Multi-view integration of microbiome data for identifying disease-associated modules

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; elbo@tauex.tau.ac.il

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 aimed to investigate the impact of various integration approaches on the ability to predict disease state based on multiple microbiome views, and primarily to generate multifaceted biological hypotheses linking features across views. We first implemented an "early integration" ML pipeline to serve as a baseline for disease predictability and informative features' detection, and compared the resulting multi-view models to naïve single view models. Using a dataset collection of 25 case-control metagenomic studies, we found that multi-view models typically offer prediction accuracy that is comparable to that of the best-performing single view model, yet, further provide a mixed set of informative features from different views while accounting for between-views dependencies. Such early-integration multi-view models, however, are still limited in their ability to provide well-defined disease-associated multi-view modules that can be used to generate mechanistic hypotheses concerning the disease in question. To address this challenge, we developed a novel "intermediate integration" framework termed MintTea, based on sparse generalized canonical correlation analysis (CCA), to identify *multi-view modules* of features, highlighting shared disease-associated trends in the data expressed by the different views. We showed that this framework identified multiple modules that both are highly predictive of the disease, and exhibit strong within-module associations across features from different views. We further demonstrated that MintTea has significantly lower false discovery rates compared to other CCA-based approaches. 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 that aim to predict 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 of the microbiome with the host phenotype[1,4,6], using model explainability methods to identify a subset of important features and generate new biological hypotheses[7–9], or developing new diagnostic tools, risk score calculators[8,10–13], or microbiome-based recommendation systems[14,15]. 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 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[1,4,5,16].

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 views: Researchers can choose to focus on taxonomic profiling (and at varying taxonomic levels) using tools such as Kraken2[17], mOTUs2[18] or MetaPhlAn4[19] for taxonomy assignment and abundance estimation. They can also use functional profiles, estimating the abundances of genes or pathways in the metagenome, as generated by tools such as HUMAnN3[20] or MetaLAFFA[21]. Additional profiles that could potentially be inferred based on shotgun data include strain markers[22], single nucleotide variation profiles[18], microbial genomic structural variants[23], bacteriophage profiles[24], and others. Moreover, advances in various high-throughput technologies have recently prompted a proliferation of *multi-omic* microbiome studies, where the same set of samples are profiled via multiple molecular assays[25–27], including, for example, transcriptomics or metabolomics. When other such omics are available, these can be similarly used as additional views for modelling, offering potentially complementary perspectives of the microbiome and its interaction with the host. Indeed, focusing on the gut microbiome, the addition of fecal metabolomics, for example, can provide valuable insights into the metabolic environment in the gut, which could in turn be incorporated into the constructed 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*. It is also important to note that 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[4,5,16,28,29].

Importantly, however, when data from such multi-omic studies are analyzed, the different views (omics) are often examined independently, or in a cross-view univariate manner using, for example, correlation networks[25,30–32]. Clearly, in the context of a classification problem, 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[33–35], and successful multi-omic classification models have been introduced, for example, in breast cancer therapy prediction[36], early diagnosis of Parkinson's disease[37], and COVID-19 severity prediction[38]. In the microbiome domain, ML methods that allow multi-view integration are still somewhat scarce, but a few such approaches have been previously used to highlight association with disease, to identify disease subtypes, or to generate mechanistic hypotheses about host-microbiome interactions involving multiple omics[25,27,39,40]. One recent study, for example, used sparse canonical correlation analysis (sCCA) to integrate taxonomic profiles and host transcriptomics to identify multi-omic associations that are either shared or specific across different gastrointestinal diseases[27]. These studies testify to the potential benefit of such models, yet, their applicability to diverse microbiome datasets, and the robustness of the obtained results, remain unclear.

In this study, we set out to explore and promote the use of multi-view machine learning for microbiome research, hypothesizing 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 could potentially be identify by advanced integration approaches. 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 both the prediction accuracy and the identification of informative features obtained by models trained on single-view data and by models trained on concatenated multi-view data (a.k.a. "early integration" approach), as a baseline for ML-based detection of microbiome-disease associations. We then developed

106 a novel "intermediate integration" approach which we term *MintTea*, to specifically identify
107 modules of features from all views that jointly shift in disease, thus reflecting both
108 interactions between views and association with disease. Unlike standard multi-view ML
109 pipelines, MintTea is focused on identifying robust and meaningful multi-view modules, and
110 not only optimizing predictive performance (and see also Supplementary Note 1). We
111 showed that some of the modules identified by MintTea are in agreement with previous
112 observations concerning the role of the microbiome in disease, and provided a complete
113 catalog of identified multi-view modules across all datasets analyzed. This work thus serves
114 as a proof of concept for the benefit of integrating multiple microbiome data types in a single
115 ML pipeline for disease classification, and for going beyond naïve, early-integration,
116 approaches.

117 Results

118 Overview

119 To examine different multi-view integration approaches across diverse microbiome datasets,
120 we obtained 25 different datasets of shotgun metagenomics sequencing. All datasets are from
121 case-control studies, spanning multiple health conditions and diseases (Supplementary Table
122 S1). For each dataset, we generated 3 different feature sets based on the metagenomic data,
123 namely: a taxonomic profile ("T"), a gene family profile ("G"), and a pathway-level profile
124 ("P") (see Methods, Figure 1A). For 7 of the 25 datasets, metabolomics data was also
125 available, and we hence additionally included a metabolite profile ("M") for these datasets.
126 We specifically focused initially on 4 datasets, describing cases-control studies of ulcerative
127 colitis (UC) and Crohn's disease (CD)[41], late-stage colorectal cancer (CRC)[26], and liver
128 cirrhosis[42]. We then also describe result highlights for all of the additional datasets.

129 In addition to analyzing these multi-view data without integrating the different profiles, we
130 focused on two different integration approaches (Figure 1B), with a two-pronged objective in
131 mind: First, optimizing disease classification based on multiple different microbiome views,
132 and second, generating biological hypotheses that capture the interactions between the
133 different microbiome views and between these views and the disease state. We followed
134 common terminology in this domain and labeled the two approaches as "early integration"
135 and "intermediate integration", where the former refers to a naïve concatenation of different

136 views into a single joint table ("T+G+P" or "T+G+P+M"), and the latter refers to a joint
137 analysis of interactions between views and between each view and the phenotype, ultimately
138 identifying multi-view modules of features with a concordant disease-associated behavior.

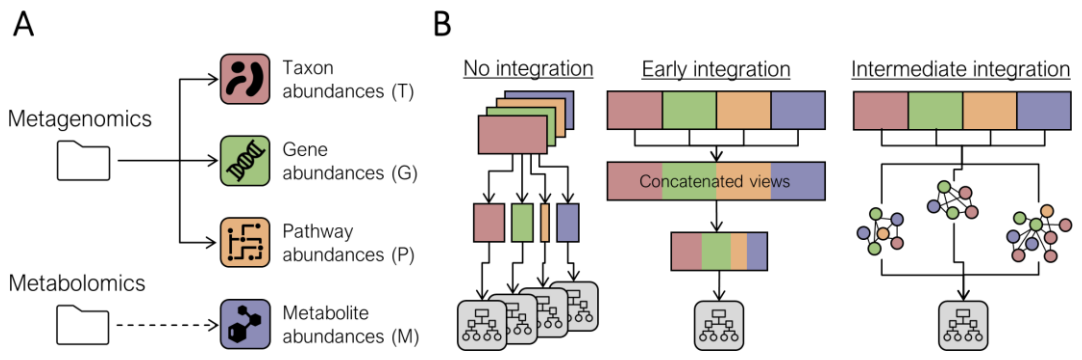

Figure 1: Multi-view microbiome data and integration approaches. (A) Data types used in this analysis. Throughout the manuscript, taxonomic, gene, pathway and metabolite abundances are labeled with the letters 'T', 'G', 'P', and 'M', respectively. (B) Different integration approaches carried out using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

139 ## Single-view vs. early integration multi-view disease models

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

155 examined the stability of each feature selection method – an important property of feature
156 selection that is often overlooked[44] – by comparing the selected features when the method is
157 applied to different subsets of the data. We found that different methods varied significantly
158 in their stability (Supplementary Figure S1B-C), and that higher stability was, perhaps not
159 surprisingly, strongly correlated with higher model AUCs (Pearson correlation > 0.7, $P < 10^{-7}$
160 across all feature selection methods). We finally ran this pipeline once for each feature set
161 independently ("single-view" models), and once using a concatenated table of all feature sets
162 (the "multi-view" model, using an "early integration" approach as described above). Receiver
163 operating characteristic (ROC) area under the curve (AUC) was averaged over cross-
164 validation iterations for each dataset and each feature set (See Methods).

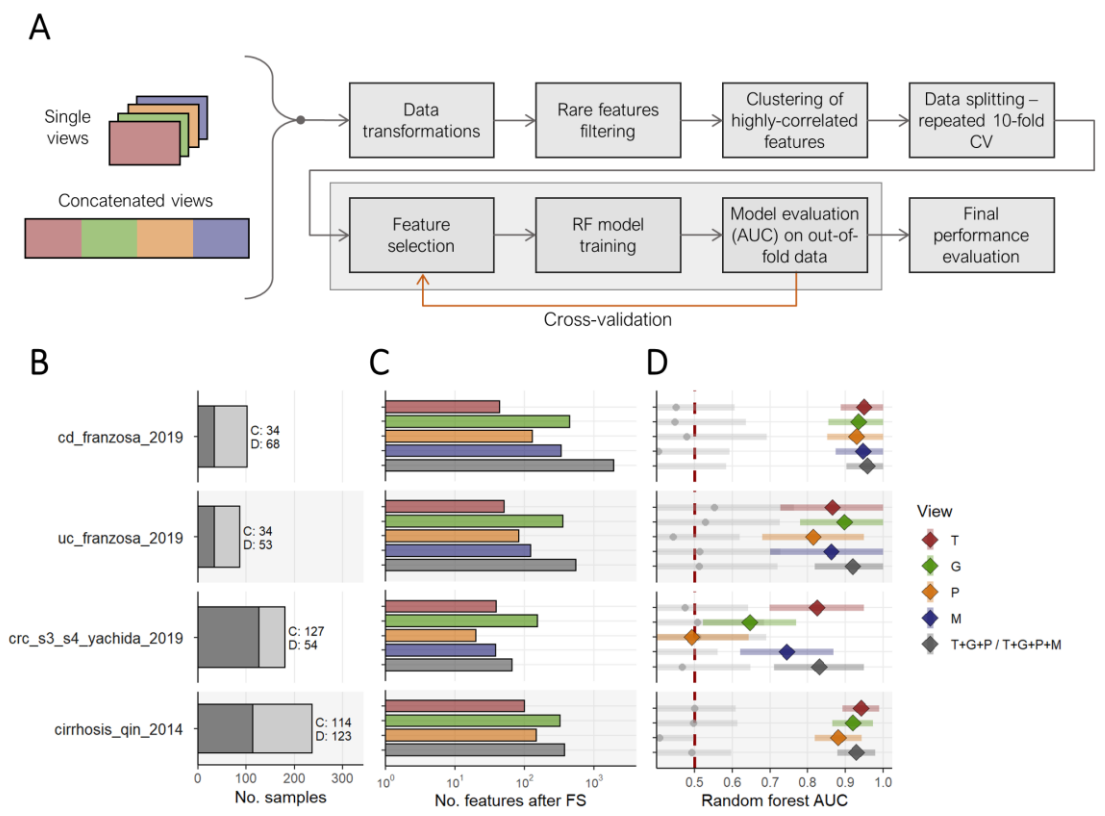

Figure 2: Multi-view machine learning (ML) modelling of host diseases using no integration vs. early integration. (A) A diagram of a ML pipeline for microbiome data. Model evaluation is carried out through a repeated cross-validation procedure. Final estimations of model performance are averaged over 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 Supplementary 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.

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

179 To benchmark our entire pipeline we also compared our results to previously published
180 studies which applied similar ML pipelines to the same datasets[1,4,6,26]. Notably, although
181 several major differences exist between the pipelines in terms of modelling approaches,
182 feature transformation, and evaluation strategy, our pipeline resulted in comparable AUC's to
183 other pipelines across most datasets (Supplementary Figure S4). We also compared our
184 model to null models, where study group labels were shuffled, confirming that all AUC
185 values obtained by the true models were significantly higher than those obtained by the null
186 models (Figure 2D, Supplementary Figure S3C).

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

Identifying disease-associated features in single-view and early-integration multi-view models

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 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]. Multiple other taxonomic (e.g. *Eubacterium siraeum*, *Dorea formicigenerans*, *Oscillibacter sp. 57_20*), functional (e.g. pyrimidine deoxyribonucleoside salvage, methylglyoxal degradation), and metabolite (e.g. chenodeoxycholate / deoxycholate, urobilin, adrenic acid, nicotinic acid) features with high importance scores were in agreement with differential analysis results on various IBD cohorts worldwide[25,53–55]. We also note that CD was 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 gut microbiome and metabolome in CD compared to UC[56].

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 was 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. A similar analysis of the CRC dataset can be found in Supplementary note 3.

