 3), we found that they span both taxonomic features, gene families, and metabolites, indicating that the microbial signature in IBD may be best captured by a mixed set of features (also see Supplementary Figure S6B,D). Also, we note that among these top 10 multi-view features, the number of metabolite features is somewhat higher, likely reflecting the fact that these features originated from metabolomic assays which are completely independent from the other metagenomic-based views and thus less redundant compared to them.

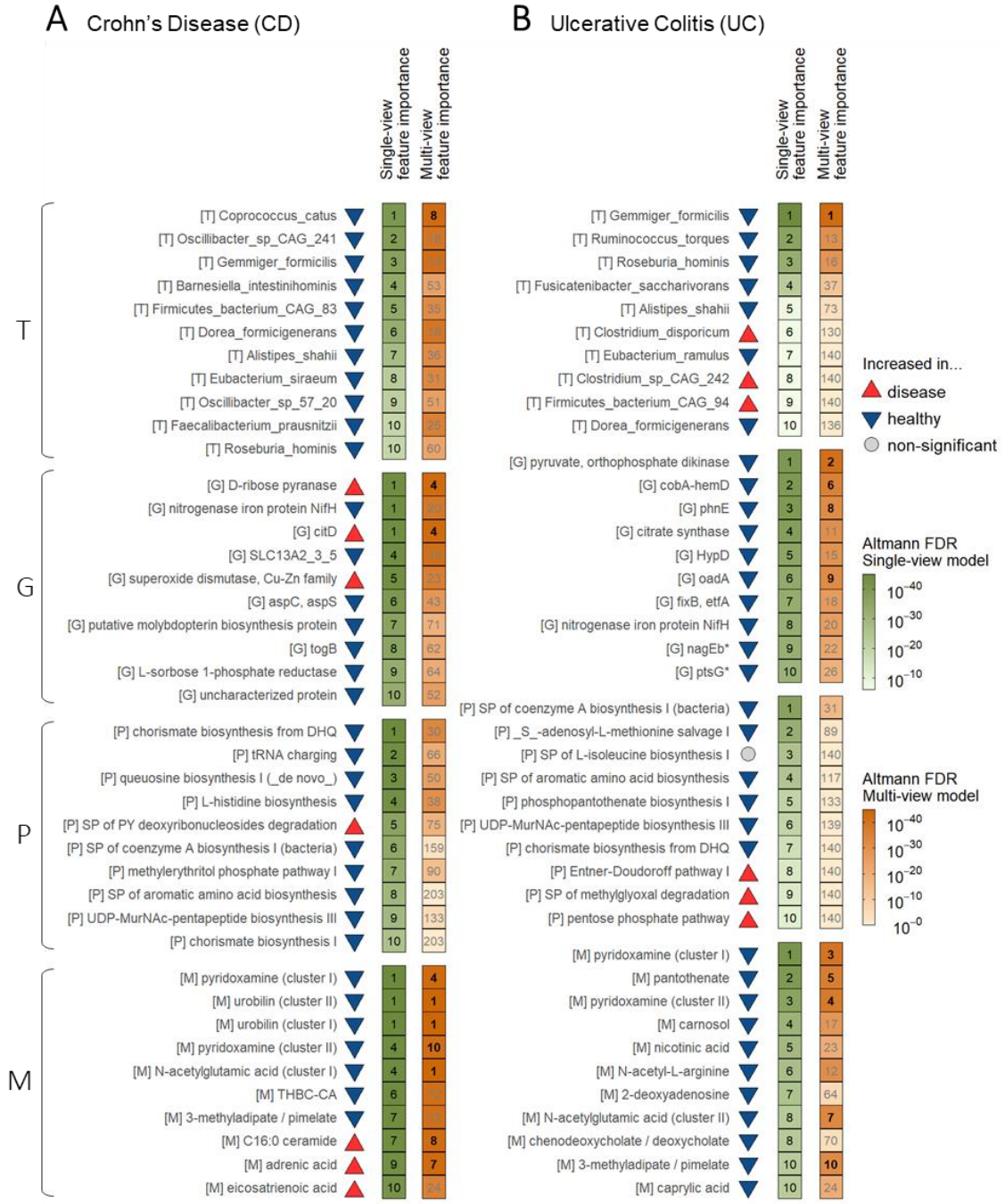

Figure 3: Multiple views of feature importance in predicting IBD. Feature importance analysis is presented for CD (a) and UC (b), using Altmann's method with the single-view machine learning models (green shades), and Altmann's method with the multi-view machine learning models (brown shades; Methods). A union of the top 10 single-/multi-view contributors are shown for each dataset, grouped by feature type (T/G/P/M prefix). Numbers represent the feature's importance rank in either the single- or multi-view model (ties were all assigned the minimal rank within the tied group of values). Ranks up to 10 are colored black, while higher ranks are in grey. Symbols next to each feature name indicate whether that feature was significantly elevated (up triangle) or decreased (down triangle) in individuals with the disease (grey circles indicate no significant difference). A

star at the end of a feature name indicates that this is a cluster representative (see Supplementary Table S8).

CD: Crohn's Disease; UC: Ulcerative Colitis; SP: Superpathway; THBC-CA: 1,2,3,4-tetrahydro-beta-carboline-3-carboxylic acid; HypD: hydrogenase expression/formation protein HypD; citD: citrate lyase subunit gamma (acyl carrier protein); aspC, aspS: nondiscriminating aspartyl-tRNA synthetase; togB: oligogalacturonide transport system substrate-binding protein; SLC13A2_3_5: solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2/3/5; ptsG: glucose PTS system EIICBA or EIICB component; cobA-hemD: uroporphyrinogen III methyltransferase / synthase; fixB, etfA: electron transfer flavoprotein alpha subunit; nagEb: N-acetylglucosamine PTS system EIIB component; phnE: phosphonate transport system permease protein; oadA: oxaloacetate decarboxylase (Na+ extruding) subunit alpha; MurNAc: N-acetylmuramyl; DHQ: 3-Dehydroquinic acid; PY: pyrimidine

As another example, in the late-stage CRC data, we again found well-established CRC associated species included among the top important taxonomic features in both the taxonomy-only and multi-view models, including several sub-species of the oral commensal *Fusobacterium nucleatum*[59,60] (which were previously shown to colonize tumors[61]), as well as *Peptostreptococcus stomatis*[60], *Peptostreptococcus anaerobius*[62], and *Parvimonas micra*[60], (Supplementary Table S4). Of the measured metabolites, top contributing metabolites included fumarate (elevated in CRC), a known oncometabolite[63], and alpha-Aminoadipate, previously detected at higher levels in CRC tumor samples compared to adjacent mucosa[64]. While both of these metabolites are likely not microbiome-related, other top contributing metabolites such as dimethylglycine (elevated in CRC), previously suggested as an early marker for CRC[65], hydroxyproline (depleted in CRC), known for intestinal health benefits[66], and urocanate (elevated in CRC), again in line with previous reports[67], are potentially driven by microbiome activity[65,68]. Among functional microbiome features, though overall performance of these models was poor, genes involved in tryptophan biosynthesis (e.g. K01696, K00766; depleted in CRC patients) were identified as main contributors, in agreement with accumulating evidence showing the role of tryptophan metabolism in CRC development[69].

An intermediate integration approach: Identifying multi-view disease-associated modules

While the identification of important features in the early integration multi-view models above provided intriguing insights into microbial signatures of disease across different layers, it may fail to account for crucial information concerning the interaction between these layers or

concordant disease-associated shifts in different views. In this section, we therefore resorted to an alternative integration approach, termed *intermediate integration*, in which each view is first transformed into an intermediate representation that captures both within- and between- view interactions and can potentially be used for disease classification. The transformation process itself can further take the disease state of each sample into account, thus extracting representations that highlight both interactions across different views and between views and the disease.

Following this approach and inspired by a previously introduced method termed DIABLO[70], we therefore developed an intermediate integration framework, termed *MintTea* (Multi-view INTegration Tool for microbiomE Analysis), based on sparse generalized canonical correlation analysis (sGCCA). MintTea's analysis aims to identify "multi-view modules" – sets of features from multiple different views that are both strongly associated with the disease and with each other (Methods). Briefly, the framework receives two or more feature tables describing different views (or omics) for the same set of samples, and a label for each sample (e.g., healthy vs. disease). It then searches for a *sparse* linear transformation per feature table that yields maximal correlations between the respective latent variables, as well as between these variables and the label. This first set of latent variables is referred to as the first component, and additional components can be computed iteratively such that each new component is orthogonal to the previous ones. This sGCCA variant, which encodes the label as an additional view, was ran using DIABLO's implementation of this variant[70]. Importantly, to identify subsets that are robust to small changes in the input data, the entire process is repeated multiple times on random data subsets (e.g., 90% of the samples), and the final *multi-view modules* are defined as sets of features from different views that consistently co-occurred in the same component. To assess the overall association of each identified module with the disease, MintTea further tests the combined predictive power of each module, using the first principle component (PC) calculated on all module's features and subsequently computing its AUC. Lastly, MintTea compares the true AUC to AUCs obtained using randomly selected modules (of the same size and distribution over views), effectively serving as null models (Methods). The final outcome for each dataset is therefore multiple multi-view modules, each capturing features from all views that are highly associated within view, between views, and between views and phenotype. The entire pipeline, including data preparations steps, is depicted in Figure 4A.

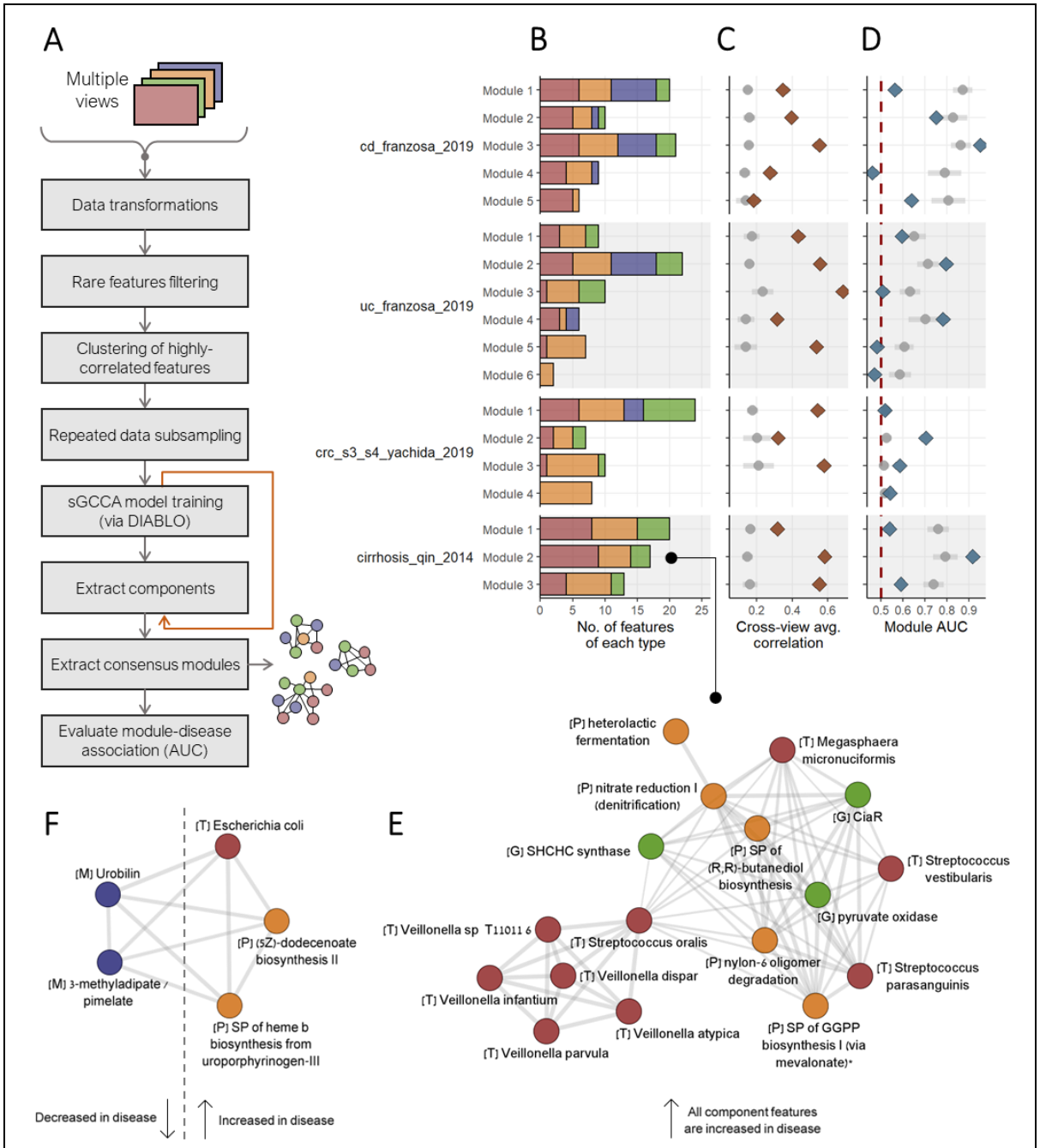

Figure 4: Identifying disease-associated multi-view modules via MintTea intermediate integration framework. (A) A diagram of the MintTea pipeline. (B-D) Properties of the multi-view modules obtained by applying MintTea to several key datasets, including (B) the number of features in each module, stratified by feature type, (C) cross-view correlation per module, calculated as the average pairwise correlation between features of different views (with grey points and lines indicating the average and standard deviation of correlations obtained from random modules of the same size), and (D) AUC per module, calculated using the first principal component values against disease state of each sample (again compared to random modules, in grey). (E) A multi-view module predictive of liver cirrhosis. (F) A sub-module of mixed features, shared between a CD-predictive module (module 3) and a UC-predictive module (module 2). Node colors represent feature type (T/G/P). Edges