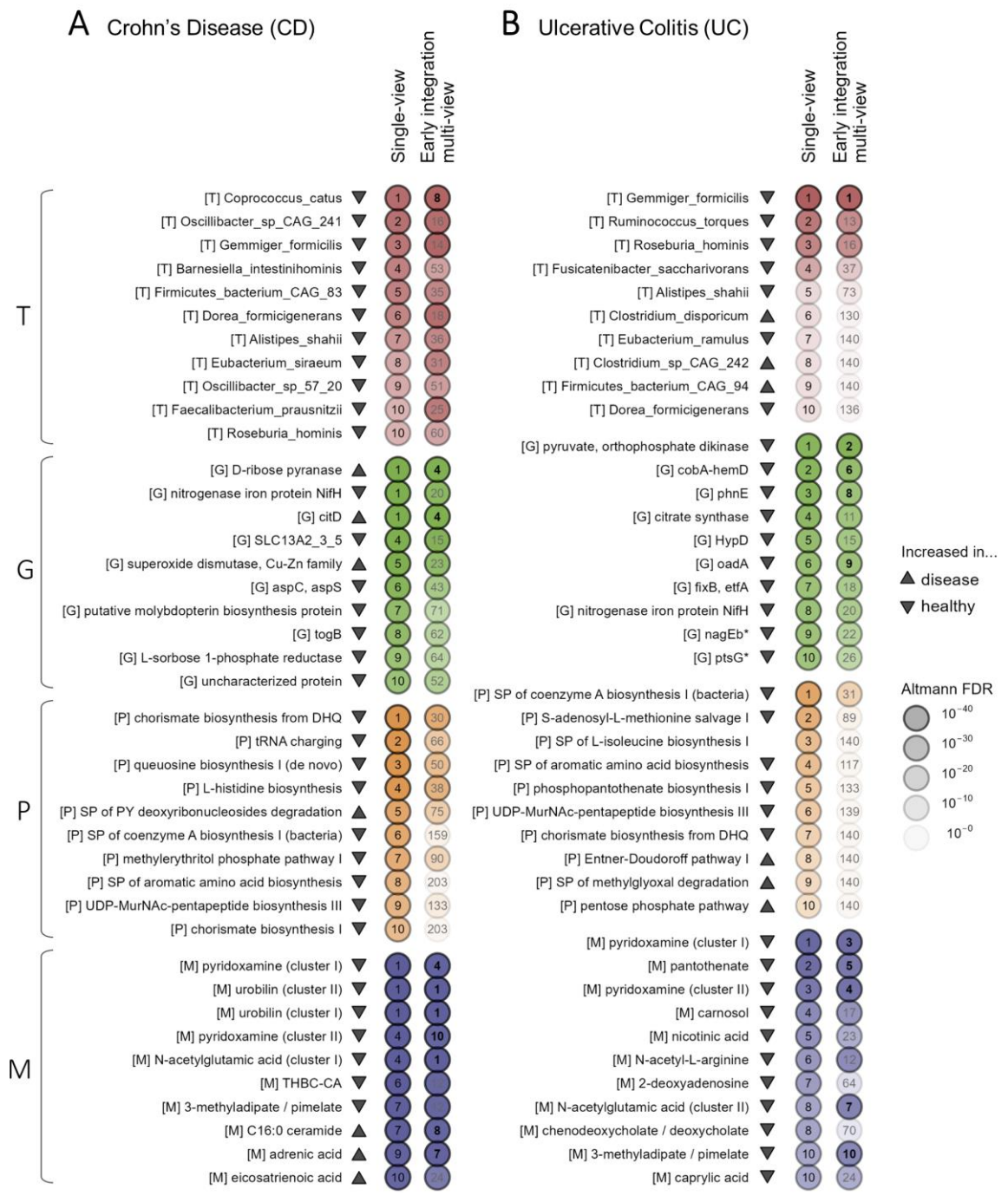

Figure 3: **Single-view vs. early integration multi-view feature importance in predicting IBD**. Feature importance analysis is presented for CD (A) and UC (B), using Altmann's method applied to single-view ML models (left columns), or to the early integration multi-view ML models (right columns; Methods). A union of the top 10 single-/multi-view contributors are shown for each dataset, grouped by feature type (T/G/P/M prefix, colors as in Figure 1A). 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 (no triangle indicates no statistically 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: Super-pathway; 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: non-discriminating 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

From early to intermediate integration: Introducing *MintTea*

While the identification of important features in the early integration multi-view models above provided insights into microbial markers of disease across different layers, it may fail to account for important information concerning the interaction between these layers or concordant disease-associated shifts in different views. This caveat is especially concerning when aiming to identify cohesive, distinct, and interpretable multi-omic disease signatures, that go beyond one mixed pool of informative features from the various views. Put differently, we wish to put forward a method that can identify multi-omic *modules*, each representing a set of features from multiple views that capture both shared trends across the various omics and association with disease.

One alternative integration approach that can be utilized for such a task is 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. Canonical correlation analysis (CCA), for example, is one popular intermediate integration method that was previously extended and applied to microbiome-related multi-omic data, specifically using sparse CCA (sCCA) – a CCA extension that included sparsity constraints, or sparse generalized CCA (sGCCA), which generalizes sCCA to further support more than two views[57,58]. Importantly, the

243 transformation process itself can further take the disease state of each sample into account,
244 thus extracting representations that highlight both interactions across different views and
245 between views and the disease (as in, for example, Refs[58,59]). Unfortunately, CCA and its
246 extensions are highly sensitive to small perturbations in the data, atypical samples,
247 collinearity between variables, and potentially non-unique solutions[57,60].

248 To address this challenge and present a framework that can identify cohesive, disease-
249 associated, multi-omic *modules,* we developed an intermediate integration framework, termed
250 *MintTea* (Multi-view INTegration Tool for microbiomE Analysis), based on sparse
251 generalized canonical correlation analysis (sGCCA) and inspired by previously introduced
252 methods[58,59]. Specifically, MintTea's analysis identifies sets of features from multiple
253 different views that are both strongly associated with the disease and with each other.
254 Moreover, MintTea-identified modules aim to account for noisy data and to ensure robust and
255 confident results. Briefly, the framework receives two or more feature tables describing
256 different views (or omics) for the same set of samples, and a label for each sample (e.g.,
257 healthy vs. disease). It then searches for a *sparse* linear transformation per feature table that
258 yields maximal correlations between the respective latent variables, as well as between these
259 variables and the label. This first set of latent variables is referred to as the first component,
260 and additional components can be computed iteratively such that each new component is
261 orthogonal to the previous ones. This sGCCA variant, which encodes the label as an
262 additional view, was ran using DIABLO's implementation of this variant[59]. Importantly, to
263 identify subsets that are robust to small changes in the input data, the entire process is
264 repeated multiple times on random data subsets (e.g., 90% of the samples), and the final
265 *multi-view modules* are defined as sets of features from different views that consistently co-
266 occurred in the same component ("consensus modules"; see Methods). To assess the overall
267 association of each identified module with the disease, MintTea further tests the combined
268 predictive power of each module, using the first principal component (PC) calculated on all
269 module's features and subsequently computing its AUC. Lastly, MintTea compares the true
270 AUC to AUCs obtained using randomly selected modules (of the same size and distribution
271 over views), effectively serving as null models (Methods). The final outcome for each dataset
272 is therefore multiple multi-view modules, each capturing features from all views that are
273 highly associated within view, between views, and between views and phenotype. The entire
274 pipeline, including data preparations steps, is depicted in Figure 4A.

275 In Supplementary note 1, we discuss the conceptual differences between MintTea and other
276 closely related approaches, and specifically compare MintTea to DIABLO[59] and sGCCA[58] in
277 terms of module-disease association, between-view associations, stability of obtained
278 modules, false discovery rate, and orthogonality of modules. We found that MintTea provides
279 a good balance across multiple desired properties, while achieving a substantially lower false
280 discovery rate (Supplementary Figure S7).

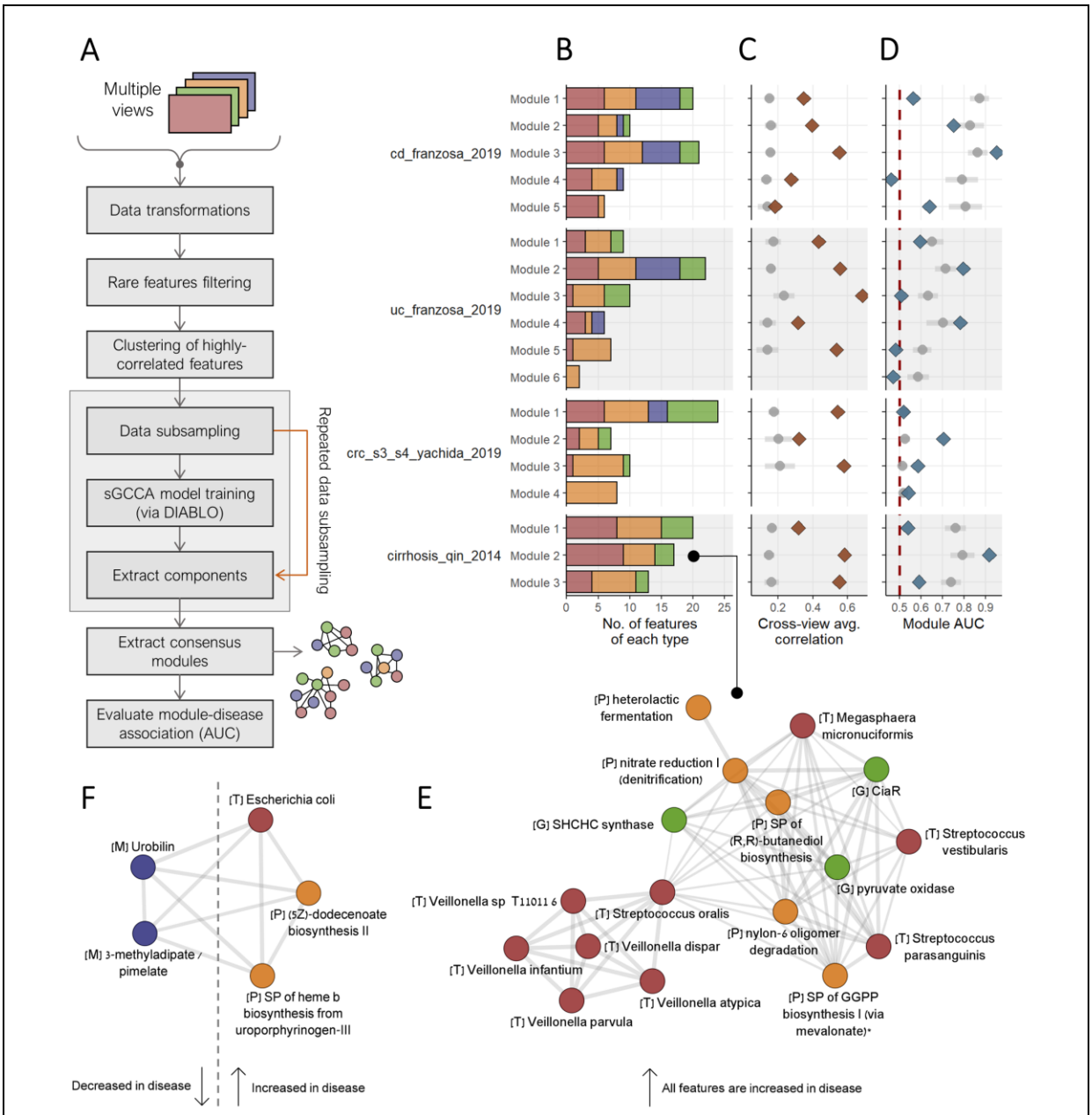

Figure 4: Identifying disease-associated multi-view modules via MintTea intermediate integration framework. (A) A diagram of the MintTea pipeline. (B-D) Properties of the multi-view modules obtained by applying MintTea to several key datasets, including (B) the number of features

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. Node colors represent feature type (T/G/P). Edges connect features that appeared together in a sGCCA component over 80% of data subsampling iterations. (F) A sub-module of mixed features, shared between a CD-predictive module (module 3) and a UC-predictive module (module 2).

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.

Using MintTea for identifying multi-view disease-associated modules

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 4D, 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 a single feature's contribution to disease prediction, but also take into account interactions between features, balancing between these two objectives. A detailed list of features in each module is available in Supplementary Table S6.

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 oxidative stress (OS), a key driver of cirrhosis[61–63]. The bacterial species driving this module are typically found in the human oral cavity, including several *Streptococcus*, *Veillonella*, and *Megasphaera* species (both commensal and

302 opportunistic pathogens)[64], all of which previously found to be increased in stool samples of
303 individuals with liver diseases[65–67], as well as other gastrointestinal disorders[68–70]. We also
304 observe that several functional features (genes and pathways) in the module are *Streptococci*-
305 related and/or oxidative stress related, including the pyruvate oxidase gene, shown to be
306 possessed by multiple *Streptococcus* species and produce ROS[71], CiaR, a signaling
307 mechanism highly conserved in *Streptococci* that has a central role in the regulation of stress
308 tolerance[72], and nitrate reduction, which promotes oxidative stress and has been confirmed in
309 almost all of the *Streptococcus* and *Veillonella* species appearing in the module[73]. We
310 hypothesize that the *Veillonella* species take part in the module mostly due to their strong co-
311 abundance patterns with various Streptococcus species. Indeed, a recent study found that
312 *Streptococcus* and *Veillonella* species were co-abundant across multiple large-scale cohorts[74],
313 also in line with known metabolic interactions between them[75]. *Veillonella* species have also
314 been shown to reduce oxidative microenvironments, thus facilitating the growth of
315 *Streptococci* and others[76]. Other features such as the (R,R)-butanediol biosynthesis
316 superpathway are most likely included in the module as they are primarily driven by
317 *Streptococcus* species, as implied by HUMAnN3's taxon-stratified output. Overall, this
318 module suggests a potential multi-layer stress-related mechanism of specific bacterial species
319 that is highly associated with liver cirrhosis.