connect features that appeared together in a sGCCA component over 80% of data subsampling iterations.
sGCCA: Sparse Generalized Canonical Correlation Analysis; SP: super-pathway; GGPP: geranylgeranyl diphosphate; SHCHC: 2-succinyl-6-hydroxy-2,4- cyclohexadiene-1-carboxylate; CiaR: two-component system, OmpR family, response regulator CiaR.

Applying MintTea to the 4 datasets previously described, we found 5 modules in the CD data, 6 in UC, 4 in late-stage CRC, and 3 in cirrhosis (Figure 4B), the majority of which capturing strong correlations (Pearson's $r > 0.3$) between features from different views (Figure 4C). Moreover, of these modules, 1, 2, 1, and 1, respectively, were further highly predictive of the disease state, exhibiting an AUC that was both above 0.7 and significantly higher than random modules (see Methods, Figure 4C, Supplementary Table S5). Intriguingly, however, we found that the features included in the highly predictive modules noted above did not necessarily overlap with those ranked highest by the single-view and early integrated multi-view models discussed in the previous section. Specifically, only 28.6%-77.8% of the features in each module were also significant contributors in the *single-view* models. Similarly, 14.3%-77.8% of module features were significant contributors in the *multi-view* models. This may confirm the ability of the intermediate integration approach to not only consider single feature's association with the disease, but also take into account interactions between features, balancing between these two objectives.

Figure 4E presents one example of such a multi-view module associated with liver cirrhosis (metagenomics only, no metabolomics). Indeed, multiple features in this module are related to bacterial contribution and response to intestinal oxidative stress (OS), driven by OS in the liver, a key driver of cirrhosis[71–74]. The bacterial species driving this module are typically found in the human oral cavity, including several *Streptococcus*, *Veillonella*, and *Megasphaera* species (both commensal and opportunistic pathogens)[75], all of which previously found to be increased in stool samples of individuals with liver diseases[76–78], as well as other gastrointestinal disorders[79–81]. We also observe that several functional features (genes and pathways) in the module are *Streptococci*-related and/or oxidative stress related, including the pyruvate oxidase gene, shown to be possessed by multiple *Streptococcus* species and produce ROS[82], CiaR, a signaling mechanism highly conserved in *Streptococci* that has a central role in the regulation of stress tolerance[83], and nitrate reduction, which promotes oxidative stress and has been confirmed in almost all of the *Streptococcus* and *Veillonella* species appearing in the module[84].

We hypothesize that *Veillonella* species take part in the module mostly due to their strong co-abundance patterns with various Streptococcus species. Indeed, a recent study found that *Streptococcus* and *Veillonella* species were co-abundant across multiple large-scale cohorts[85], also in line with known metabolic interactions between them[86]. *Veillonella* species have also been shown to reduce oxidative microenvironments, thus facilitating the growth of *Streptococci* and others[87]. Other features such as the (R,R)-butanediol biosynthesis superpathway are most likely included in the module as they are primarily driven by *Streptococcus* species, as implied by HUMAnN3's stratified output. Overall, this module suggests a potential multi-layer stress-related mechanism of specific bacterial species that is highly associated with liver cirrhosis.

We further checked whether there are significant overlaps between modules from the four different datasets and disease states (Supplementary Table S7). One significant overlap (Fisher's exact test FDR value $7 \cdot 10^{-7}$) included 5 features shared between CD-associated module 3 and UC-associated module 2 (both of which were highly predictive of disease, AUC > 0.8, Figure 4F), including *Escherichia Coli*, heme b biosynthesis pathway (MetaCyc PWY0-1415), (5Z)-dodecenoate biosynthesis pathway (PWY-7858), urobilin metabolite and 3-methyladipate/pimelate metabolite. Of these, the first 3 are significantly elevated in IBD fecal samples and the latter 2 are decreased. Significant correlations were detected between these features (either positive or negative) even when controlling for study group (linear regression FDR values $< 1*10^{-4}$, for all feature pairs). We also observed that the majority of metagenomic reads mapped to the two pathways were indeed contributed by *E. coli* (as reported by taxon-stratified HUMAnN3 outputs). We postulate that this set of features reflect a heme-related mechanism led by specific heme-metabolizing *E. Coli* strains. Indeed, intestinal inflammation has been shown to result in decreased enteric iron and increased enteric heme availability, promoting an environment favorable for heme utilizing bacteria such as *E. Coli*[88–90]. Furthermore, fecal urobilin, a downstream product of heme degradation (via bilirubin), is significantly decreased in IBD patients, both in this dataset and in independent cohorts[31,80], in accordance with reduced host heme catabolism. *E. Coli* has been repeatedly reported to be overrepresented in IBD patients, based on both fecal samples[31] and mucosal biopsies[90,91], often with indications of mucosa invasion in CD specifically[92]. It has been further suggested that *E. coli* has a causative role in intestinal inflammation, as opposed to merely colonizing inflamed mucosa[93]. We also note that Urobilin, Pimelic acid, and 3-Methyladipate were all found to be highly associated with the gut microbiome in multiple independent datasets, in a meta-analysis

study we have previously conducted[68]. Though fecal Pimelic acid / 3-Methyladipate were also previously shown to be significantly decreased in IBD patients[31,80], we could not recover a specific mechanistic connection to *E. Coli* and heme utilization.

Lastly, we evaluated the overall predictive power of all modules in each dataset, by training machine learning models on the 1st PCs of all modules, and estimating their combined performance with cross-validation. We found that the multi-view modules by themselves achieved an AUC comparable to that of the previous models in CD and cirrhosis datasets, while utilizing a significantly smaller number of features in total (Supplementary Table S5, Supplementary Figure S7). In the liver cirrhosis dataset, for example, MintTea modules achieved an AUC of 0.905, based on 50 features in total that were further translated into just 3 values, representing the 1st PC of each of the 3 modules that these 50 features were grouped into. In comparison, the early integration approaches achieved a slightly better AUC of 0.929, but used 383 features for training the ML model after feature selection. Similarly, the overall AUC of the CD modules was 0.961, compared to 0.958 using early-integration, utilizing 66 features (in 5 modules) compared to over 1900 used for ML training. Finally, though AUC performance in the UC and CRC datasets was lower than that of previous early integration models, AUC estimations in all cases were significantly higher than those reached when training similar models on random modules (Supplementary Figure S7).

The landscape of multi-view modules across cohorts and disease states

Having shown that MintTea was capable of identifying multi-view modules exhibiting associations both in-between views and with a disease, we finally applied MintTea to all shotgun datasets described above (i.e. a total of 17 datasets), to map the landscape of such multi-view modules across cohorts in general and across disease states. This analysis has identified several multi-view modules per dataset (average: 5.24), with a varying number of features (ranging between 2 and 95 with an average of 14.43, Figure 5A). All modules are described in Supplementary Tables S5 (modules' descriptive statistics) and S6 (features in each module). As expected, the vast majority of modules captured strong between-view correlations that were significantly higher than expected by chance (Figure 5B). We also examined which of these multi-view modules were further predictive of the disease state (AUC > 0.7 and significantly higher than random modules), identifying 1-2 such modules in some datasets (Figure 5C). Importantly, although no highly predictive modules were identified for certain

datasets, different pipeline settings may have resulted in predictive modules for these datasets as well (see Methods).

Following the example in Figure 4E, we next mapped all significantly overlapping modules across all 17 datasets (including both disease-associated and not disease-associated modules). We identified a total of 119 significant overlaps (Fisher's exact test FDR-corrected p-value < 0.1, and at least 3 features in common), suggesting highly prevalent multi-view associations between microbiome features (Supplementary Table S7, Figure 5D). These could potentially represent cohesive and universal building blocks of the microbiome's composition and function. We also identified specific bacterial species that appeared in multiple modules of different datasets. *Bifidobacterium longum*, for example, a key commensal in the human gastrointestinal tract and a common probiotic, appeared in modules of 10 different datasets (though not disease-associated in their corresponding datasets). Perhaps not surprisingly, *B. longum* tended to appear with metabolic pathways related to dietary carbohydrate metabolism, including glucose fermentation to lactate (also termed '*Bifid* shunt'[94]) and starch biosynthesis. Indeed it is estimated that approximately 12% of the annotated open reading frames in *Bifidobaterium* genomes are assigned to carbohydrate-metabolism related functions[95]. The methylerythritol phosphate pathway (MetaCyc: NONMEVIPP-PWY), and its superpathway of geranylgeranyl diphosphate biosynthesis were also consistently included in these modules, in accordance with fact that the vast majority of reads assigned to this pathway were contributed by *Bifidobaterium* species, as determined by stratified HUMAnN3 outputs. Interestingly, other species commonly appearing with *B. longum* in MintTea modules were also from the *Bifidobacteriaceae* family, including *B. adolescentis*, *B. pseudocatenulatum*, and *Aeriscardovia aeriphila* (previously classified as belonging to the *Bifidobacterium* genus as well[96]). Taken together, these modules suggest a common axis of variation shared between multiple cohorts and characterized mainly by *Bifidobacterium* species and carbohydrate metabolism. Two other examples of species appearing in multiple modules, namely *Flavonifractor plautii* (which was also consistently over-abundant across different disease states) and *Klebsiella pneumoniae*, are provided in Supplementary Figure S9. Though some features consistently co-occurred with these species across modules, most features appeared in a single module only, suggesting that different datasets contain different feature dependencies even in relation to the same bacterial species.