320 We further checked whether there are significant overlaps between modules from the four
321 different datasets and disease states (Supplementary Table S7). One significant overlap
322 (Fisher's exact test FDR value $7 \cdot 10^{-7}$) included 5 features shared between CD-associated
323 module 3 and UC-associated module 2 (both of which were highly predictive of disease,
324 AUC > 0.8, Figure 4F), including *Escherichia Coli*, heme b biosynthesis pathway (MetaCyc
325 PWY0-1415), (5Z)-dodecenoate biosynthesis pathway (PWY-7858), urobilin metabolite and
326 3-methyladipate/pimelate metabolite. Of these, the first 3 are significantly elevated in IBD
327 fecal samples and the latter 2 are decreased. Significant correlations were detected between
328 these features (either positive or negative) even when controlling for study group (linear
329 regression FDR values $< 1 \cdot 10^{-4}$, for all feature pairs). We also observed that the majority of
330 metagenomic reads mapped to the two pathways were indeed contributed by *E. coli* (as
331 reported by taxon-stratified HUMAnN3 outputs). We postulate that this set of features reflect
332 a heme-related mechanism led by specific heme-metabolizing *E. Coli* strains. Indeed,
333 intestinal inflammation has been shown to result in decreased enteric iron and increased
334 enteric heme availability, promoting an environment favorable for heme utilizing bacteria

335 such as *E. Coli*[77–79]. Furthermore, fecal urobilin, a downstream product of heme degradation
336 (via bilirubin), is significantly decreased in IBD patients, both in this dataset and in
337 independent cohorts[25,69], in accordance with reduced host heme catabolism. *E. Coli* has been
338 repeatedly reported to be overrepresented in IBD patients, based on both fecal samples[25] and
339 mucosal biopsies[79,80], often with indications of mucosa invasion in CD specifically[81]. It has
340 been further suggested that *E. coli* has a causative role in intestinal inflammation, as opposed
341 to merely colonizing inflamed mucosa[82]. We also note that Urobilin, Pimelic acid, and 3-
342 Methyladipate were all found to be highly associated with the gut microbiome in multiple
343 independent datasets, in a meta-analysis study we have previously conducted[83]. Though fecal
344 Pimelic acid / 3-Methyladipate were also previously shown to be significantly decreased in
345 IBD patients[25,69], we could not recover a specific mechanistic connection to *E. Coli* and heme
346 utilization.

347 Lastly, we evaluated the overall predictive power of all modules in each dataset, by training
348 ML models on the 1st PCs of all modules, and estimating their combined performance with
349 cross-validation. We found that the multi-view modules by themselves achieved an AUC
350 comparable to that of the previous models in CD and cirrhosis datasets, while utilizing a
351 significantly smaller number of features in total (Supplementary Table S5, Supplementary
352 Figure S8). In the liver cirrhosis dataset, for example, MintTea modules achieved an AUC of
353 0.905, based on 50 features in total that were further translated into just 3 values, representing
354 the 1st PC of each of the 3 modules that these 50 features were grouped into. In comparison,
355 the early integration approaches achieved a slightly better AUC of 0.929, but used 383
356 features for training the ML model after feature selection. Similarly, the overall AUC of the
357 CD modules was 0.961, compared to 0.958 using early-integration, utilizing 66 features (in 5
358 modules) compared to over 1900 used for ML training. Finally, though AUC performance in
359 the UC and CRC datasets was lower than that of previous early integration models, AUC
360 estimations in all cases were significantly higher than those reached when training similar
361 models on random modules (Supplementary Figure S8).

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

364 Having shown that MintTea was capable of identifying multi-view modules exhibiting
365 associations both in-between views and with a disease, we finally applied MintTea to all

366 shotgun datasets described above (i.e. a total of 17 datasets), to map the landscape of such
367 multi-view modules across cohorts in general and across disease states. This analysis has
368 identified several multi-view modules per dataset (average: 5.24), with a varying number of
369 features (ranging between 2 and 95 with an average of 14.43, Figure 5A). All modules are
370 described in Supplementary Tables S5 (modules' descriptive statistics) and S6 (features in
371 each module). As expected, the vast majority of modules captured strong between-view
372 correlations that were significantly higher than expected by chance (Figure 5B). We also
373 examined which of these multi-view modules were further predictive of the disease state
374 (AUC > 0.7 and significantly higher than random modules), identifying 1-2 such modules in
375 most datasets (Figure 5C). Importantly, although no highly predictive modules were
376 identified for certain datasets, different pipeline settings may have resulted in predictive
377 modules for these datasets as well (see Methods).

378 Notably, examining the modules in the above catalogue may provide novel insight into
379 microbiome-disease associations. For example, in a cohort of individuals from rural
380 Cameroon populations[84], some of whom suffer from soil-transmitted helminths (STH;
381 intestinal worms that infect humans through contaminated soil), we identified two STH-
382 associated modules (AUC 0.71 and 0.82 for modules 1 and 2, respectively) that may hint at
383 two distinct or complementary infection signatures, possibly related to different STH types or
384 different population characteristics. Supportive of this hypothesis, module 1 includes
385 *Treponema succinifaciens* which is increased in hunter-gatherers and was also found to be
386 positively associated with Trichuris (whipworm) infections specifically[84,85]. The module also
387 includes known *Treponema* genes (K10190, K10843). Remarkably, only 2 of the 14 features
388 in this module contributed to the early-integration ML model, further demonstrating the
389 potential value of intermediate integration approaches. As another example, module 2 in the
390 myalgic encephalomyelitis / chronic fatigue syndrome (ME/CFS) cohort[86] achieved an AUC
391 of 0.73 and included 4 species, 3 pathways, and 1 KO, with 7 of the 8 features significantly
392 elevated in ME/CFS patients (one pathway was not significantly differentially abundant). The
393 module includes the *Bacteroides fragilis* species, which was specifically elevated in patients
394 that do not also suffer from irritable bowel syndrome (IBS), a highly-common comorbidity of
395 ME/CFS that may confound ME/CFS-microbiome associations. Two additional species in the
396 module, as well as pathways of aliphatic amino acid biosynthesis (PWY-5101, P163-PWY),
397 were previously implicated in depression and/or fibromyalgia, both of which closely related
398 to ME/CFS[87–89].

399 Following the example in Figure 4E, we next mapped all significantly overlapping modules
400 across all 17 datasets (including both disease-associated and not disease-associated modules).
401 We identified a total of 119 significant overlaps (Fisher's exact test FDR-corrected *P* value <
402 0.1, and at least 3 features in common), suggesting highly prevalent multi-view associations
403 between microbiome features (Supplementary Table S7, Figure 5D). These could potentially
404 represent cohesive and universal building blocks of the microbiome's composition and
405 function. We also identified specific bacterial species that appeared in multiple modules of
406 different datasets. *Bifidobacterium longum*, for example, a key commensal in the human
407 gastrointestinal tract and a common probiotic, appeared in modules of 10 different datasets
408 (though not disease-associated in their corresponding datasets). Perhaps not surprisingly, *B.*
409 *longum* tended to appear with metabolic pathways related to dietary carbohydrate
410 metabolism, including glucose fermentation to lactate (also termed '*Bifid* shunt'[90]) and starch
411 biosynthesis. Indeed it is estimated that approximately 12% of the annotated open reading
412 frames in *Bifidobaterium* genomes are assigned to carbohydrate-metabolism related
413 functions[91]. The methylerythritol phosphate pathway (MetaCyc: NONMEVIPP-PWY), and
414 its superpathway of geranylgeranyl diphosphate biosynthesis were also consistently included
415 in these modules, in accordance with fact that most of the reads assigned to this pathway
416 were contributed by *Bifidobaterium* species, as determined by stratified HUMAnN3 outputs.
417 Interestingly, other species commonly appearing with *B. longum* in MintTea modules were
418 also from the *Bifidobacteriaceae* family, including *B. adolescentis*, *B. pseudocatenulatum*,
419 and *Aeriscardovia aeriphila* (previously classified as belonging to the *Bifidobacterium* genus
420 as well[92]). Taken together, these modules suggest a common axis of variation shared between
421 multiple cohorts and characterized mainly by *Bifidobacterium* species and carbohydrate
422 metabolism. Two other examples of species appearing in multiple modules, namely
423 *Flavonifractor plautii* (which was also consistently over-abundant across different disease
424 states) and *Klebsiella pneumoniae*, are provided in Supplementary Figure S9. Though some
425 features consistently co-occurred with these species across modules, most features appeared
426 in a single module only, suggesting that different datasets contain different feature
427 dependencies even in relation to the same bacterial species.

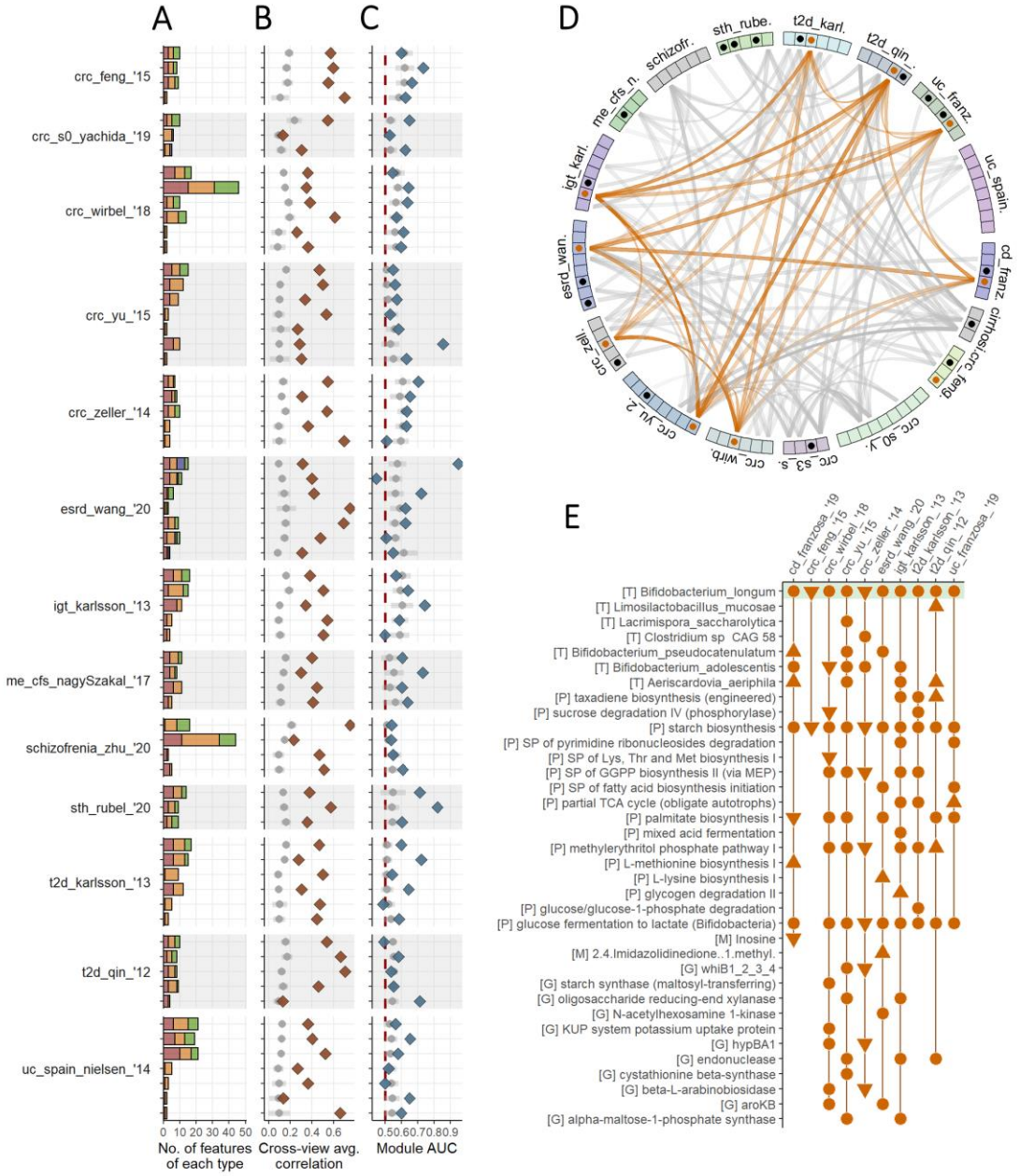

Figure 5: Multi-view modules across datasets and disease states. (A-C) Overview of multi-view modules identified for the 13 additional shotgun datasets (adding to the 4 datasets shown in Figure 4B-D). See legend of Figure 4B-D. Modules with features from a single view only were discarded and are not 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: super-pathway; 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[93] 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[26,41,94]. For each of these studies we ran HUMAnN3[20] to obtain KEGG Orthology (KO) profiles and MetaCyc[95] pathway profiles. MetaPhlAn3[20] 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.[41] were split into Chron's disease (CD) and Ulcerative Colitis (UC) and CRC data from Yachida et al.[26] 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 ML modelling for each sub-type/stage separately, as we performed here.