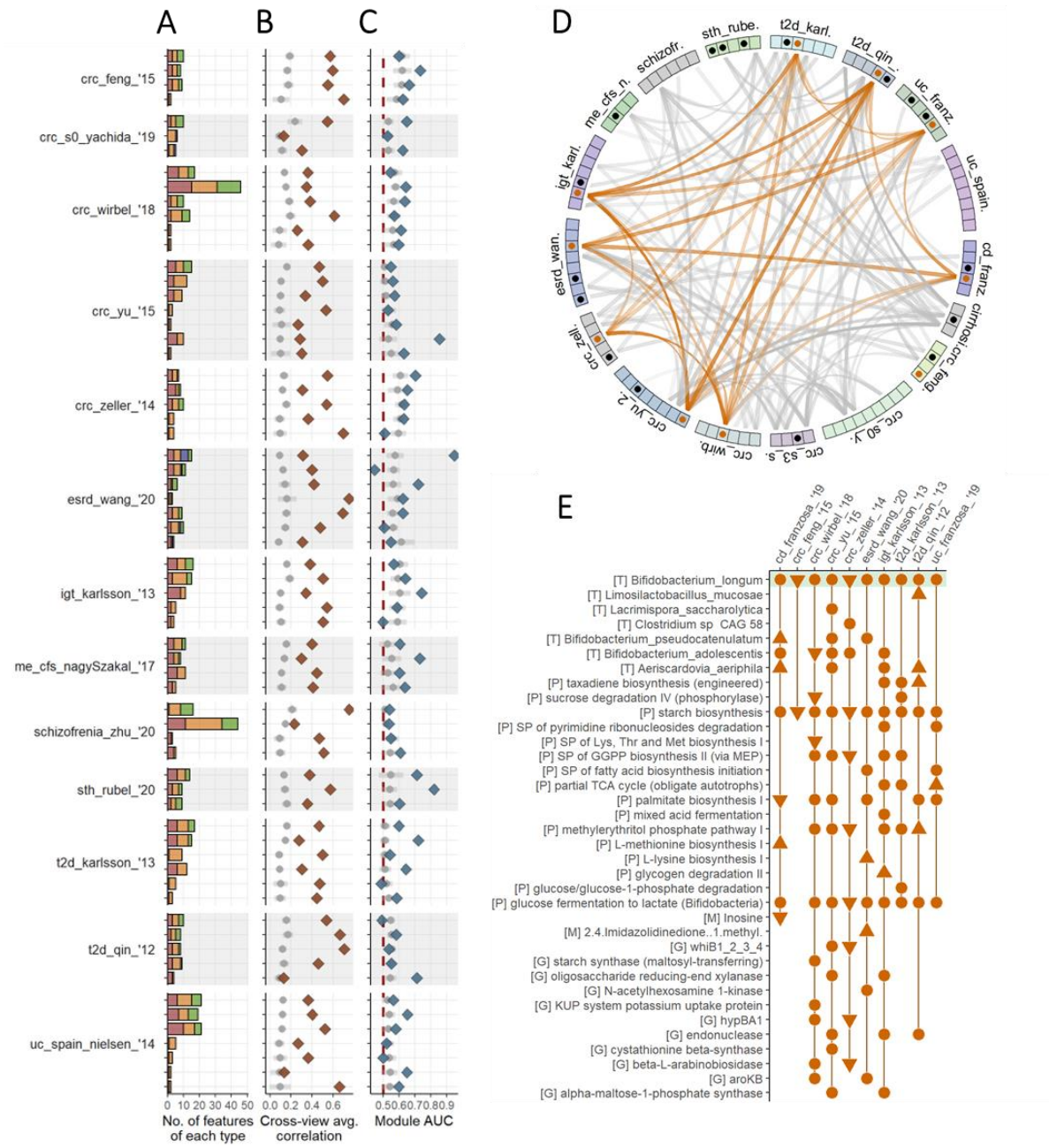

Figure 5: Multi-view modules across datasets and disease states. (A-C) Overview of multi-view modules identified for the 13 additional shotgun datasets (adding to the 4 datasets shown in Figure 4B-D). See legend of Figure 4B-D. Modules with features from a single view only were discarded and are not shown here. (D) Significant overlaps between multi-view modules of different datasets (Fisher's exact test FDR < 0.1, and at least 3 features in common). Each sector in the circus plot represents a multi-view module, grouped by the datasets to which they belong. Black dots represent disease-associated modules. Links between modules indicate significant overlaps. Modules including *Bifidobacterium longum* are further marked with orange dots and links. (E) Multi-view modules from multiple datasets that include *Bifidobacterium longum*. For each dataset listed on top, all module features are marked by points or triangles. Triangles pointing up indicate that the feature levels were

significantly increased in disease, while triangles pointing down indicate an opposite trend. Circles indicate no significant difference between study groups. None of these modules were highly predictive of disease state (AUC < 0.7), but all modules had significant cross-view correlations (r > 0.3, Methods); SP: superpathway; Lys: L-Lysine; Thr: L-Threonine; Met: L-Methionine; GGPP: geranylgeranyl diphosphate; hypBA1: non-reducing end beta-L-arabinofuranosidase; aroKB: shikimate kinase / 3-dehydroquinate synthase; whiB1_2_3_4: WhiB family transcriptional regulator, redox-sensing transcriptional regulator;

Methods

Dataset collection

In this work we collected, processed, and analyzed data from two previously published large-scale microbiome data repositories, namely curatedMetagenomicData[97] and MicrobiomeHD[6], as well as from additional studies that included both metagenomics and metabolomics data. Below are details describing the data obtained from each of these sources and their processing.

We first processed data from 3 case-control studies that profiled both stool shotgun metagenomics and metbaolomics[40,41,98]. For each of these studies we ran HUMAnN3[21] to obtain KEGG Orthology (KO) profiles and MetaCyc[99] pathway profiles. MetaPhlAn3[21] was used to obtain taxonomic profiles at the species level. Processed metabolomics data was obtained from the supplementary information of each of the original publications or online repositories. Only identified metabolites (i.e., annotated with metabolite name by the original study's authors) were included in the analysis, to allow downstream interpretation. Metadata per subject was obtained from the supplementary information. When distinct disease subtypes were present, we split the data into separate datasets per disease sub-type, taking each sub-type as the "case" patients, and the same set of control subjects as shared controls. Specifically, the IBD data from Franzosa et al.[40] were split into Chron's disease (CD) and Ulcerative Colitis (UC) and CRC data from Yachida et al.[41] were split into CRC stages. In both of these cases, the original publications indicated unique microbiome signatures for each disease sub-type/stage, and performed machine learning modelling for each sub-type/stage separately, as we performed here.

Eighteen processed shotgun metagenomics datasets were extracted from the *curatedMetagenomicData* R package[97], version 3.2.3, together with the corresponding sample metadata. Specifically, MetaPhlAn3[21] species-levels abundances and HUMAnN3[21]

unstratified gene-level (UniRef) and pathway-level (MetaCyc) abundances were obtained. We only considered datasets with data from at least 40 stool samples from different subjects in each study group (as defined by the ”study_condition” variable). Longitudinal studies were excluded. We mapped UniRef gene abundances into KO abundances using the mapping file provided by HUMAnN3. Additional details for each dataset are available in Supplementary Table 1.

We further downloaded 12 stool 16S amplicon sequencing datasets from the *MicrobiomeHD* collection[6]. Only studies with at least 40 cases and at least 40 controls were taken. Each dataset included sample metadata, 16S OTU abundances (clustered by 100% similarity using USEARCH[100]), and taxonomy assignments at the genus level (computed by RDP classifiers[101]). Additional details on each dataset are available in Supplementary Table 1. To further generate functional profiles for each 16S dataset, we ran PICRUSt2[49] with default settings and obtained KO and MetaCyc pathway profiles.

In total, data analyzed in this study included 5,702 samples, obtained from 28 studies, and partitioned into 37 case-control datasets, covering 16 disease states. Of these 5,702 samples, 759 included shotgun metagenomics and metabolome data, 2,389 shotgun metagenomics data only, and 2,554 16S data only. For most analyses, only shotgun metagenomic datasets were used.

Shared data preprocessing

For each of the datasets described above, we obtained sample metadata, taxonomy abundances (T), gene abundances (G), pathway abundances (P), and, when available, metabolite abundances (M) (Figure 1A). For all datasets: T,G, and P profiles were converted into relative abundances (by total sum scaling); Metabolite values were log-transformed; Constant and rare features (defined as those with <15% non-zero abundance values, or a mean abundance <0.005%) were discarded; Highly correlated features were clustered such that absolute correlations between any two cluster members is higher than 0.99, and using one random feature as cluster representative (Supplementary Table S8); Lastly, non-bacterial MetaCyc pathways were removed from the P profile, as annotated in the MetaCyc database within the “Expected Taxonomic Range” field.

Predicting disease state using microbiome features

We implemented a standard machine learning pipeline to assess disease predictability in each dataset and using each feature set. Training was performed with a repeated 5-fold cross

validation (10 repeats) procedure and included a feature selection step (as described below), training a random forest (RF) classifier, evaluating model performance on out-of-fold data using receiver operating characteristic (ROC) area under the curve (AUC), and evaluating feature importance (as described below). We used the 'ranger' R package[102] for RF modelling with a "min_n" of 8 and otherwise default parameters. We opted for RF due to its simplicity on the one hand and its tendency to perform well on microbiome abundance data as consistently shown in previous studies[1,4,6,48,103–105]. Hyper-parameter tuning for the RF classifiers with a nested cross-validation approach did not result in significantly improved performance for most datasets and was therefore discarded.

Feature selection

To reduce data dimensionality and risk of overfitting, we dropped non-relevant features, using 4 different 'all-relevant' feature selection approaches (including simple univariate tests, feature importance-based methods, and Boruta[43]), as well as no feature selection for comparison. We selected the "all-relevant" approach so that downstream analysis of the top-performing features will not be subjected to arbitrary removal of redundant features that may nonetheless be biologically relevant. Details about the different feature selection methods are described in Supplementary Note 1. Eventually, feature selection was carried out with the repeated 10-fold cross-validation, running independently for each fold. The best-performing method per dataset was maintained for downstream analysis (see Supplementary Table S2).

Model evaluation and comparison

Final model performances, for each dataset and each feature set, were determined by averaging AUC over the 5 repeats and 10 folds (i.e. a total of 50 trained models). The entire pipeline was also executed on a shuffled version of the data where disease labels were permuted, to compare true classification performance to that of a null (random) model. Datasets where all AUC's were below 0.7 were discarded, leaving a total of 17 datasets.

For each dataset, single- and multi-view models' performances were compared using a repeated measures ANOVA to determine whether any significant differences in AUC are present. Subsequent paired t-tests were used to determine which specific models outperformed others. We specifically used a *repeated measures* framework to link models trained and tested on the same data folds.

We also verified our AUC results against previous studies that trained microbiome-based classification models for the same datasets and using similar features, either in the original

studies from which data were obtained or in meta-analyses studies[1,4,6,41]. We only performed comparisons when the authors reported an average AUC using some form of cross-validation (as opposed to an AUC on an external validation dataset). Notably, the reference models were often based on different data transformations, different feature selection methods, and different machine learning algorithms. See Supplementary Figure S4.

Feature importance analysis

We examined the features importance in each model that achieved an AUC > 0.7, as follows: Within each fold, we calculated permutation-based feature importance scores implemented in 'ranger' package[102]. We additionally applied the method introduced by Altmann *et al.*[50] to assign a p-value for each feature in each model. Final feature importance per feature was aggregated over repeats and folds. Final p-values were FDR-corrected over all features in each feature set. We refer to features that were selected by the feature selection method in at least 50% of repeats/folds and obtained a final Altmann FDR < 0.1 as "contributors".

Identifying multi-view disease modules

To identify multi-view sets of features that are both tightly associated with each other and collectively associated with the disease, we developed a framework based on sparse generalized canonical correlation analysis (sGCCA)[106,107]. We also utilized the recently introduced DIABLO method, which extends sGCCA to a supervised scenario[70,108]. Supplementary Note 2 provides detailed explanation and formalization of sGCCA and its variants. Briefly, when applied to multiple views and a label of interest, DIABLO identifies sparse linear transformations of each view (termed 'components') so that the resulting latent variables are highly correlated between views and with the label. One key limitation of such linear models, however, is that components tend to be highly unstable, especially in the presence of thousands of features and relatively small sample sizes. Indeed, we observed that small changes in the input data (for example excluding a few samples) resulted in substantially different components. We therefore wrapped DIABLO's core sGCCA computation with a repeated sub-sampling procedure to identify "consensus components" shared across multiple runs. Specifically, we ran 100 iterations, in each one only a random sample of 90% of the samples was used for training. For each such run we recorded the list of features with non-zero coefficients from each view, for each of the first 5 components (adding additional components did not change downstream consensus components as they tend to be highly unstable). We then constructed a network of multi-view features where 2 features are connected by an edge if they

appeared together in a component in at least 80% of the iterations. Connected components of size ≥ 2 were considered "consensus components" which we term "multi-view modules".