Eighteen processed shotgun metagenomics datasets were extracted from the *curatedMetagenomicData* R package[93], version 3.2.3, together with the corresponding

452 sample metadata. Specifically, MetaPhlAn3[20] species-levels abundances and HUMAnN3[20]
453 unstratified gene-level (UniRef) and pathway-level (MetaCyc) abundances were obtained.
454 We only considered datasets with data from at least 40 stool samples from different subjects
455 in each study group (as defined by the ”study_condition” variable). Longitudinal studies were
456 excluded. We mapped UniRef gene abundances into KO abundances using the mapping file
457 provided by HUMAnN3. Additional details for each dataset are available in Supplementary
458 Table 1.

459 We further downloaded 12 stool 16S amplicon sequencing datasets from the *MicrobiomeHD*
460 collection[6]. Only studies with at least 40 cases and at least 40 controls were taken. Each
461 dataset included sample metadata, 16S OTU abundances (clustered by 100% similarity using
462 USEARCH[96]), and taxonomy assignments at the genus level (computed by RDP
463 classifiers[97]). To further generate functional profiles for each 16S dataset, we ran PICRUSt2[49]
464 with default settings and obtained KO and MetaCyc pathway profiles. Additional details on
465 each dataset are available in Supplementary Table 1.

466 In total, data analyzed in this study included 5,702 samples, obtained from 28 studies, and
467 partitioned into 37 case-control datasets, covering 16 disease states. Of these 5,702 samples,
468 759 included shotgun metagenomics and metabolome data, 2,389 shotgun metagenomics data
469 only, and 2,554 16S data only. For most analyses, only shotgun metagenomic datasets were
470 used.

471 *Shared data preprocessing*

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

482 *Predicting disease state using microbiome features*

483 We implemented a standard ML pipeline to assess disease predictability in each dataset and
484 using each feature set. Training was performed with a repeated 5-fold cross validation (10
485 repeats) procedure and included a feature selection step (as described below), training a
486 random forest (RF) classifier, evaluating model performance on out-of-fold data using
487 receiver operating characteristic (ROC) area under the curve (AUC), and evaluating feature
488 importance (as described below). We used the 'ranger' R package[98] for RF modelling with a
489 "min_n" of 8 and otherwise default parameters. We opted for RF due to its simplicity on the
490 one hand and its tendency to perform well on microbiome abundance data as consistently
491 shown in previous studies[1,4,6,48,99,100]. Hyper-parameter tuning for the RF classifiers with a
492 nested cross-validation approach did not result in significantly improved performance for
493 most datasets and was therefore discarded.

494 *Feature selection*

495 To reduce data dimensionality and risk of overfitting, we dropped non-relevant features,
496 using 4 different 'all-relevant' feature selection approaches (including simple univariate tests,
497 feature importance-based methods, and Boruta[43]), as well as no feature selection for
498 comparison. We selected the "all-relevant" approach so that downstream analysis of the top-
499 performing features will not be subjected to arbitrary removal of redundant features that may
500 nonetheless be biologically relevant. Details about the different feature selection methods are
501 described in Supplementary Note 2. Eventually, feature selection was carried out with the
502 repeated 10-fold cross-validation, running independently for each fold. The best-performing
503 method per dataset was maintained for downstream analysis (see Supplementary Table S2).

504 *Model evaluation and comparison*

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

510 For each dataset, single- and multi-view models' performances were compared using a
511 repeated measures ANOVA to determine whether any significant differences in AUC are
512 present. Subsequent paired t-tests were used to determine which specific models

513 outperformed others. We specifically used a *repeated measures* framework to link models
514 trained and tested on the same data folds.

515 We also verified our AUC results against previous studies that trained microbiome-based
516 classification models for the same datasets and using similar features, either in the original
517 studies from which data were obtained or in meta-analyses studies[1,4,6,26]. We only performed
518 comparisons when the authors reported an average AUC using some form of cross-validation
519 (as opposed to an AUC on an external validation dataset). Notably, the reference models were
520 often based on different data transformations, different feature selection methods, and
521 different ML algorithms. See Supplementary Figure S4.

522 *Feature importance analysis*

523 We examined the features importance in each model that achieved an AUC > 0.7, as follows:
524 Within each fold, we calculated permutation-based feature importance scores implemented in
525 'ranger' package[98]. We additionally applied the method introduced by Altmann et al.[50] to
526 assign a *P* value for each feature in each model. Final feature importance per feature was
527 aggregated over repeats and folds. Final *P* values were FDR-corrected over all features in
528 each feature set. We refer to features that were selected by the feature selection method in at
529 least 50% of repeats/folds and obtained a final Altmann FDR < 0.1 as "contributors".

530 *Identifying multi-view disease modules*

531 To identify multi-view sets of features that are both tightly associated with each other and
532 collectively associated with the disease, we developed a framework based on sparse
533 generalized canonical correlation analysis (sGCCA)[57,58]. We also utilized the recently
534 introduced DIABLO method, which extends sGCCA to a supervised scenario[59,101].
535 Supplementary Note 4 provides detailed explanation and formalization of sGCCA and its
536 variants. Briefly, when applied to multiple views and a label of interest, DIABLO identifies
537 sparse linear transformations of each view (termed 'components') so that the resulting latent
538 variables are highly correlated between views and with the label. One key limitation of such
539 linear models, however, is that components tend to be highly unstable, especially in the
540 presence of thousands of features and relatively small sample sizes. Indeed, we observed that
541 small changes in the input data (for example excluding a few samples) resulted in
542 substantially different components. We therefore wrapped DIABLO's core sGCCA
543 computation with a repeated sub-sampling procedure to identify "consensus components"
544 shared across multiple runs. Specifically, we ran 100 iterations, in each one only a random

545 sample of 90% of the samples was used for training. For each such run we recorded the list of
546 features with non-zero coefficients from each view, for each of the first 5 components
547 (adding additional components did not change downstream consensus components as they
548 tend to be highly unstable). We then constructed a network of multi-view features where 2
549 features are connected by an edge if they appeared together in a component in at least 80% of
550 the iterations. Connected components of size ≥ 2 were considered "consensus components"
551 which we refer to as "multi-view modules". We term the entire pipeline MintTea (Multi-view
552 INTegration Tool for microbiomE Analysis).

553 To assess the extent to which each MintTea-identified module is associated with disease
554 state, we applied principal component analysis (PCA) to all module features and took the first
555 PC as a representative of the module. We then computed the AUC for this first PC of each
556 module, and compared it to "null" modules where the same number of features, from the
557 same views, were randomly selected. Lastly, we trained a RF model on all modules to
558 estimate the overall predictive power of modules in each dataset. Classification performance
559 (measured by ROC AUC), of both single modules and all modules together, was estimated as
560 before using repeated 10-fold cross-validation.

561 To assess MintTea's sensitivity to various settings, we ran the entire pipeline using various
562 parameter combinations. The selected setting tended to perform well across most datasets as
563 measured by overall AUC's (Supplementary Figure S10A). We additionally confirmed that
564 using different settings tended to either expand, narrow, or split the original modules, rather
565 than producing unrelated ones (Supplementary Figure S10B-C). For a few datasets in which
566 the obtained multi-view modules were extremely large, we used pipeline settings that were
567 slightly more conservative (Supplementary Table S5).

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

571 The pipeline was developed in R using the packages "mixOmics"[101] for DIABLO running
572 ("block.splsda" function), and "igraph"[102] for graph operations. Supplementary Note 1 further
573 describes a conceptual and quantitative comparison between MintTea and related methods.

574

575 # Discussion

576 In this work, we focused on the application of methods from the domain of multi-view
577 learning to microbiome multi-omic data. Specifically, after exploring the performances of
578 models that do not integrate multi-omic data or that use a naïve "early" integration approach,
579 we developed a novel integrated ML algorithm for simultaneously predicting host's disease
580 based on multiple gut microbiome views and identifying cohesive, multi-omic disease
581 signatures. Our "views" included taxonomic, gene-family, and metabolic pathway profiles
582 derived from shotgun metagenomics data, as well as metabolomic profiles. Analyzing 25
583 case-control datasets, we found that simply early integration multi-view models typically
584 resulted in performance comparable to the best single-view model, but further provided a
585 mixed set of informative features, spanning different views. Importantly, however, using our
586 new "intermediate integration" MintTea method, we were able to go beyond a pooled list of
587 top informative multi-view features and pinpoint cohesive multi-view modules of features
588 that reflected coordinated trends in the data across views, while also being highly predictive
589 of disease. Indeed, MintTea serves as a proof-of-concept for the potential and value of
590 applying advanced intermediate integration methods to microbiome multi-view data to
591 generate transparent and multifaceted biological hypothesis (see Supplementary note 1 for a
592 discussion about differences and similarities between MintTea and other related approaches).

593 Our systematic analyses and assessment of various integration approaches (including no
594 integration, early integration, and intermediate integration) also provided several more
595 general and intriguing insights concerning the association between microbiome features and
596 disease. For example, a perhaps surprising observation in our study was that combining
597 multiple views for host disease prediction with early integration did not necessarily improve
598 prediction accuracy. Nonetheless, similar results have been reported on various biological
599 domains[103,104]. We assume that the lack of increased performance, in our analyses, stems
600 from the large redundancy between different microbiome representations, especially in the
601 presence of mechanisms that involve all types of features. In a way, this makes the use of
602 intermediate integration methods even more fitting, as it targets both inter-view associations
603 and disease-related information simultaneously. Notably, different ML pipeline choices, such
604 as an alternative feature selection approach or ML algorithm, may have resulted in slightly
605 different results. Comparing the single-view models to one another, we also observed that
606 different views performed best for single-view prediction for different datasets, though the
607 taxonomic view tended to outperform the functional ones. We emphasize that this result does

608 not necessarily imply that the former is truly more clinically important than the latter.
609 Functional profiles at the gene-level are likely too-high-resolution (e.g., include thousands of
610 gene families), making it challenging for many ML models (even when applying feature
611 selection), or too-low resolution (e.g., aggregated at the pathway level), masking more subtle
612 disease-related signals[105].

613 Intermediate integration methods, an alternative (or complementary) approach to early
614 integration, have been gaining increased attention due to their ability to uncover intricate
615 multivariable interactions, as desperately required when dealing with complex biological
616 systems[106]. Our motivation to use a method based on sCCA was the assumption that gut
617 microbiome-disease associations are explained by one or more "modules" of features from all
618 views that take part in a specific mechanism effected by or promoting host disease.
619 Furthermore, sCCA-based frameworks have been previously applied to various microbiome
620 multi-view data with insightful results[27,107]. To overcome challenges related to the high-
621 dimensionality of the data and consequent instability of results, however, we further wrapped
622 this analytical technique with repeated iterations using data subsamples and consensus
623 analysis, to identify *robust* multi-view modules based on multiple sCCA components. Using
624 this approach, we were able to pinpoint several such multi-view modules in each dataset.
625 Taking the liver cirrhosis dataset as one example, we identified a cirrhosis-associated module
626 of features including bacterial species typically characteristic of the human oral microbiota,
627 as well as oxidative-stress related functions. Notably, we also showed that using solely the
628 first principal component of this one module, *on its own*, achieved an AUC of 0.916, showing
629 our framework's ability to simultaneously expose both inter-view and disease-related
630 associations and further hint at biologically meaningful mechanisms.

631 Applying MintTea to our previously described collection of datasets, and comparing the
632 modules identified in each dataset, we identified multiple modules with a significantly high
633 overlap, as well as specific microbiome features that appeared in modules of multiple
634 datasets. *B. longum*, for example, appeared in modules of 10 different datasets, consistently
635 co-occurring with dietary carbohydrate metabolism and other pathways. These modules,
636 alongside additional groups of overlapping modules, suggest common axes of variation in the
637 gut's microbiome composition and function, shared across multiple cohorts and representing
638 cohesive, multi-layer microbiome modules. In a way, these modules can thus be thoughts of
639 as the microbiome's fundamental building blocks or "modes" of operation.

640 Clearly, any attempt to model disease based on metagenomic and other features can only be
641 as accurate and reliable as the underlying features. Though there are several well-established
642 tools for taxonomic and functional profiling from metagenomics data, they may each induce
643 biases and inaccuracies[108]. Moreover, additional, possibly invaluable features can be
644 recovered from metagenomics and metabolomics data, for example by using assembly-based
645 methods to add unknown-genome abundances to taxonomic profiles[108], or adding
646 unannotated metabolites from untargeted metabolomic profiles. Another limitation is the
647 insufficient control for major confounders known to exert microbiome shifts (e.g., stool
648 consistency and medication usage)[109]. This limitation could be potentially mitigated by
649 careful sample stratification, inclusion of these factors as covariates in models, or even
650 adding such data as an additional "view". Diet, for example, is likely a substantial mediator
651 of many of the identified microbiome-metabolome modules, and it could therefore be
652 especially interesting to include "diet"-view profiles in future analyses to elucidate potential
653 diet-driven mechanisms. Lastly, we note that a major limitation of intermediate integration
654 methods in general is the difficulty to systematically validate the identified modules, largely
655 since systematic mapping of such multi-omic mechanisms is lacking. In this work for
656 example, we suggest methods to evaluate the between-views associations in each module, as
657 well as the overall module's association with the disease, but a validation of the biological
658 coherence of the modules is nonetheless limited.