To assess the extent to which each module is associated with disease state, we applied principal component analysis (PCA) to all module features and took the first PC as a representative of the module. We then computed the AUC for this first PC of each module, and compared it to "null" modules where the same number of features, from the same views, were randomly selected. Lastly, we trained a RF and logistic regression models on all modules to estimate the overall predictive power of modules in each dataset. Classification performance was estimated as before using repeated 10-fold cross-validation.

To assess pipelines' sensitivity to various settings, we ran the entire pipeline using various parameter combinations. The selected setting tended to perform well across most datasets as measured by overall AUC's (Supplementary Figure S8A). We additionally confirmed that using different settings tended to either expand, narrow, or split the original modules, rather than producing unrelated ones (Supplementary Figure S8B-C). For a few datasets in which the obtained multi-view modules were extremely large, we used pipeline settings that were slightly more conservative (Supplementary Table S5).

Lastly, to identify significantly overlapping modules from different datasets, we first narrowed modules to the set of features available in both datasets, and then ran Fisher's exact test to determine odds ratios and p-values. Results are listed in Supplementary Table S7.

The pipeline was developed in R using the packages "mixOmics"[108] for DIABLO running ("block.splsda" function), and "igraph"[109] for graph operations.

Discussion

In this work, we explored the application of methods from the domain of multi-view learning to microbiome data. We specifically constructed an integrated ML analysis pipeline for predicting host's disease state using multiple gut microbiome views and applied it to data from 25 case-control studies. Our "views" included taxonomic, gene-family, and metabolic pathway profiles derived from shotgun metagenomics data, as well as metabolomic profiles when available. Overall, for 17 of the 25 datasets, we were able to predict host disease based on at least one microbiome "view" with an AUC over 0.7. We found that simple multi-view models, based on naïve concatenation of features from different views ("early integration"), typically

resulted in performance comparable to the best single-view model, but importantly, further provided a mixed set of top informative features, spanning different views. Developing and applying a more sophisticated framework (MintTea) that utilizes an "intermediate integration" approach, based on sparse generalized canonical correlation analysis and a repeated subsampling procedure, we were able to identify cohesive multi-view modules of features that reflected coordinated trends in the data across views, while also being highly predictive of disease.

A perhaps surprising observation in our study was that combining multiple views for host disease prediction did not necessarily improve prediction accuracy. Nonetheless, similar results have been reported on various biological domains, using other multi-view biomedical data[110,111]. We assume that the lack of increased performance, in our analyses, stems from the large redundancy between different microbiome representations, especially in the presence of mechanisms that involve all types of features. In a way, this makes the use of intermediate integration methods even more apt, as it targets both inter-view associations and disease-related information simultaneously. Notably, different ML pipeline choices, such as an alternative feature selection approach or ML algorithm, may have resulted in slightly different results.

We also observed that comparing the single-view models to one another, different views performed best for different datasets, though the taxonomic view tended to outperform the functional ones. We emphasize that this result does not necessarily imply that the former is truly more clinically important than the latter. Functional profiles at the gene-level are likely too-high-resolution (e.g., include thousands of gene families), making it challenging for many ML models (even when applying feature selection), or too-low resolution (e.g., aggregated at the pathway level), masking more subtle disease-related signals[112]. Also, different databases used for gene-family annotation vary in their scope, curation level, inclusion of uncharacterized genes, etc., ultimately introducing additional noise to the data.

Examining the top contributing features in each multi-view model, we found that they typically spanned all input views. Though this common practice of post-hoc feature importance analysis suffers from various biases and inaccuracies (the most prominent one being that *model feature importance* is not equivalent to *biological feature importance*, especially given imperfect models) it nonetheless demonstrated that the naïve multi-view models trained on concatenated data can successfully utilize mixed sets of features, and can potentially hint at multi-view disease signatures.

Intermediate integration methods, an alternative (or complementary) approach to early integration, have been gaining increased attention due to their ability to uncover intricate multivariable interactions, as desperately required when dealing with complex biological systems[113]. Our motivation to use a method based on sCCA was based on the assumption that gut microbiome-disease associations are explained by one or more "modules" of features from all views that take part in a specific mechanism effected by or promoting host disease. Furthermore, sCCA-based frameworks have been previously applied to various microbiome multi-view data with insightful results[38,114]. To overcome challenges related to the high-dimensionality of the data and consequent instability of results, however, we wrapped sCCA with multiple repeats on data subsamples and consensus analysis, to identify *robust* multi-view modules based on multiple sCCA components. Using this approach, we identified several modules per dataset spanning multiple views, some of which were also highly predictive of the disease. Taking the liver cirrhosis dataset as one example, we identified a cirrhosis-associated module of features including bacterial species typically characteristic of the human oral microbiota, as well as oxidative-stress related functions. Notably, the first principal component of the module, *on its own*, achieved an AUC of 0.916, showing our framework's ability to simultaneously expose both inter-view and disease-related associations and further hint at biologically meaningful mechanisms.

Applying MintTea to a diverse set of 17 datasets, we identified several significantly overlapping modules, as well as specific microbiome features that appeared in modules of multiple datasets. *B. longum*, for example, appeared in modules of 10 different datasets, consistently co-occurring with dietary carbohydrate metabolism and other pathways. These modules, alongside additional groups of overlapping modules, suggest common axes of variation in the gut's microbiome composition and function, shared across multiple cohorts and representing cohesive, multi-layer microbiome modules. In a way, these modules can thus be thoughts of as the microbiome's fundamental building blocks or "modes" of operation.

Clearly, any attempt to model disease based on metagenomic and other features can only be as accurate and reliable as the underlying features. Though there are several well-established tools for taxonomic and functional profiling from metagenomics data, they may each induce biases and inaccuracies[115]. Moreover, additional, possibly invaluable features can be recovered from metagenomics and metabolomics data, for example by using assembly-based methods to add unknown-genome abundances to taxonomic profiles[115], or adding unannotated metabolites from untargeted metabolomic profiles. Furthermore, as the different taxonomic and functional

views were all derived from the same raw sequencing files, we recognize that many associations in between these views are likely expected and that taxon-gene and taxon-pathway associations within our multi-view modules may be amplified due to the specificity of certain functions to certain clades. Another limitation is the insufficient control for major confounders known to exert microbiome shifts (e.g., stool consistency and medication usage)[116]. This limitation could be potentially mitigated by careful sample stratification, inclusion of these factors as covariates in models, or even adding such data as an additional "view". Diet, for example, is likely a substantial mediator of many of the identified microbiome-metabolome modules, and it could therefore be especially interesting to include "diet"-view profiles in the pipeline to elucidate these diet-driven mechanisms. Lastly, we note that a major limitation of intermediate integration methods in general is the difficulty to systematically validate the identified modules, largely since systematic mapping of such multi-omic mechanisms is lacking. In this work for example, we suggest methods to evaluate the between-views associations in each module, as well as the overall module's association with the disease, but a validation of the biological coherence of the modules is nonetheless lacking. Future studies should also examine the generalizability of the modules to multiple cohorts of the same disease.

Overall, this work serves as a proof of concept for integrating multiple microbiome data types in a single machine learning pipeline for disease classification, and going beyond naïve, early-integration approaches by applying intermediate integration methods. We advocate for using such multi-view models to capture complementary biological insights into the complex mechanisms underlying microbiome-disease associations.

References

1. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 1–27 (2021).

2. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-Associated Microbial Signatures: Examining Their Predictive Value. *Cell Host & Microbe* **10**, 292–296 (2011).

3. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).

4. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Comput. Biol.* **12**, e1004977 (2016).

5. Le Goallec, A. *et al.* A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLOS*

Comput. Biol. **16**, e1007895 (2020).

6. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).

7. Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* **4**, (2019).

8. Beck, D. & Foster, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. *PLoS One* **9**, e87830 (2014).

9. Gou, W. *et al.* Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. *Diabetes Care* **44**, 358–366 (2021).

10. Bar, N. *et al.* A reference map of potential determinants for the human serum metabolome. *Nature* **588**, (2020).

11. Montassier, E. *et al.* Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med.* **8**, 1–11 (2016).

12. Braun, T. *et al.* Individualized Dynamics in the Gut Microbiota Precede Crohn's Disease Flares. *Am. J. Gastroenterol.* **114**, 1142–1151 (2019).

13. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 1–10 (2016).

14. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

15. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).

16. Zmora, N., Zeevi, D., Korem, T., Segal, E. & Elinav, E. Taking it Personally: Personalized Utilization of the Human Microbiome in Health and Disease. *Cell Host & Microbe* **19**, 12–20 (2016).

17. Pechal, J. L. *et al.* The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *Int J Leg. Med* **128**, 193–205 (2014).

18. LaPierre, N., Ju, C. J. T., Zhou, G. & Wang, W. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* **166**, 74–82 (2019).

19. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, (2019).

20. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun. 2019 101* **10**, 1–11 (2019).

21. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife* **10**, (2021).

22. Eng, A., Verster, A. J. & Borenstein, E. MetaLAFFA: a flexible, end-to-end, distributed computing-

compatible metagenomic functional annotation pipeline. *BMC Bioinformatics* **21**, 1–9 (2020).

23. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).

24. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).

25. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).

26. Ning, J. & Beiko, R. G. Phylogenetic approaches to microbial community classification. *Microbiome* **3**, 47 (2015).

27. Larsen, P. E. & Dai1, Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *Gigascience* **4**, 1–16 (2015).

28. Tang, Z.-Z. *et al.* Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals Microbiome-Dependent Relationships Between Diet and Metabolites. *Front. Genet.* **10**, 454 (2019).

29. Blanco-Miguez, A., Fdez-Riverola, F., Sanchez, B. & Lourenço, A. Resources and tools for the high-throughput, multi-omic study of intestinal microbiota. *Brief. Bioinform.* **20**, 1032–1056 (2017).

30. Mars, R. A. T. *et al.* Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell* **182**, 1460-1473.e17 (2020).

31. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

32. Feldner-Busztin, D. *et al.* Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* **39**, (2023).

33. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **49**, 107739 (2021).

34. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience* **25**, 103798 (2022).

35. Sammut, S. J. *et al.* Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2021).

36. Makarious, M. B. *et al.* Multi-modality machine learning predicting Parkinson's disease. *npj Park. Dis. 2022 81* **8**, 1–13 (2022).

37. Overmyer, K. A. *et al.* Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst.* **12**, 23-40.e7 (2021).

38. Priya, S. *et al.* Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat. Microbiol.* **7**, 780–795 (2022).

39. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics

data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).

40. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).

41. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).

42. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).

43. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13 (2010).

44. Kalousis, A., Prados, J. & Hilario, M. Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowl. Inf. Syst.* **12**, 95–116 (2007).

45. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).

46. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661–672 (2014).

47. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573 (2017).

48. Oh, T. G. *et al.* A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metab.* **32**, 878-888.e6 (2020).

49. Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).

50. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).

51. Lavelle, A. & Sokol, H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 223–237 (2020).

52. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**, 29–41 (2017).

53. Takahashi, K. *et al.* Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. *Digestion* **93**, 59–65 (2016).

54. Machiels, K. *et al.* A decrease of the butyrate-producing species roseburia hominis and faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283 (2014).

55. Serrano-Gómez, G. *et al.* Dysbiosis and relapse-related microbiome in inflammatory bowel disease: A shotgun metagenomic approach. *Comput. Struct. Biotechnol. J.* **19**, 6481–6489 (2021).

56. Kowalska-Duplaga, K. *et al.* Differences in the intestinal microbiome of healthy children and patients with newly diagnosed Crohn's disease. *Sci. Rep.* **9**, 31–34 (2019).

57. Galipeau, H. J. *et al.* Novel Fecal Biomarkers That Precede Clinical Diagnosis of Ulcerative Colitis. *Gastroenterology* **160**, 1532–1545 (2021).

58. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 1–7 (2017).

59. Kostic, A. D. *et al.* Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).

60. Osman, M. A. *et al.* Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer. *Sci. Reports 2021 111* **11**, 1–12 (2021).

61. Ternes, D. *et al.* The gut microbial metabolite formate exacerbates colorectal cancer progression. *Nat. Metab.* **4**, 458–475 (2022).

62. Long, X. *et al.* Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.* **4**, 2319–2330 (2019).

63. Yang, M., Soga, T., Pollard, P. J. & Adam, J. The emerging role of fumarate as an oncometabolite. *Front. Oncol.* **2 JUL**, 85 (2012).