659 Overall, this work serves as a proof of concept for integrating multiple microbiome data types
660 in a single machine learning pipeline for disease classification and for generating hypotheses
661 concerning multi-view disease-associated modules. We advocate for using such multi-view
662 intermediate integration models to capture complementary biological insights into the
663 complex mechanisms underlying microbiome-disease associations.

664

665 References

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

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

670 3. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS*
671 *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. Beck, D. & Foster, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. *PLoS One* **9**, e87830 (2014).

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

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

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

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

12. 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).

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

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

15. 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).

16. 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).

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

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

19. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* (2023). doi:10.1038/s41587-023-01688-w

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

21. 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).

22. 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).

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

24. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phage-inclusive profiling of human gut microbiomes with Phanta. *Nat. Biotechnol.* 21–24 (2023). doi:10.1038/s41587-023-01799-4

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

26. 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).

27. 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).

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

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

30. 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).

31. 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).

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

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

34. 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).

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

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

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

741 38. Overmyer, K. A. *et al.* Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst.* **12**, 23-40.e7
742 (2021).

743 39. Chen, Y. *et al.* Human gut microbiome aging clocks based on taxonomic and functional signatures
744 through multi-view learning. *Gut Microbes* **14**, (2022).

745 40. Zhou, W. *et al.* Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**, 663–
746 671 (2019).

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

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

750 43. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13
751 (2010).

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

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

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

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

760 48. Oh, T. G. *et al.* A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metab.* **32**,
761 878-888.e6 (2020).

762 49. Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688
763 (2020).

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

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

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

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

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

774 55. Galipeau, H. J. *et al.* Novel Fecal Biomarkers That Precede Clinical Diagnosis of Ulcerative Colitis.

775 *Gastroenterology* **160**, 1532–1545 (2021).

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

778 57. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**,
779 569–583 (2014).

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

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

784 60. Wilms, I. & Croux, C. Robust sparse canonical correlation analysis. *BMC Syst. Biol.* **10**, 1–13 (2016).

785 61. Borrelli, A. *et al.* Role of gut microbiota and oxidative stress in the progression of non-alcoholic fatty
786 liver disease to hepatocarcinoma: Current and innovative therapeutic approaches. *Redox Biol.* **15**, 467–
787 479 (2018).

788 62. Cichoz-Lach, H. & Michalak, A. Oxidative stress as a crucial factor in liver diseases. *World J.*
789 *Gastroenterol.* **20**, 8082–8091 (2014).

790 63. Jones, R. M. & Neish, A. S. Redox signaling mediated by the gut microbiota. *Free Radic. Biol. Med.*
791 **105**, 41–47 (2017).

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

794 65. Wei, X. *et al.* Abnormal gut microbiota metabolism specific for liver cirrhosis. *Front. Microbiol.* **9**, 1–4
795 (2018).

796 66. Davis, B. C. & Bajaj, J. S. The Human Gut Microbiome in Liver Diseases. *Semin. Liver Dis.* **37**, 128–
797 140 (2017).

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

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

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

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

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

808 72. He, L. Y., Le, Y. J., Guo, Z., Li, S. & Yang, X. Y. The Role and Regulatory Network of the CiaRH

809 Two-Component System in Streptococcal Species. *Front. Microbiol.* **12**, 1–11 (2021).

810 73. Rosier, B. T. *et al.* The Importance of Nitrate Reduction for Oral Health. *J. Dent. Res.* **101**, 887–897
811 (2022).

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

814 75. Chalmers, N. I., Palmer, R. J., Cisar, J. O. & Kolenbrander, P. E. Characterization of a Streptococcus
815 sp.-Veillonella sp. community micromanipulated from dental plaque. *J. Bacteriol.* **190**, 8145–8154
816 (2008).

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

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

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

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

825 80. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**,
826 382–392 (2014).

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

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

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

832 84. Rubel, M. A. *et al.* Lifestyle and the presence of helminths is associated with gut microbiome
833 composition in Cameroonians. *Genome Biol.* **21**, 1–32 (2020).

834 85. Kupritz, J., Angelova, A., Nutman, T. B. & Gazzinelli-Guimaraes, P. H. Helminth-Induced Human
835 Gastrointestinal Dysbiosis: A Systematic Review and Meta-Analysis Reveals Insights into Altered
836 Taxon Diversity and Microbial Gradient Collapse. *MBio* **12**, 1–17 (2021).

837 86. Nagy-Szakal, D. *et al.* Fecal metagenomic profiles in subgroups of patients with myalgic
838 encephalomyelitis/chronic fatigue syndrome. *Microbiome* **5**, 44 (2017).

839 87. Minerbi, A. *et al.* Altered microbiome composition in individuals with fibromyalgia. *Pain* **160**, 2589–
840 2602 (2019).

841 88. Liu, R. T. *et al.* Reductions in anti-inflammatory gut bacteria are associated with depression in a sample
842 of young adults. *Brain. Behav. Immun.* **88**, 308–324 (2020).

843 89. Wang, Y. *et al.* Gut microbiota is involved in the antidepressant-like effect of (S)-norketamine in an
844 inflammation model of depression. *Pharmacol. Biochem. Behav.* **207**, 173226 (2021).

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

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

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

852 93. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**,
853 1023–1024 (2017).

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

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

858 96. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–
859 2461 (2010).

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

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

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

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

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

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

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

874 104. Culley, C., Vijayakumar, S., Zampieri, G. & Angione, C. A mechanism-aware and multiomic machine-
875 learning pipeline characterizes yeast cell growth. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 18869–18879
876 (2020).

877 105. Manor, O. & Borenstein, E. Revised computational metagenomic processing uncovers hidden and

878 biologically meaningful functional variation in the human microbiome. *Microbiome* **5**, 1–11 (2017).

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

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

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

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

886 110. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC*
887 *Bioinformatics* **9**, (2008).

888 111. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: Criteria of Max-
889 Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–
890 1238 (2005).

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

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

895 114. Osman, M. A. *et al.* Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and
896 Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer. *Sci. Reports 2021 111*
897 **11**, 1–12 (2021).

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

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

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

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

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

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

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

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

123. 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).

124. 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[93] (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[26,41,94] (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. 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: Conceptual and quantitative comparison between MintTea and other related methods

As described in the main text, MintTea's objective is to identify cohesive disease-associated multi-view microbiome modules, and as such aims to simultaneously capture both associations between views and associations with the host disease state. Though many methods that focus on either of these two aims exist, only a few methods tackle both objectives simultaneously, each taking a unique approach to achieve these two tasks, with various, and sometime subtle, differences. In the table below, we outlined the capabilities and key objectives of several different common approaches for identifying associations between the microbiome and a given disease and for pinpointing disease-associated signatures in the microbiome using multi-view data. Importantly, in this comparison, we focused on the *conceptual* differences between these approaches, rather than on their specific performances as captured by some evaluation metric. We specifically indicated for each method whether it is explicitly designed for multi-view data, whether it directly attempts to optimize classification ability, and whether it is easily interpretable in terms of pinpointing the most important features. We further noted whether each method also directly aims to optimize associations between the different views (or between features from different views). Assuming that informative features are identified by the method, we further indicated whether these features are partitioned or clustered into modules to generate distinct disease signatures, and also whether these modules are forced to include all views (even at the cost of reduced association) or not. Lastly, given that a major limitation of many statistical and machine learning (ML) methods is their sensitivity to noisy and high-dimensional data (as is the case for microbiome data), which in turn impacts reproducibility of the obtained results, we indicated whether each method inherently incorporates mechanisms to reduce likely spurious results.

To span the space of potential approaches, we included in the comparison both ML approaches for multi-view data, CCA extensions as previously discussed (see Supplementary note 4), as well as a reasonable heuristic approach that can be used to achieve a similar goal, based on first filtering out non-informative features and then clustering the remaining features

970 based on pairwise associations (using WGCNA for example, a highly popular method for
971 clustering highly correlated features).

Method	Description	Specific for multi-view data (≥ 2 views)	Optimizes classification accuracy	Pinpoints informative features (interpretable)	Optimizes inter-view associations	Partitions informative features into modules	Supports modules with a subset of views[1]	Inherently controls for high-dimensional and noisy data
ML: no integration	A ML model is trained for each view independently	✗	✓[2]	✓[3]	✗	✗	NA	✓
ML: early integration	A ML model is trained on a concatenation of all views	✓	✓	✓[3]	✗	✗	✓	✓
ML: late integration	A ML model is trained for each view independently, with a final ensemble model for final classification	✓	✓	✓[3]	✗	✗	✓	✓[4]
Feature selection + WGCNA[110]	Features are filtered based on disease-association, and WGCNA is applied to find correlated modules	✗	✓[5]	✓	✓	✓	✓	✗
sCCA / sGCCA / sMCCA[57,58]	A sparse version of classic CCA, and extensions of sCCA for ≥ 2 views	✓	✗	✓	✓	✓	✗	✗
DIABLO[59]	An extension of sGCCA to a supervised framework. See Ref[59]	✓	✓	✓	✓	✓	✗	✗
MintTea	Our framework	✓	✓	✓	✓	✓	✓	✓

ML: Machine learning; CCA: Canonical correlation analysis; sCCA: Sparse CCA; sGCCA/sMCCA: sparse generalized/multiple CCA; WGCNA: Weighted correlation network analysis.
[1] As opposed to requiring each module to contain features from all views.
[2] Does not leverage the availability of multiple views, hence optimization is only within each view.
[3] Depends on the selected ML algorithm and the selected explainability methods.
[4] Depends on the selected ML algorithm (e.g., whether bootstrapping/bagging mechanisms are involved).
[5] Features are filtered by their univariate association with disease. As a consequence, some of the discarded features may have actually contributed to classification accuracy when combined with others.

972 As clearly evident from this comparison, only a few of these approaches directly target, as
973 does MintTea, both aims described above. It is also important to note that while it may be
974 possible, in principle, to first apply some ML method that optimizes predictability and then,
975 as a second step, try to cluster informative features into modules (as in the "Feature selection
976 + WGCNA" entry in the table above), such a two-step approach may yield different or partial
977 modules, failing, for example, to identify features that are not sufficiently informative by
978 themselves but that can be incorporated into informative modules to enhance predictability
979 (as discussed in the main text).

980 A few other properties further differentiate MintTea from other approaches, including
981 specifically its ability to generate modules that do not incorporates all views (which may be
982 crucial in future studies when the number of omics further increases), and its inherent

983 subsampling and consensus detection procedure, which allows MintTea to better handle the
984 extremely high-dimensionality and noise associated with microbiome data. Indeed, this last
985 property is a key differentiator of MintTea from other, closely related, CCA expansions. Such
986 a repeated sub-sampling procedure forces outputted modules to be more robust to small
987 changes in input data and ultimately to provide more robust results. To quantify the impact of
988 this subsampling procedure on the obtained modules' quality, robustness, and false discovery
989 rates, we next compared MintTea's modules to those produced by the two CCA-based
990 method listed in the above table: sMCCA (as implemented by the 'MultiCCA' function in the
991 'PMA' R package[58]) and DIABLO[59]. For all comparisons, we used the liver cirrhosis dataset,
992 as it was one of the largest datasets at hand (237 samples), but application to other datasets
993 resulted in qualitatively similar trends. To allow comparison, we also tuned the penalty
994 parameters of each method so that the average component/module size (i.e. non-zero loadings
995 in MultiCCA/DIABLO) is approximately 20 features (from all views combined), with
996 DIABLO and MultiCCA further set to output 3 modules only.

997 We first compared the modules in terms of their ability to classify disease and the strength of
998 between-view associations in unseen data. To this end, we ran each model on 70% of the
999 samples, serving as a training set, and then computed the resulting modules on the remaining
1000 left-out 30% test set. For DIABLO and MultiCCA, the test set modules were computed using
1001 the trained loadings (feature weights), and for MintTea we took the 1st PC of all features from
1002 all views (as done in the main text). Notably, DIABLO and MultiCCA result in one predicted
1003 variable *per view* (3 views in this case) that are supposedly highly correlated, and so
1004 evaluation is carried out on each such vector separately and then averaged. The entire
1005 procedure was repeated 100 times to ensure the results consistency and significance. We
1006 found that modules from all 3 methods achieved similar disease classification performance
1007 (with a notable, yet not statistically significant advantage for MintTea), with an AUC of 0.69
1008 (standard deviation $\pm$0.11), 0.67 ($\pm$0.08) and 0.74 ($\pm$0.14) for DIABLO, MultiCCA and
1009 MintTea, respectively. This suggests that while all these methods aim to capture cohesive
1010 modules, this is not achieved at the cost of prediction accuracy. Similarly, modules were
1011 comparable in their average correlation between features from different views (0.47$\pm$0.1,
1012 0.49$\pm$0.14, 0.45$\pm$0.1, respectively), suggesting that methods capture similar between-view
1013 associations.