64. Brown, D. G. *et al.* Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. *Cancer Metab.* **4**, (2016).

65. Lin, Y. *et al.* NMR-based fecal metabolomics fingerprinting as predictors of earlier diagnosis in patients with colorectal cancer. *Oncotarget* **7**, 29454 (2016).

66. Wu, G. Important roles of dietary taurine, creatine, carnosine, anserine and 4-hydroxyproline in human nutrition and health. *Amino Acids* **52**, 329–360 (2020).

67. Goedert, J. J. *et al.* Fecal metabolomics: assay performance and association with colorectal cancer. *Carcinogenesis* **35**, 2089–2096 (2014).

68. Muller, E., Algavi, Y. M. & Borenstein, E. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. *Microbiome* **9**, 1–18 (2021).

69. Zhang, H. L. *et al.* Targeting regulation of tryptophan metabolism for colorectal cancer therapy: a systematic review. *RSC Adv.* **9**, 3072–3080 (2019).

70. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).

71. Li, S. *et al.* The role of oxidative stress and antioxidants in liver diseases. *Int. J. Mol. Sci.* **16**, 26087–26124 (2015).

72. Borrelli, A. *et al.* Role of gut microbiota and oxidative stress in the progression of non-alcoholic fatty liver disease to hepatocarcinoma: Current and innovative therapeutic approaches. *Redox Biol.* **15**, 467–479 (2018).

73. Cichoz-Lach, H. & Michalak, A. Oxidative stress as a crucial factor in liver diseases. *World J.*

Gastroenterol. **20**, 8082–8091 (2014).

74. Jones, R. M. & Neish, A. S. Redox signaling mediated by the gut microbiota. *Free Radic. Biol. Med.* **105**, 41–47 (2017).

75. Chen, H. & Jiang, W. Application of high-throughput sequencing in understanding human oral microbiome related with health and disease. *Front. Microbiol.* **5**, 508 (2014).

76. Wei, X. *et al.* Abnormal gut microbiota metabolism specific for liver cirrhosis. *Front. Microbiol.* **9**, 1–4 (2018).

77. Davis, B. C. & Bajaj, J. S. The Human Gut Microbiome in Liver Diseases. *Semin. Liver Dis.* **37**, 128–140 (2017).

78. Zhong, X. *et al.* Streptococcus, the Predominant Bacterium to Predict the Severity of Liver Injury in Alcoholic Liver Disease. *Front. Cell. Infect. Microbiol.* **11**, 1–10 (2021).

79. Rigsbee, L. *et al.* Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome. *Am. J. Gastroenterol.* **107**, 1740–1751 (2012).

80. Santoru, M. L. *et al.* Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. *Sci. Rep.* **7**, 1–14 (2017).

81. Qi, Y. feng *et al.* Intestinal Microbiota Is Altered in Patients with Gastric Cancer from Shanxi Province, China. *Dig. Dis. Sci.* **64**, 1193–1203 (2019).

82. Redanz, S. *et al.* Pyruvate secretion by oral streptococci modulates hydrogen peroxide dependent antagonism. *ISME J.* **14**, 1074–1088 (2020).

83. He, L. Y., Le, Y. J., Guo, Z., Li, S. & Yang, X. Y. The Role and Regulatory Network of the CiaRH Two-Component System in Streptococcal Species. *Front. Microbiol.* **12**, 1–11 (2021).

84. Rosier, B. T. *et al.* The Importance of Nitrate Reduction for Oral Health. *J. Dent. Res.* **101**, 887–897 (2022).

85. Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11**, 1–12 (2020).

86. Chalmers, N. I., Palmer, R. J., Cisar, J. O. & Kolenbrander, P. E. Characterization of a Streptococcus sp.-Veillonella sp. community micromanipulated from dental plaque. *J. Bacteriol.* **190**, 8145–8154 (2008).

87. Zhou, P., Manoil, D., Belibasakis, G. N. & Kotsakis, G. A. Veillonellae: Beyond Bridging Species in Oral Biofilm Ecology. *Front. Oral Heal.* **2**, 1–11 (2021).

88. Celis, A. I., Relman, D. A. & Huang, K. C. The impact of iron and heme availability on the healthy human gut microbiome in vivo and in vitro. *Cell Chem. Biol.* **30**, 110-126.e3 (2023).

89. Maharshak, N. *et al.* Escherichia coli heme oxygenase modulates host innate immune responses. *Microbiol. Immunol.* **59**, 452–465 (2015).

90. Pilarczyk-Zurek, M. *et al.* Possible role of Escherichia coli in propagation and perpetuation of chronic inflammation in ulcerative colitis. *BMC Gastroenterol.* **13**, (2013).

91. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).

92. Martin, H. M. *et al.* Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology* **127**, 80–93 (2004).

93. Rhodes, J. M. The role of Escherichia coli in inflammatory bowel disease. *Gut* **56**, 610–612 (2007).

94. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).

95. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol.* **7**, (2016).

96. Simpson, P. J., Ross, R. P., Fitzgerald, G. F. & Stanton, C. Bifidobacterium psychraerophilum sp. nov. and Aeriscardovia aeriphila gen. nov., sp. nov., isolated from a porcine caecum. *Int. J. Syst. Evol. Microbiol.* **54**, 401–406 (2004).

97. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).

98. Wang, X. *et al.* Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. *Gut* **69**, 2131–2142 (2020).

99. Caspi, R. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **34**, 511–516 (2006).

100. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

101. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

102. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, (2015).

103. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

104. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

105. DiMucci, D., Kon, M. & Segrè, D. Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks. *mSystems* **3**, (2018).

106. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* **8**, (2009).

107. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**, 569–583 (2014).

108. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).

109. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* 1–9 (2006).

110. Ray, B. *et al.* Information content and analysis methods for Multi-Modal High-Throughput Biomedical Data. *Sci. Reports 2014 41* **4**, 1–10 (2014).

111. Culley, C., Vijayakumar, S., Zampieri, G. & Angione, C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 18869–18879 (2020).

112. Manor, O. & Borenstein, E. Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. *Microbiome* **5**, 1–11 (2017).

113. Wang, Y. *et al.* Multi-modal intermediate integrative methods in neuropsychiatric disorders: A review. *Comput. Struct. Biotechnol. J.* **20**, 6149–6162 (2022).

114. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host & Microbe* **17**, 260–273 (2015).

115. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

116. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

117. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).

118. Iguyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182 (2003).

119. Sun, L., Ji, S., Yu, S. & Ye, J. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. in *IJCAI International Joint Conference on Artificial Intelligence* 1230–1235 (2009).

120. Tenenhaus, A. & Tenenhaus, M. Regularized Generalized Canonical Correlation Analysis. *Psychom. 2011 762* **76**, 257–284 (2011).

Data and code availability

Data used in this study was retrieved from either the curatedMetagenomicData package[97] (available at: https://waldronlab.io/curatedMetagenomicData/), the MicrobiomeHD collection[6] (available at: https://zenodo.org/record/569601), or supplementary tables and deposited data from 3 specific studies[40,41,98] (exact source of each table and accession numbers are listed in Supplementary Table S1). The R code used for this analysis, including the MintTea pipeline, is available on GitHub (https://github.com/borenstein-lab/multi_view_integration_analysis).

Acknowledgements

We would like to thank the Borenstein lab members for helpful feedback and insightful discussions during the process. We thank all authors of the multiple microbiome studies whose data was included in our analysis, for generating invaluable data and making them publicly available. We are also grateful to Professor Uri Gophna for his important inputs, support, and assistance. Lastly, we also thank the Tel Aviv University Computer Science IT team for their technical support during computational work. This work was supported in part by National Institutes of Health grant U19AG057377, and Israel Science Foundation grant 2435/19 to EB. EB is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. EM and IS are supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary notes

Supplementary note 1: "All-relevant" feature selection approaches

We compared 4 different feature selection approaches, all of which are designed to retain all informative features (weakly informative as well). Notably, in cases where classification accuracy is the only objective, other approaches that reduce redundancy should probably be tested and even favored (e.g. mRMR[117], or Recursive Feature Elimination), as machine learning models may be prone to overfit in face of large redundancy in the data[118]. The first approach was based on a univariate statistical test of differential abundance between cases and controls. Such an approach is commonly referred to as a "filter" feature selection method. We specifically chose the Mann Whitney U-test, running it on either relative abundances or CLR-transformed values and taking all features that obtained a p-value below a specified threshold.

The second approach was a modified Boruta feature selection[43]. Roughly, the Boruta algorithm adds "shadow features" to the original dataset, which are permutations of the original features. This way the distributions of the true data are maintained but the associations with disease, if existed, are removed. The machine learning algorithm of choice is then trained (here: random forests, RFs), and all features are scored for their importance by the default RF method. The maximal importance score among all shadow features is considered a threshold distinguishing truly informative features (features with scores above the threshold) and non-informative ones (scores below the threshold). This permutation and evaluation process is repeated multiple times and a statistical criterion is used for final definition of informative and uninformative features. Due to the conservative nature of the Boruta method, we relaxed the threshold used for defining informative features, as implemented in the python Boruta implementation (https://pypi.org/project/Boruta/). The third approach was an ensemble of the previous two approaches, taking the intersection of selected features.

Lastly, the fourth approach was based on the Altmann method[50], computing p-values per feature reflecting the significance of the feature's contribution to the model and taking features with a p-value below a specified threshold. We additionally ran our pipeline with no feature selection at all, to serve as a baseline of performance.

The entire evaluation of feature selection methods was incorporated in a repeated 5-fold cross-validation framework. Overall, different feature selection methods performed best for different

datasets and different feature sets (see Supplementary Table S2), though differences were typically non-significant.

Supplementary note 2: Explanation and formulation of the intermediate integration pipeline

The intermediate integration approach implemented in this study is based on sparse canonical correlation analysis (sCCA) and specifically a recently introduced framework named DIABLO[70]. The modifications and expansions of sCCA which we implemented in the MintTea framework are described in the Methods section. Explanations and formulations of sCCA and its relevant expansions are provided below.

Sparse canonical correlation analysis (sCCA)

Briefly, sCCA is a method used to identify sparse linear combinations of two sets of features describing the same set of samples, so that the resulting latent variables are maximally correlated with each other. Formally, we denote 2 standardized feature tables ('views') as $X^{(1)} \in \mathbb{R}^{n \times P_1}$ and $X^{(2)} \in \mathbb{R}^{n \times P_2}$, P_1 and P_2 being the number of features in each view respectively, and n denoting the number of samples. Sparsity is enforced using a lasso (ℓ_1) penalty function on the loading vectors, essentially making most of the loading values equal to zero, with a hyper-parameter λ determining the degree of sparsity. It is particularly helpful in cases where the number of features substantially exceeds the number of samples. sCCA then solves the following optimization problem iteratively per component $h \in 1, \dots, H$, H being the total number of requested components (a user-defined hyper-parameter):

$$\underset{a_h^{(1)}, a_h^{(2)}}{\operatorname{argmax}} \ \operatorname{cov}\left(X_h^{(1)} \cdot a_h^{(1)}, X_h^{(2)} \cdot a_h^{(2)}\right)$$
$$s.t. \left\|a_h^{(1)}\right\|_2 = 1, \left\|a_h^{(2)}\right\|_2 = 1, \left\|a_h^{(1)}\right\|_1 \leq \lambda^{(1)}, \left\|a_h^{(2)}\right\|_1 \leq \lambda^{(2)}$$

In the above formula, $a_h^{(q)}$ is the loading vector for view q and component h, and $X_h^{(q)}$ is the residual matrix of the regression between $X_{h-1}^{(q)}$ *and* $a_{h-1}^{(q)}$. The use of residual matrices guarantees orthogonality of components within each block. We note that the use of *covariance* in the objective function, instead of *correlation* as in the original CCA, is clarified in Tenenhaus et al[107] and is closely related to the PLS method[119].