1014 Second, we compared each method's stability when trained on different data subsets
1015 (Supplementary Figure S7A-B). Here, we ran each method on two randomly selected data

1016 subsets of the same size and compared the resulting modules, assuming that more similar
1017 modules imply higher reproducibility/robustness of the method. The two sets of modules
1018 were compared both in terms of individual features included in the modules (Supplementary
1019 Figure S7A) and in terms of pairs of features that co-occurred in the same module in both sets
1020 (Supplementary Figure S7B). In the latter comparison, we used a z-score to quantify how
1021 likely were *pairs* of features co-occurring in the same module in one output to also co-occur
1022 in the same module in the other output at random. The data subsets used ranged from 50% of
1023 the samples to 100%, with stability estimations increasing as subset size increased, as
1024 expected. Overall, this analysis suggests that DIABLO tends to be less stable on this dataset,
1025 MultiCCA is slightly more stable in terms of selected features while MintTea is slightly more
1026 stable in terms of feature co-occurring in the same module.

1027 Importantly, one main caveat of CCA-based methods is that, by design, they always output a
1028 set of components, even when no correlation structure in the data actually exists. When using
1029 these methods, users are thus required to carefully evaluate the obtained modules themselves,
1030 and be aware of potentially high false-positive rates especially when the number of features
1031 greatly exceeds the number of samples (again, an unfortunately common setting in
1032 microbiome studies). To demonstrate this point, we assessed the false discovery rates of all
1033 methods by applying them to a shuffled version of the data where each view was shuffled
1034 independently (row-wise) so that any existing correlations between views (and between
1035 views and disease) were removed. We kept only DIABLO and MultiCCA components that
1036 resulted in significant between-view correlations (Bonferroni-corrected P value < 0.01), and
1037 only the features in each module that had a non-redundant loading (> 0.001). Supplementary
1038 Figure S7C illustrates the results obtained over 100 random samples of 70% of the data.
1039 Evidently, even when strictly filtering the components outputted by DIABLO/MultiCCA,
1040 their false discovery rates were still significantly higher than that of MintTea, demonstrating
1041 that MintTea is markedly more conservative and reliable.

1042 Lastly, we also evaluated the orthogonality of the different modules produced by each
1043 method (Supplementary Figure S7D). Orthogonal modules indicate complementary
1044 information captured by each module, while correlated modules indicate redundancy (i.e.,
1045 that different modules do not actually capture distinct data signatures). DIABLO, by design,
1046 outputs perfectly orthogonal components. MultiCCA, in contrast, resulted in highly
1047 correlated modules (often with high overlap of features). MintTea, perhaps not surprisingly,
1048 resulted in partial orthogonality, lying between that of DIABLO and that of MultiCC.

1049 In summary, we argue that MintTea's dual objective, both identifying associations with
1050 disease and uncovering relationships between different views, distinguishes it from many
1051 other supervised and unsupervised methods, and that even when compared to closely related
1052 CCA-based methods, MintTea represents a successful balance between multiple properties
1053 and metrics while maintaining substantially lower false discovery rates.

1054 Supplementary note 2: "All-relevant" feature selection approaches

1055 We compared 4 different feature selection approaches, all of which are designed to retain all
1056 informative features (weakly informative as well). Notably, in cases where classification
1057 accuracy is the only objective, other approaches that reduce redundancy should probably be
1058 tested and even favored (e.g. mRMR[111], or Recursive Feature Elimination), as ML models
1059 may be prone to overfit in face of large redundancy in the data[112]. The first approach was
1060 based on a univariate statistical test of differential abundance between cases and controls.
1061 Such an approach is commonly referred to as a "filter" feature selection method. We
1062 specifically chose the Mann Whitney U-test, running it on either relative abundances or CLR-
1063 transformed values and taking all features that obtained a *P* value below a specified
1064 threshold.

1065 The second approach was a modified Boruta feature selection[43]. Roughly, the Boruta
1066 algorithm adds "shadow features" to the original dataset, which are permutations of the
1067 original features. This way the distributions of the true data are maintained but the
1068 associations with disease, if existed, are removed. The ML algorithm of choice is then trained
1069 (here: random forests, RFs), and all features are scored for their importance by the default RF
1070 method. The maximal importance score among all shadow features is considered a threshold
1071 distinguishing truly informative features (features with scores above the threshold) and non-
1072 informative ones (scores below the threshold). This permutation and evaluation process is
1073 repeated multiple times and a statistical criterion is used for final definition of informative
1074 and uninformative features. Due to the conservative nature of the Boruta method, we relaxed
1075 the threshold used for defining informative features, as implemented in the python Boruta
1076 implementation (https://pypi.org/project/Boruta/). The third approach was an ensemble of the
1077 previous two approaches, taking the intersection of selected features.

1078 Lastly, the fourth approach was based on the Altmann method[50], computing *P* values per
1079 feature reflecting the significance of the feature's contribution to the model and taking

1080 features with a *P* value below a specified threshold. We additionally ran our pipeline with no
1081 feature selection at all, to serve as a baseline of performance.

1082 The entire evaluation of feature selection methods was incorporated in a repeated 5-fold
1083 cross-validation framework. Overall, different feature selection methods performed best for
1084 different datasets and different feature sets (see Supplementary Table S2), though differences
1085 were typically non-significant.

1086 Supplementary note 3: Single-view and early integration multi-view 1087 feature importance in predicting late-stage colorectal cancer

1088 Adding to the IBD example given in Figure 3, in the late-stage colorectal cancer (CRC) data,
1089 we again found well-established CRC-associated species included among the top important
1090 taxonomic features in both the taxonomy-only and multi-view models, including several sub-
1091 species of the oral commensal *Fusobacterium nucleatum*[113,114] (which were previously shown
1092 to colonize tumors[115]), as well as *Peptostreptococcus stomatis*[114], *Peptostreptococcus*
1093 *anaerobius*[116], and *Parvimonas micra*[114], (Supplementary Table S4). Of the measured
1094 metabolites, top contributing metabolites included fumarate (elevated in CRC), a known
1095 oncometabolite[117], and alpha-Aminoadipate, previously detected at higher levels in CRC
1096 tumor samples compared to adjacent mucosa[118]. While both of these metabolites are likely
1097 not microbiome-related, other top contributing metabolites such as dimethylglycine (elevated
1098 in CRC), previously suggested as an early marker for CRC[119], hydroxyproline (depleted in
1099 CRC), known for intestinal health benefits[120], and urocanate (elevated in CRC), again in line
1100 with previous reports[121], are potentially driven by microbiome activity[83,119]. Among
1101 functional microbiome features, though overall performance of these models was poor, genes
1102 involved in tryptophan biosynthesis (e.g. K01696, K00766; depleted in CRC patients) were
1103 identified as main contributors, in agreement with accumulating evidence showing the role of
1104 tryptophan metabolism in CRC development[122].

1105 Supplementary note 4: Explanation and formulation of the 1106 intermediate integration pipeline

1107 The intermediate integration approach implemented in this study is based on sparse canonical
1108 correlation analysis (sCCA) and specifically a recently introduced framework named

DIABLO[59]. 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.[57] and is closely related to the PLS method[123].

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)[124] 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:

$$\underset{a_h^{(1)},\ldots,a_h^{(Q)}}{\operatorname{argmax}} \sum_{\substack{i,j=1 \\ i \neq j}}^{Q} c_{ij} \cdot \left(\operatorname{cov}\left(X_h^{(i)} \cdot a_h^{(i)},\, X_h^{(j)} \cdot a_h^{(j)}\right)\right)$$

$$s.t. \left\| a_h^{(q)} \right\|_2 = 1 \;\; and \;\; \left\| a_h^{(q)} \right\|_1 \leq \lambda^{(q)} \;\; \forall q \in 1, \ldots, Q$$

The optimization problem is solved by an efficient iterative algorithm described in detail in Tenenhaus et al.[57]. 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[59].

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-discrimination 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

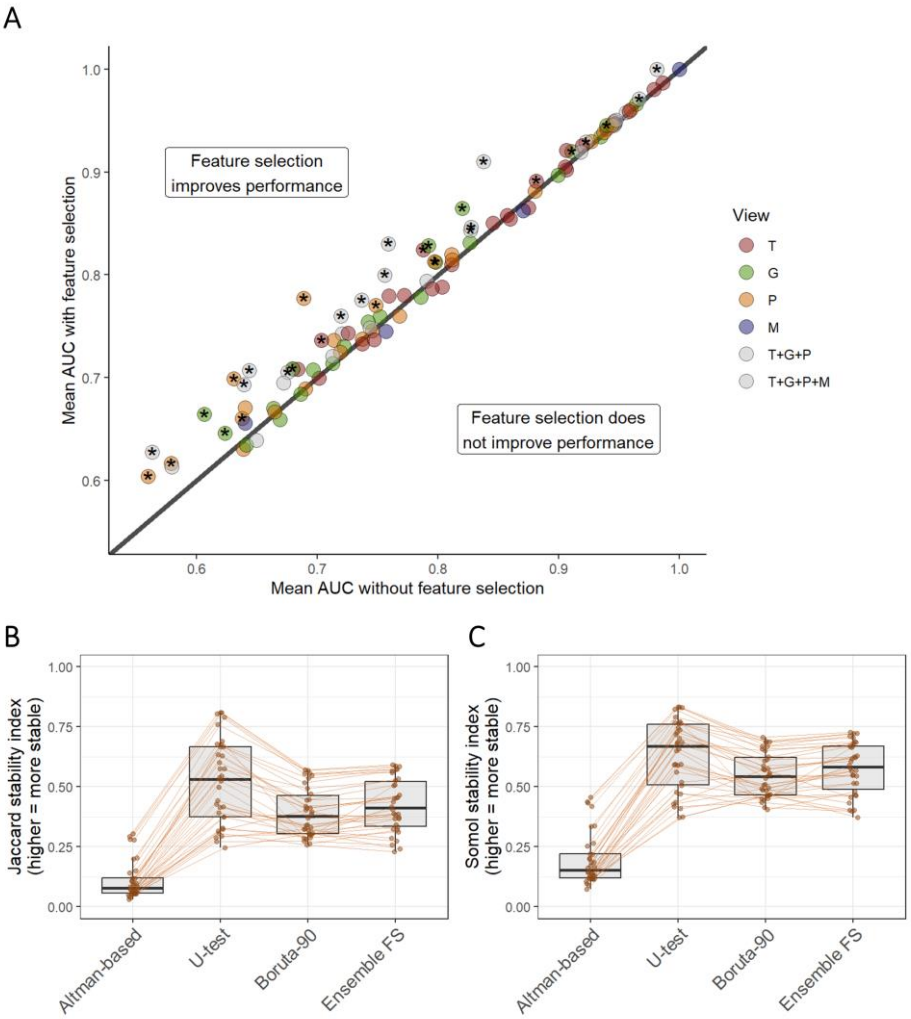

Figure S1: **Feature selection stability and performance improvement**. (A) Disease classifiers' performance (measured by a random forest AUC) with (Y axis) vs. without (X axis) applying feature selection before training. Each point represents a specific dataset (of the 26 analyzed datasets) and a specific 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.

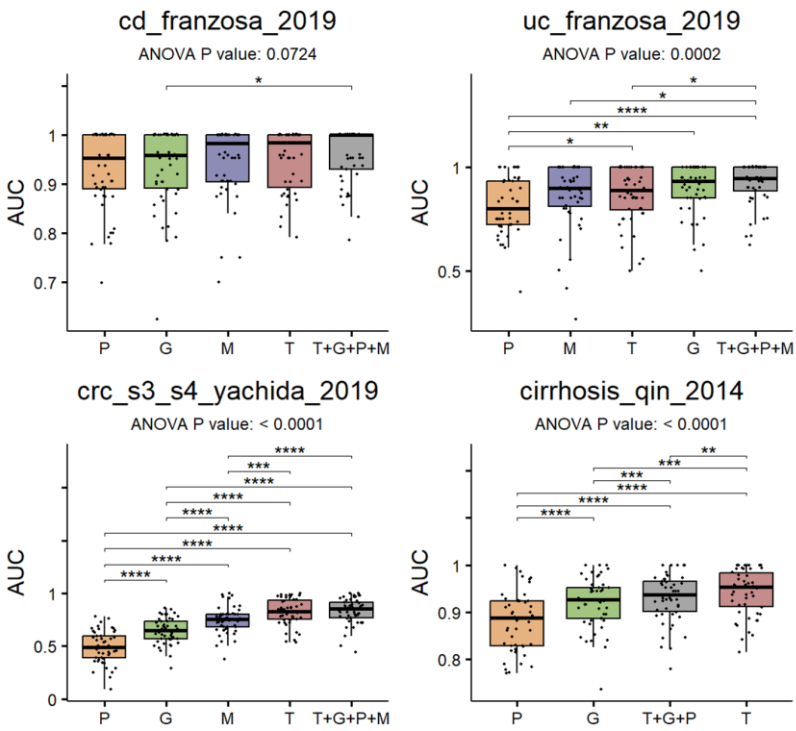

Figure S2: **Comparison of single-view and multi-view classifiers' performance**. For each of the 4 main datasets (listed at each panel's title) we show detailed performance distributions over repeated cross-validation iterations for both single- and multi-view RF models. Models are ordered within each panel from best average performance (right) to worst (left). Significant differences were determined by paired T-tests and a subsequent FDR correction. Significance marks: ****: $P < 0.0001$; ***: $P < 0.001$; **: $P < 0.01$; *: $P < 0.05$.

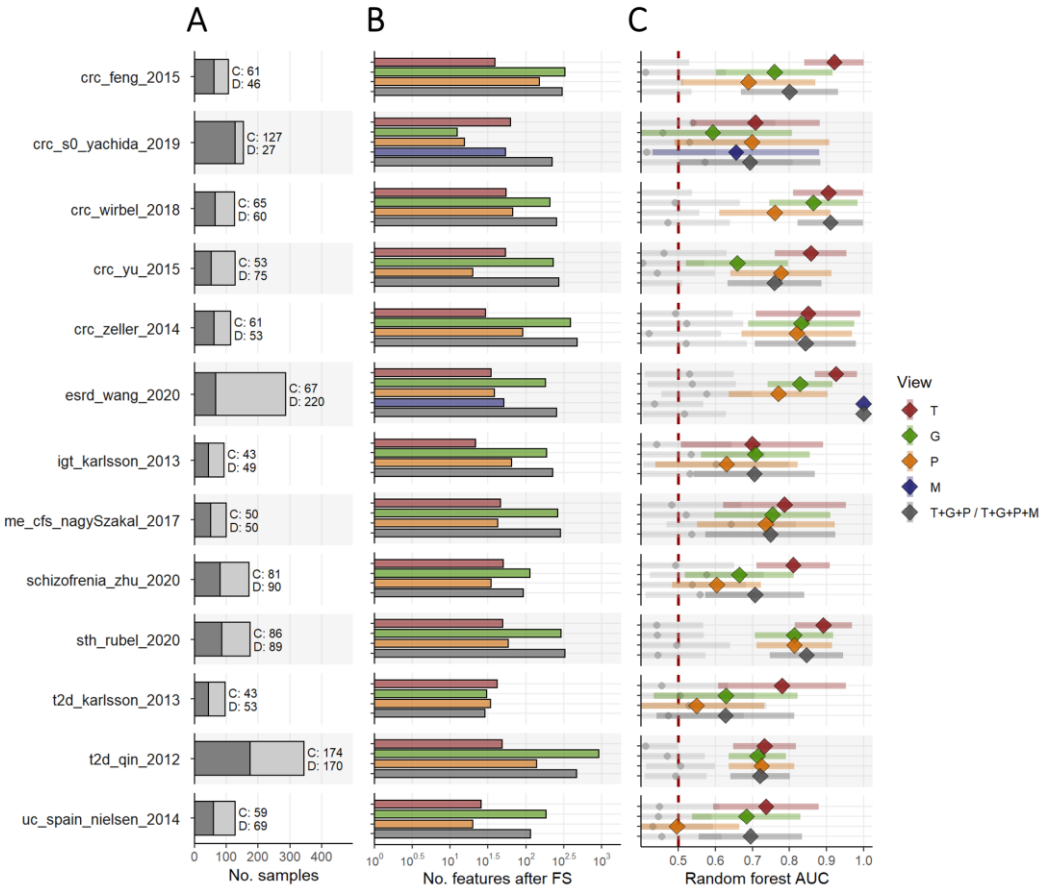

Figure S3: **Single- and multi-view disease classification based on shotgun sequencing data**. See Figure 2 legend. Datasets presented here were retrieved from the curatedMetagneomicData repository[93] or manually curated (See Supplementary Table S1).

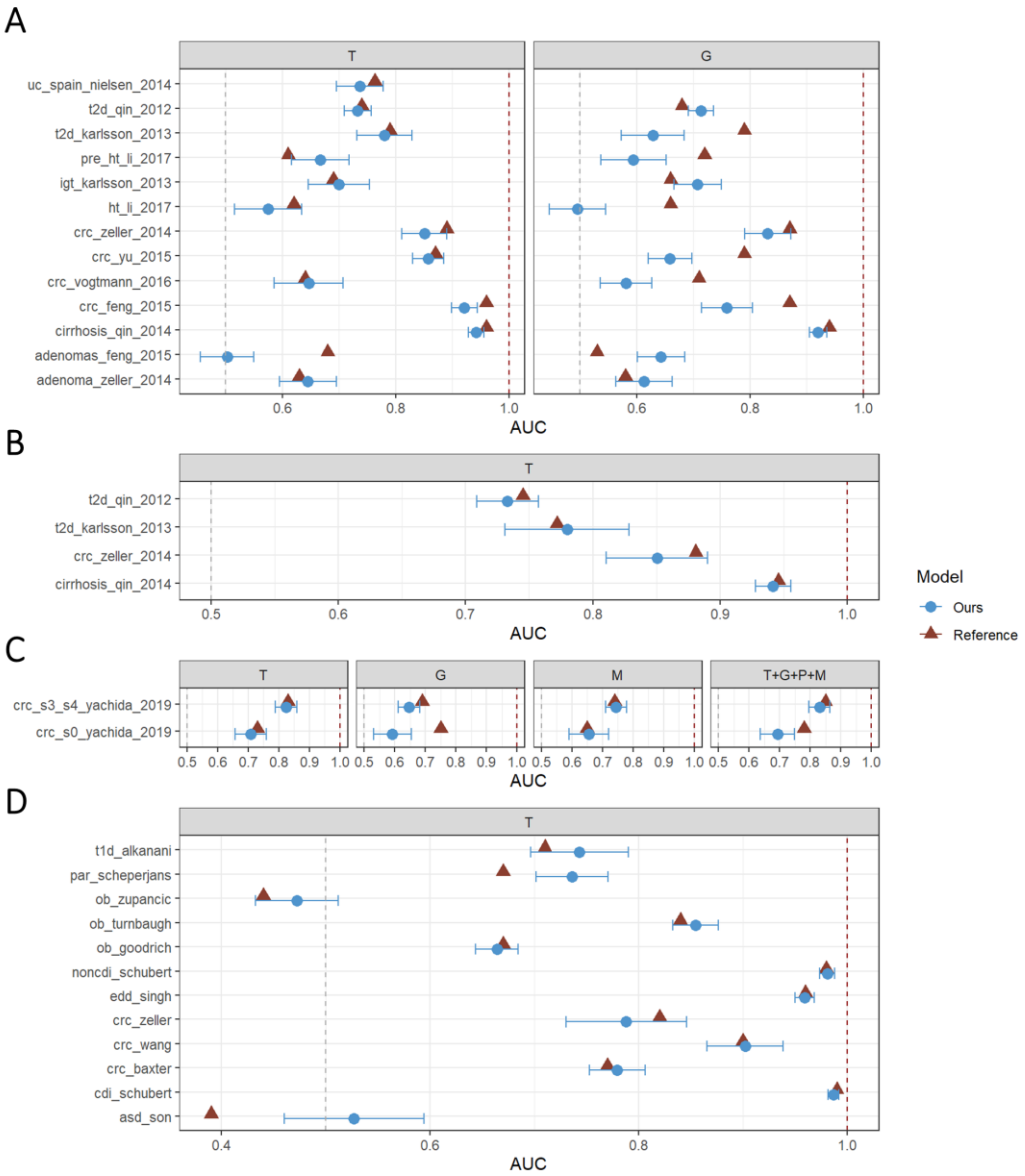

Figure S4: **AUC comparisons to previous machine learning studies using the same datasets**. We compared AUC results of our pipeline to those of similar pipelines reported in previous studies or meta-analyses, specifically Wirbel et al.[1] (A), Pasolli et al.[4] (B), Yachida et al.[26] (C) and Duvallet 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' R 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.

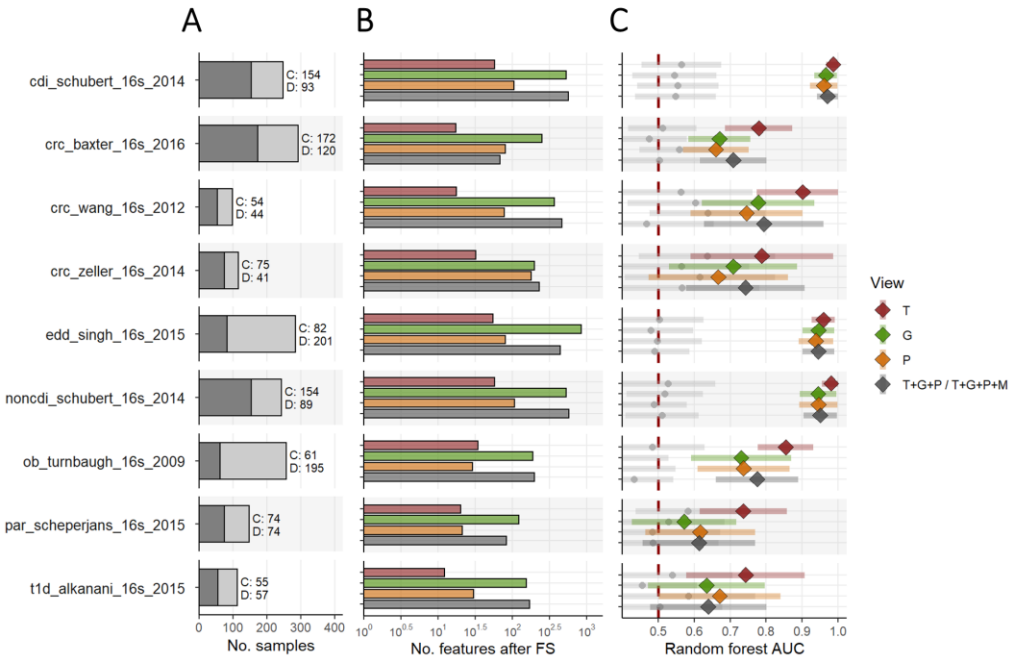

Figure S5: **Single- and multi-view disease classification based on 16S amplicon sequencing data from MicrobiomeHD resource**. See Figure 2 legend. Datasets presented here were retrieved from MicrobiomeHD[6].

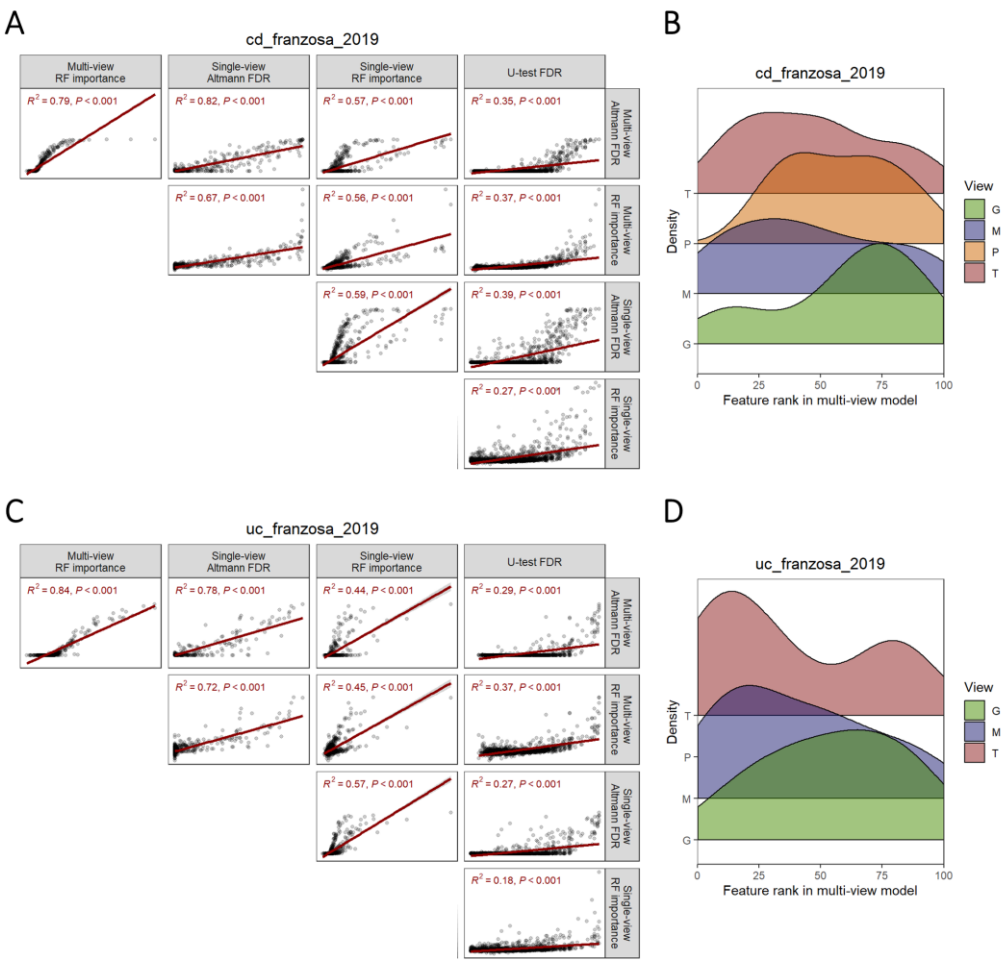

Figure S6: **Multiple estimations of feature importance**. Left panels present a pairwise comparison between different feature importance approaches, for the CD dataset (A) and the UC dataset (C). Each point in the scatter plot represents a feature, and overall correlation is illustrated by a fitted 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.