Sparse generalized canonical correlation analysis (sGCCA) and DIABLO

To expand sCCA to $Q \geq 2$ views, i.e. generalizing sCCA, the objective function shown above is set to maximize the *sum* of correlations between all pairs of latent variables. Tenenhaus and Tenenhaus (2011)[120] further introduced the notion of a design matrix, C, to weigh each pair of views according to the expected level of association between them, based on prior knowledge. Formally, a design matrix $C = \{c_{ij}\}$ is defined by the user, where $c_{ij} \in [0,1]\ \ \forall i,j \in 1,\dots,Q$, so that $c_{ij} = 1$ if view i and view j are expected to be closely related, and 0 if they are expected to be unrelated. Intermediate design values, between 0 and 1, are also supported. The resulting optimization problem for sGCCA therefore becomes:

$$\underset{a_h^{(1)},\dots,a_h^{(Q)}}{\operatorname{argmax}} \sum_{\substack{i,j=1 \\ i \neq j}}^{Q} c_{ij} \cdot \left(\operatorname{cov}\left(X_h^{(i)} \cdot a_h^{(i)},\ X_h^{(j)} \cdot a_h^{(j)}\right)\right)$$
$$s.t. \left\| a_h^{(q)} \right\|_2 = 1 \ \ and \ \ \left\| a_h^{(q)} \right\|_1 \leq \lambda^{(q)} \ \ \forall q \in 1,\dots,Q$$

The optimization problem is solved by an efficient iterative algorithm described in detail in Tenenhaus et al[107]. DIABLO extends sGCCA to classification problems by adding an additional "dummy" view, namely X^{Q+1}, with one or more indicator variables representing the label of each sample (here: host disease state). By doing so, the identified "components" reflect both inter-view relationships as well as associations with the disease. Furthermore, DIABLO substitutes the lasso's λ parameter with a "*keepX*" parameter indicating the exact number of variables to select from each view, for ease of use[70].

Overall, given Q view tables and a phenotype of interest, DIABLO returns one or more "components" which are sparse linear combinations of the original views, capturing a multi-view set of features that are both closely related with one another and jointly associated with the phenotype. Importantly, there's a tradeoff between the two (a "correlation-discrimantion tradeoff"), meaning that, at the extreme, some components may be strongly associated with the disease but have low inter-view correlations, and vice versa. Though DIABLO offers additional features, mainly label prediction based on distances in the latent space, we utilized only its core sGCCA computation, as described above.

Supplementary figures

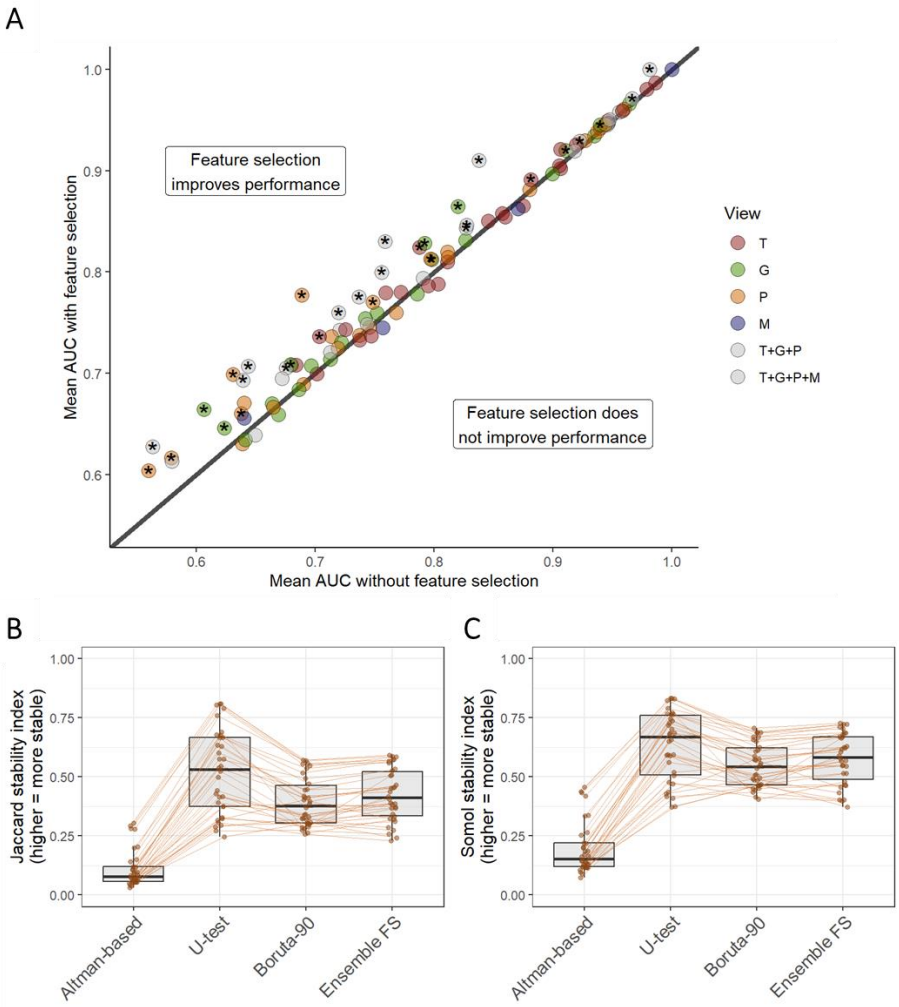

Figure S1: Feature selection stability and performance improvement. (A) Disease classifiers' performance (measured by a random forest AUC) with (Y axis) vs. without (X axis) applying feature selection before training. Each point represents a specific dataset (of the 26 analyzed datasets) and a specific feature set. Only models with AUC > 0.6 (either with or without feature selection) are presented.. Stars mark cases where the difference was significant (paired t-test FDR < 0.05). (B-C) Comparison of feature selection stability using different feature selection approaches. Each dot represents the Jaccard (B) or Somol (C) stability index for a specific dataset and using the concatenated feature set (i.e. all omics available for the dataset). Stability indices quantify how robust a feature selection is to small changes in the input data (higher values indicate greater stability), and are calculated here over data subsets as part of a cross-validation procedure.

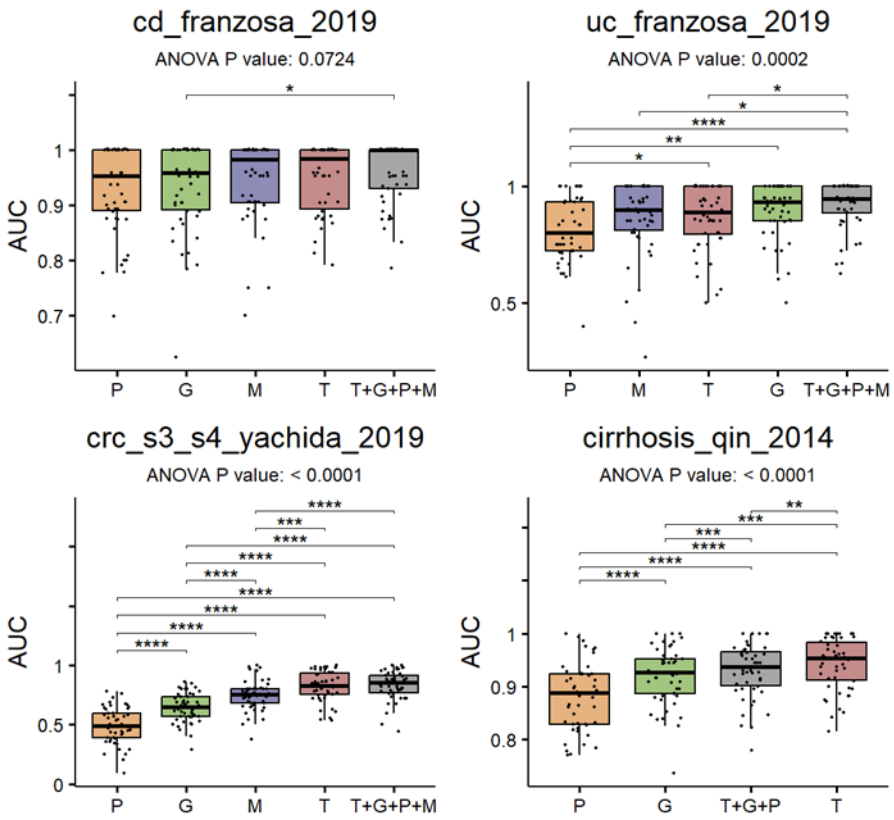

Figure S2: Comparison of single-view and multi-view classifiers' performance. For each of the 4 main datasets (listed at each panel's title) we show detailed performance distributions over repeated cross-validation iterations for both single- and multi-view RF models. Models are ordered within each panel from best average performance (right) to worst (left). Significant differences were determined by paired T-tests and a subsequent FDR correction. Significance marks:

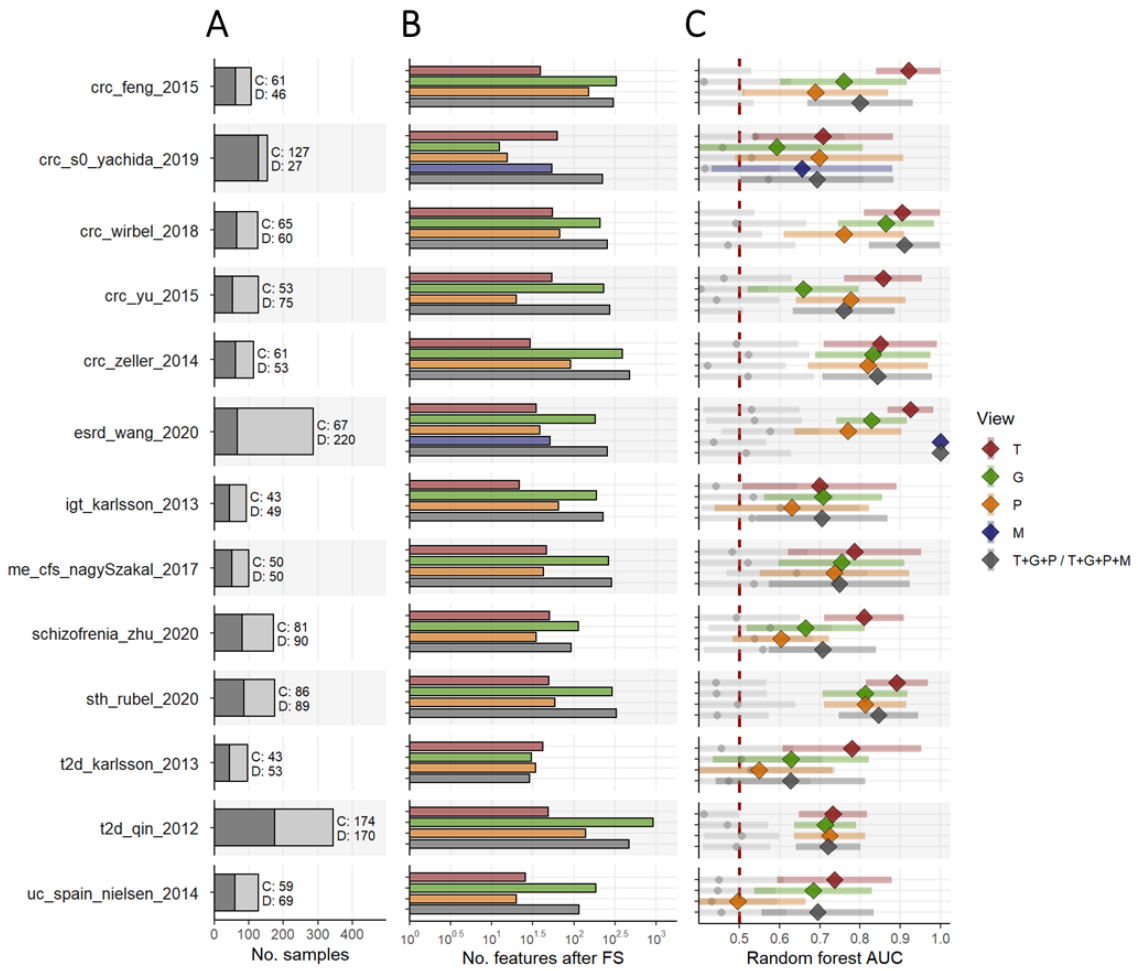

Figure S3: Single- and multi-view disease classification based on shotgun sequencing data. See Figure 2 legend. Datasets presented here were retrieved from the curatedMetagneomicData repository[97] or manually curated (See Supplementary Table S1).