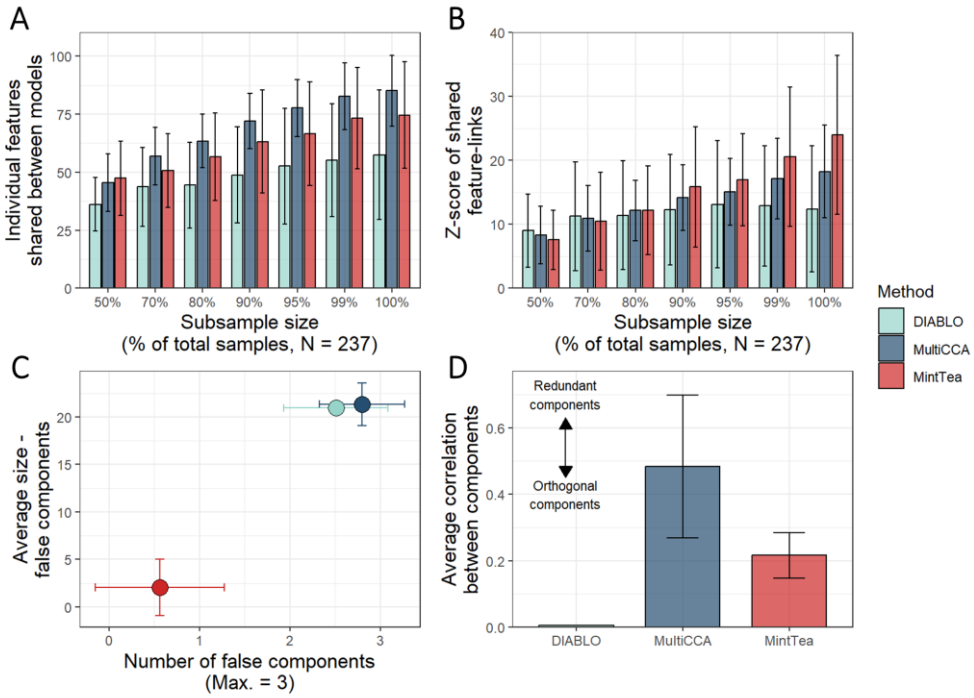

Figure S7: **MintTea comparison to DIABLO and MultiCCA**. Each panel represents a different comparison of MintTea to two other CCA-based methods: DIABLO[59] and sparse multiple CCA (implemented by the MultiCCA function from the 'PMA' R package[58]). All comparisons were made using the liver cirrhosis dataset. (A-B) Method stability assessment. For two random data subsets (ranging from 50% of the samples to 100%, noted on the x axis), components/modules were identified and compared. The more the two sets of components were similar, both in terms of individual features included in the components (A) and in terms of pairs of features that co-occurred in a component in both sets (B), we conclude that the method is more stable. See Supplementary Note 1 for details on how we computed these estimates. (C) False discovery rate analysis. All 3 methods were run on data that was shuffled across samples within each view. Higher number of identified components as well as higher numbers of features within components indicate higher false discovery rates. (D) Orthogonality assessment.

All results are averaged over 100 runs with random data subsets and error bars represent standard deviations. Penalty parameters were adjusted in each method so that the average component size (i.e. non-zero loadings in MultiCCA/DIABLO) is approximately 20 features (from all views combined).

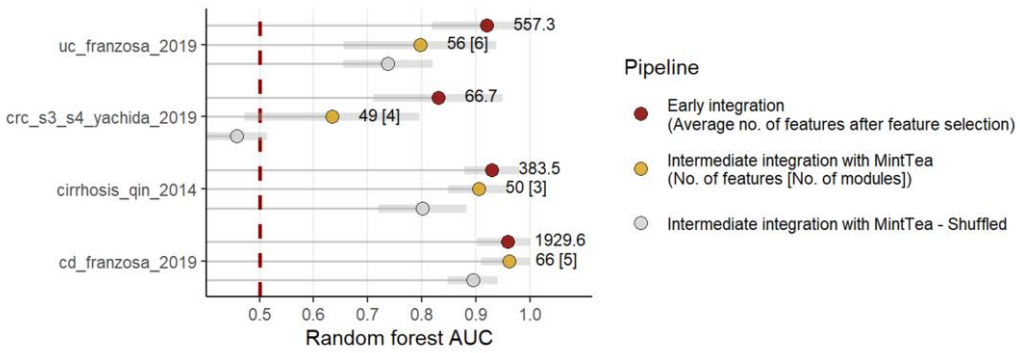

Figure S8: **Overall predictive power of multi-view modules**. For each of the four datasets, a random forest was trained on the first principal components of each module in that dataset. Cross-validated AUC's of these models are shown in yellow, and compared to standard random forest models on raw features (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.

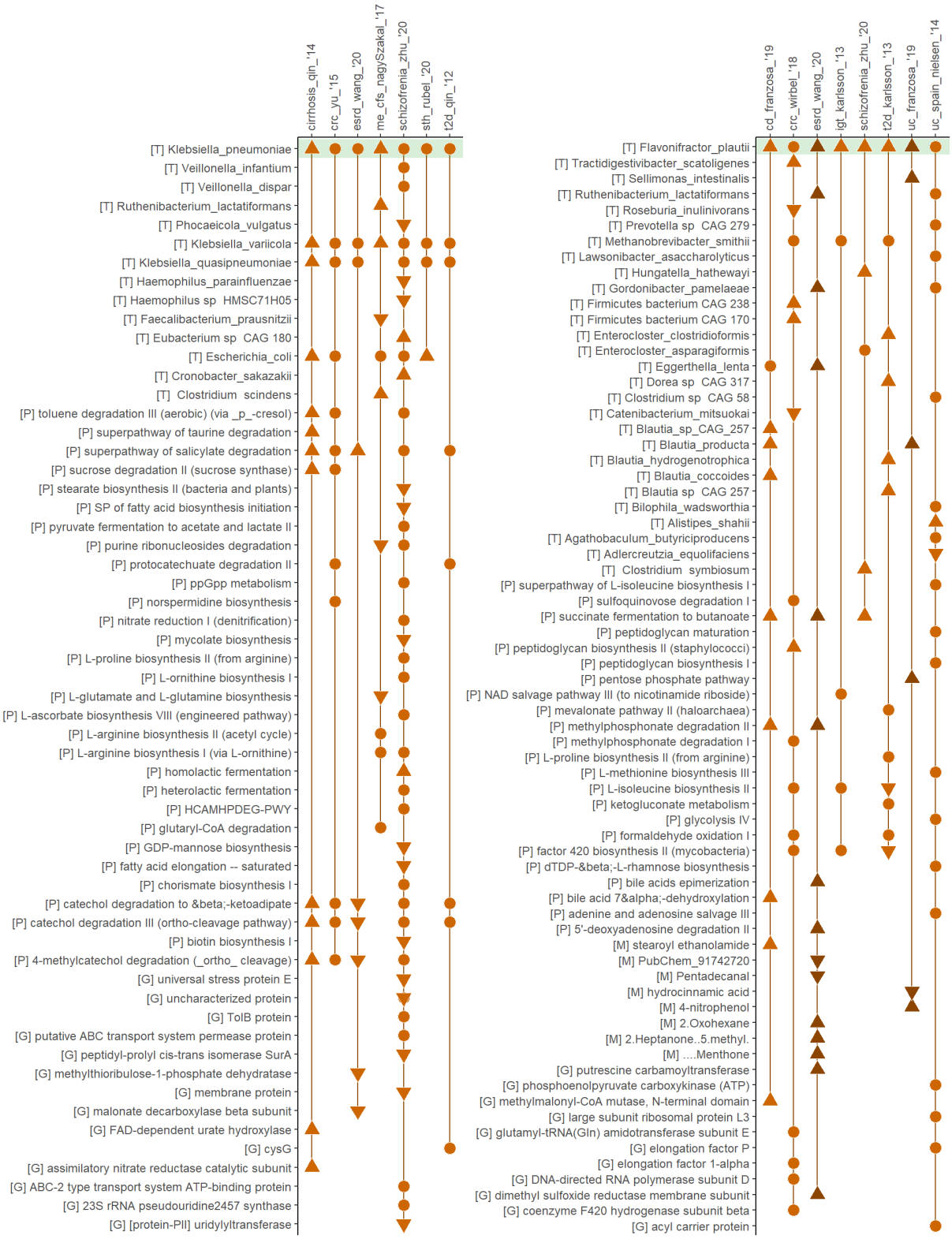

Figure S9: **Multi-view modules including the taxa *K. pneumoniae* (left) and *F. plautii* (right)**. See Figure 5E legend. Modules colored in dark orange were highly predictive of disease state (AUC > 0.7, and higher than AUC of random modules).

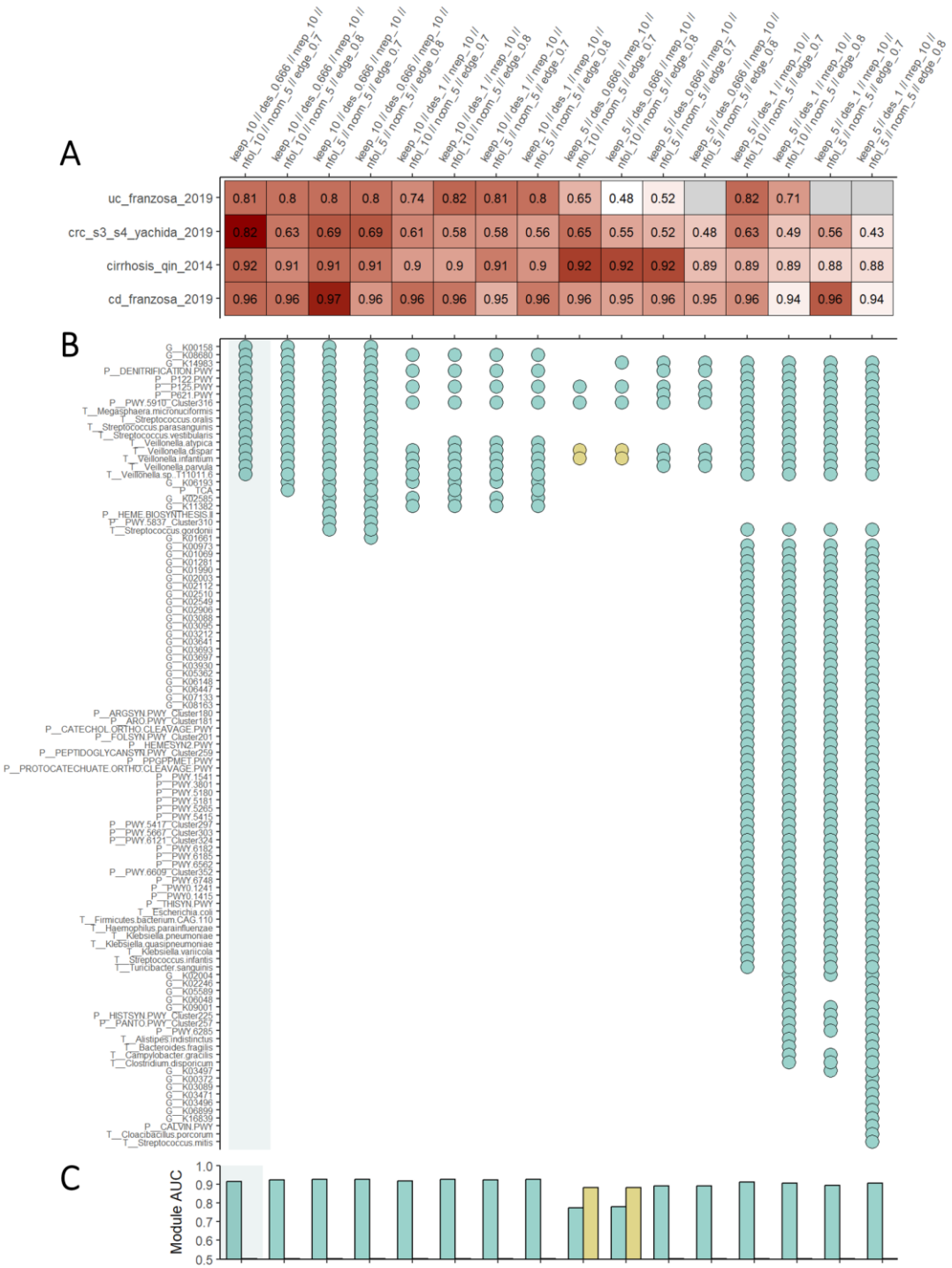

Figure S10: **Sensitivity analysis for MintTea's pipeline parameters**. The x axis for all panels are the various combinations of parameters required by our intermediate integration pipeline. Briefly, parameters include: 'keep' – the number of features to select from each view as provided to DIABLO's sGCCA computation; 'des' – the value to set as design matrix default (see Supplementary Note 4); '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.

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