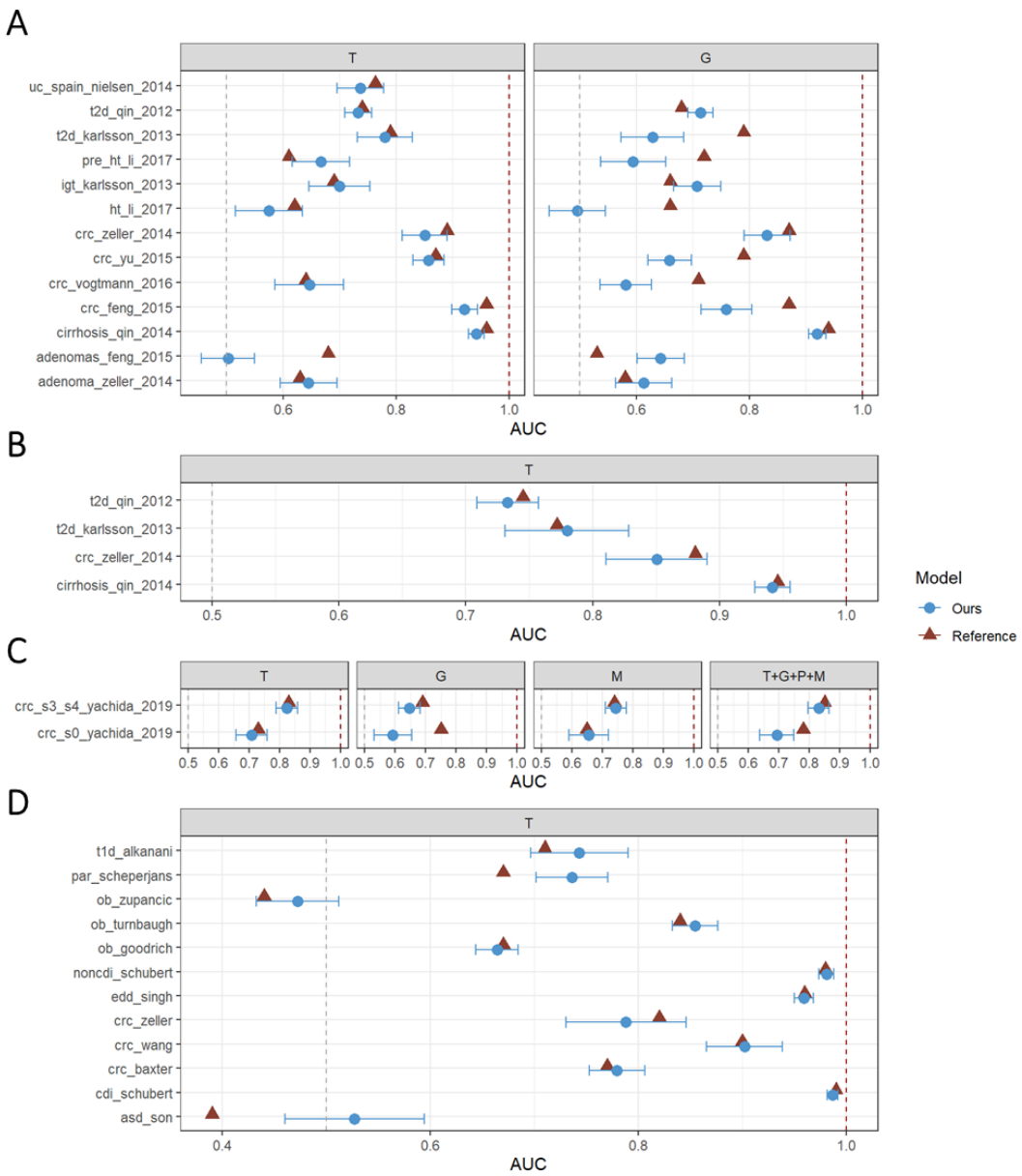

Figure S4: AUC comparisons to previous machine learning studies using the same datasets. We compared AUC results of our pipeline to those of similar pipelines reported in previous studies or meta-analyses, specifically Wirbel et al.[1] (A), Pasolli et al.[4] (B), Yachida et al.[41] (C) and Duvallet et al.[6] (D). AUC results of these studies were extracted from either publication figures or supplementary information. As some of the studies did not report confidence intervals, and others reported different types of confidence intervals or standard errors, we only included their reported average AUC's. For our AUC's we included 95% confidence intervals of the mean (using the 't.test' function). Notably, functional feature sets ("G") differed significantly between our study and Wirbel et al. and Yachida et al., as each one used a different reference database for gene annotations, substantially affecting the percent of annotated reads and the resolution of gene families.

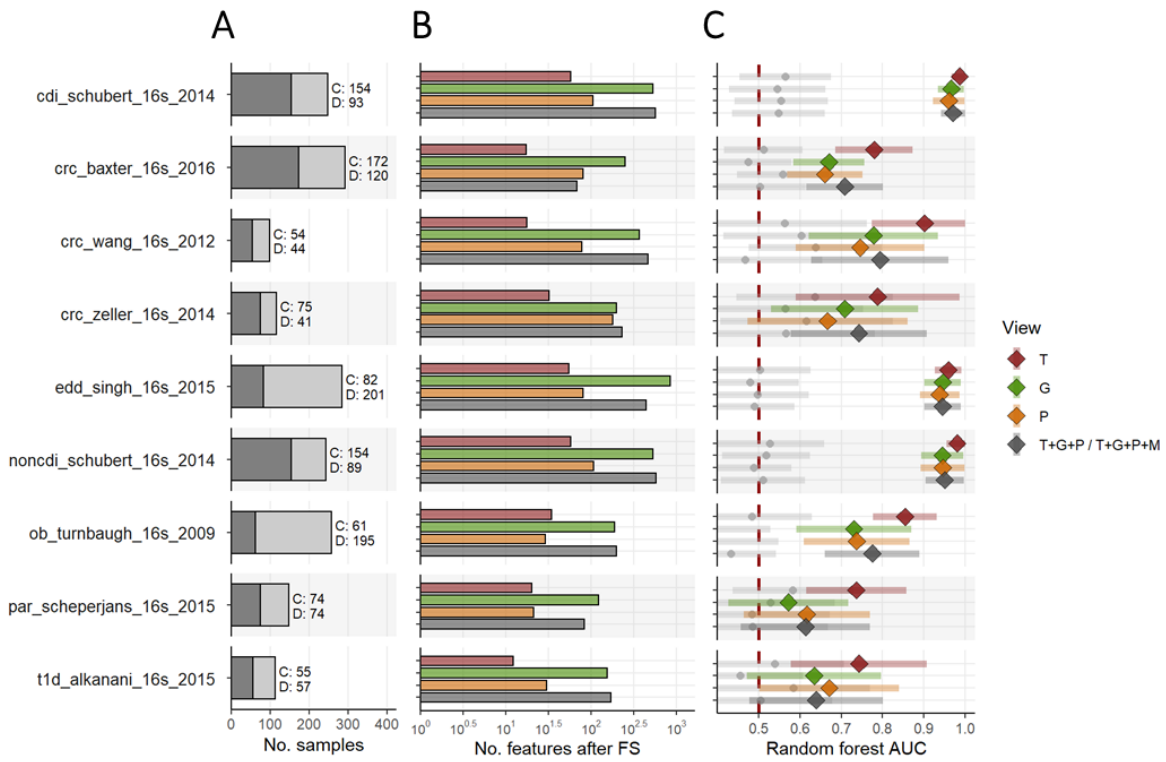

Figure S5: Single- and multi-view disease classification based on 16S amplicon sequencing data from MicrobiomeHD resource. See Figure 2 legend. Datasets presented here were retrieved from MicrobiomeHD[6].

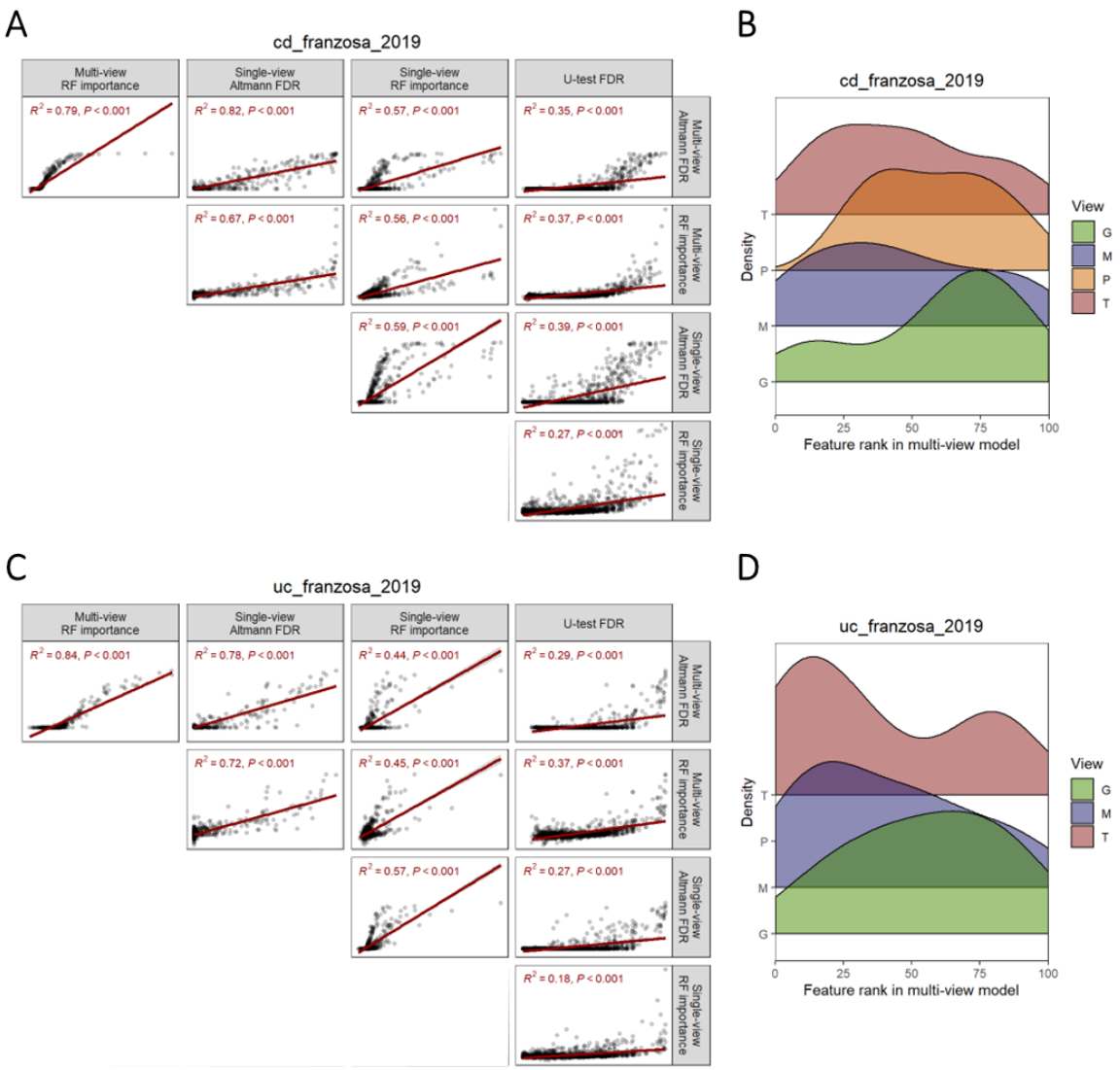

Figure S6: Multiple estimations of feature importance. Left panels present a pairwise comparison between different feature importance approaches, for the CD dataset (A) and the UC dataset (C). Each point in the scatter plot represents a feature, and overall correlation is illustrated by a fitted linear regression and its associated p-value. The following feature importance methods are included in the pairwise comparisons: "Single-view RF importance", referring to the permutation-based feature importance score calculated as part of the RF training, for each omic independently; "Single-view Altman FDR“, referring to the FDR-corrected p-value generated using the Altman method (see Methods); "Multi-view RF importance“ and "Multi-view Altman FDR“ are the equivalent metrics but using the concatenated multi-view dataset; "U-test FDR“, refers to the FDR-corrected p-value from a Mann–Whitney U test. Right panels (B, D) present the distribution of feature ranks in the CD and UC (respectively) multi-view models. Only the top 100 features are presented. RF: random forest; CD: Crohn’s Disease; UC: Ulcerative colitis.

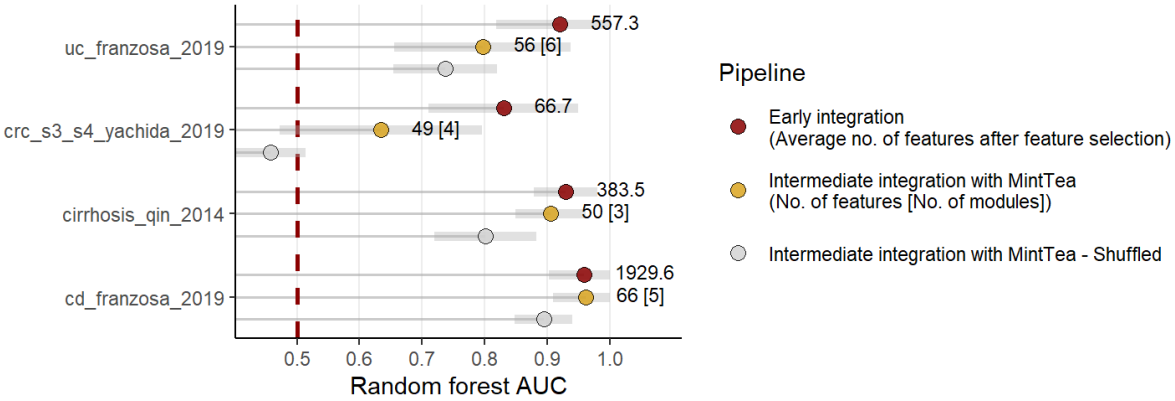

Figure S7: Overall predictive power of multi-view modules. For each of the four datasets, a random forest was trained on the first principal components of each module in that dataset. Cross-validated AUC's of these models are shown in yellow, and compared to standard random forest models on raw features (in red) and models trained on random modules (grey). Numbers next to red dots represent the average number of features used for the model (averaged over cross-validation iterations). Numbers next to yellow dots represent the total number of raw features included in modules, and in square brackets the respective number of modules in that dataset.

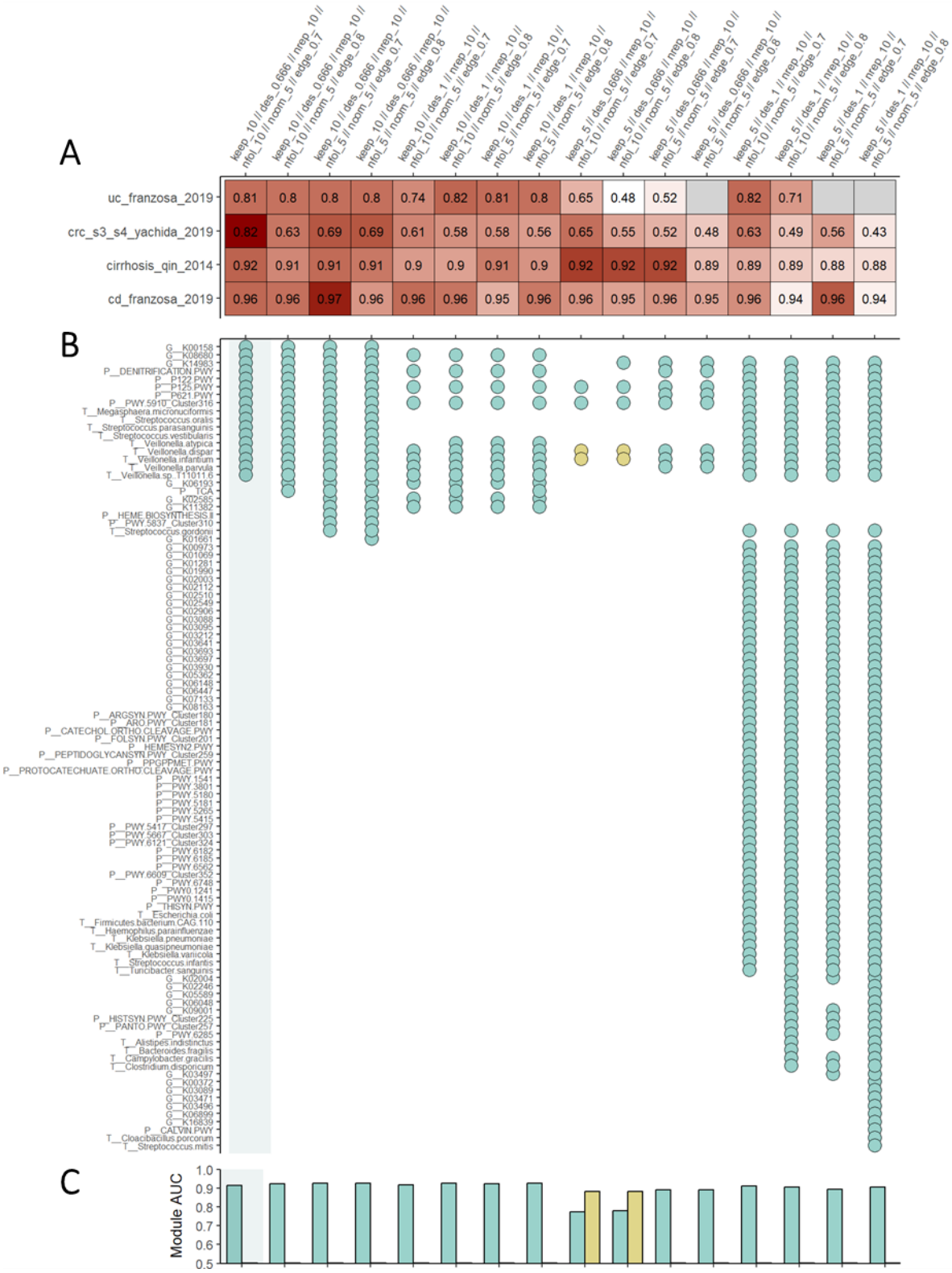

Figure S8: Sensitivity analysis for intermediate multi-view integration pipeline. The x axis for all panels are the various combinations of parameters required by our intermediate integration pipeline. Briefly, parameters include: 'keep' – the number of features to select from each view as provided to DIABLO's sGCCA computation; 'des' – the value to set as design matrix default (see Supplementary Note 2); 'nrep' – number of repetitions; 'nfol' – number of folds to which the entire data is split in order

to run DIABLO on nfol-1 folds at a time (i.e., subsampling); 'ncom' – number of components as provided to DIABLO's sGCCA; 'edge' – the threshold defining how often should a pair of features co-appear in a DIABLO component in order for them to co-appear in the final modules. The leftmost pipeline setting is the one used in the main results section of this study. (A) The overall AUC of disease classifier models, achieved for each pipeline setting, for each of the 4 main datasets (see Methods). Tiles are colored by their AUC scaled over rows (datasets), illustrating that no single pipeline setting outperforms others across all datasets. (B) Variations of module 2 from the cirrhosis dataset, as outputted by each pipeline setting. The Y axis lists features from all views, and points indicate that a feature was included in a multi-view module that resulted from using the specific pipeline setting (x axis). Only modules overlapping with the original module (highlighted, leftmost) are plotted, to illustrate how different pipeline settings tend to either expand the original module, narrow it down, or split it into two or more modules (in which case, different modules' features are colored differently). (C) Single module AUC's for each module appearing in panel B.

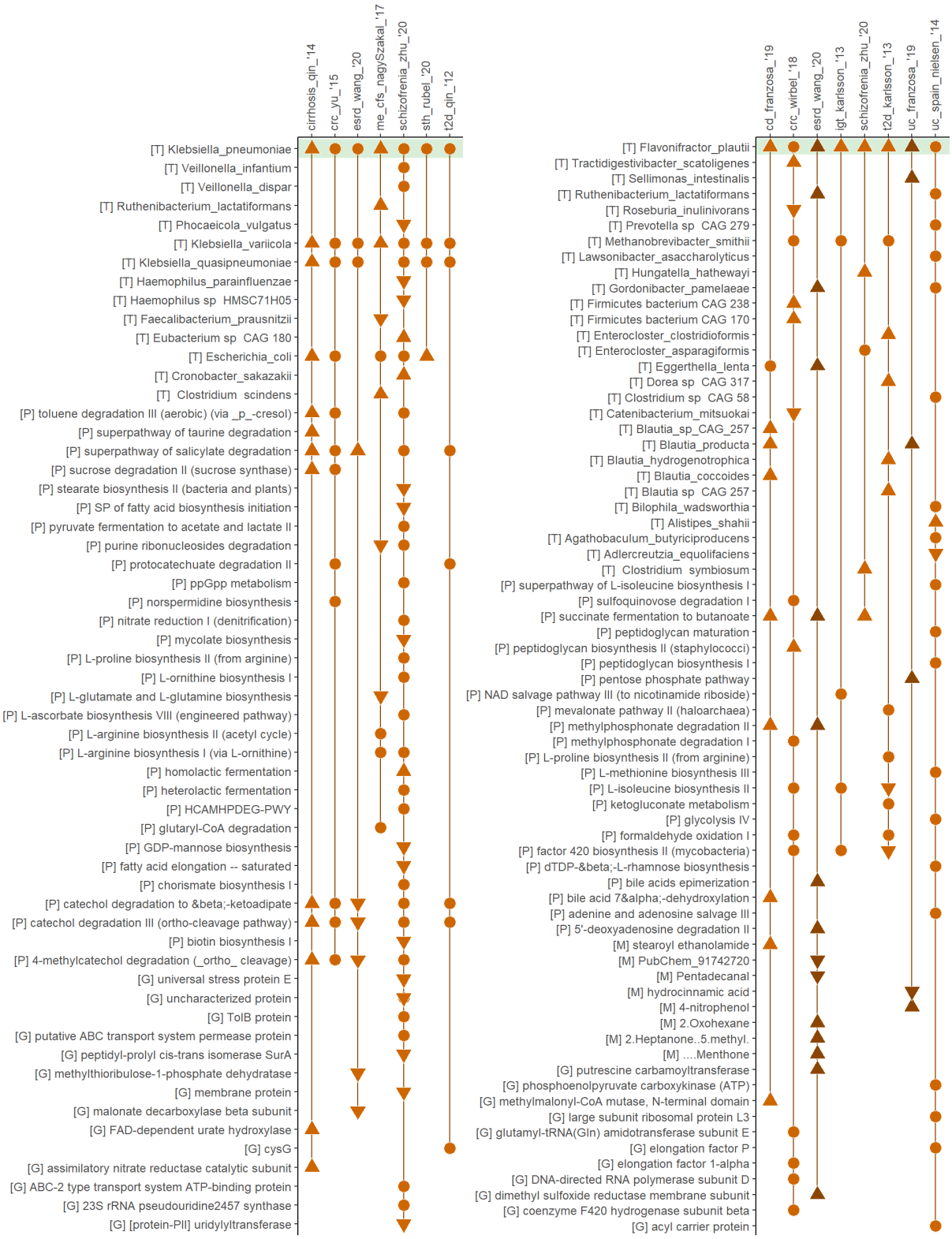

Figure S9: Multi-view modules including the taxa *K. pneumoniae* (left) and *F. plautii* (right). See Figure 5E legend. Modules colored in dark orange were highly predictive of disease state (AUC > 0.7, and higher than AUC of random modules).

Supplementary tables

Table S1: Datasets included in the analysis

Table S2: RF model performance results, for both single-view and multi-view models

Table S3: Pairwise comparisons between single- and multi-view models in each dataset

Table S4: RF top feature contributors, for both single-view and multi-view models

Table S5: Multi-view modules - Overview

Table S6: Multi-view modules - Detailed features

Table S7: Significantly overlapping modules across datasets

Table S8: Feature clusters