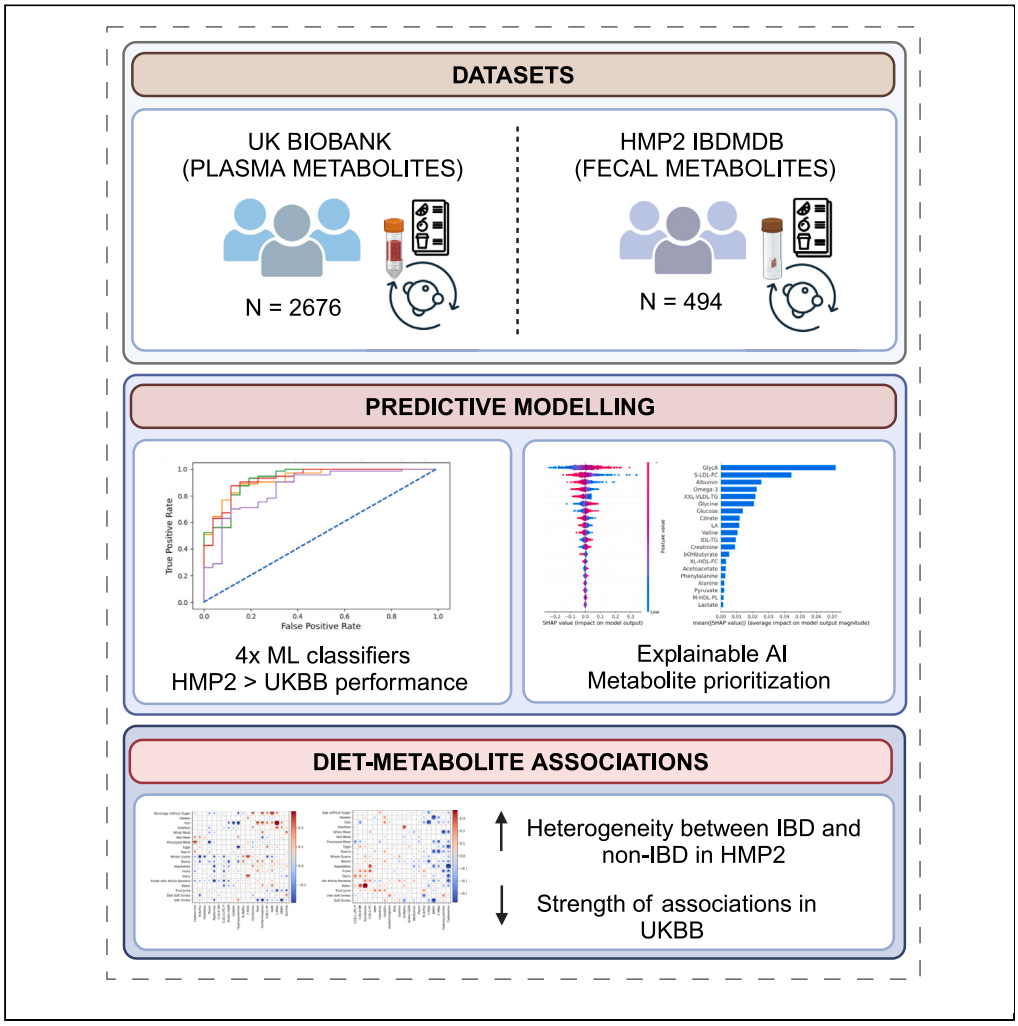
iScience

CellPress
OPEN ACCESS

Article

Explainable AI-prioritized plasma and fecal metabolites in inflammatory bowel disease and their dietary associations



Serena Onwuka, Laura Bravo-Merodio, Georgios V. Gkoutos, Animesh Acharjee

a.acharjee@bham.ac.uk

Highlights

Fecal metabolites predict IBD more effectively than plasma (AUC: 0.93 vs. 0.74)

Key features include plasma inflammatory markers, fecal nicotinic acid metabolites

Diet-metabolite associations stronger in feces, and differ between IBD and non-IBD

Onwuka et al., iScience 27, 110298
July 19, 2024
https://doi.org/10.1016/j.isci.2024.110298

iScience

CellPress
OPEN ACCESS

Article

Explainable AI-prioritized plasma and fecal metabolites in inflammatory bowel disease and their dietary associations

Serena Onwuka,[1] Laura Bravo-Merodio,[1,2,3] Georgios V. Gkoutos,[1,2,3] and Animesh Acharjee[1,2,3,4,*]

SUMMARY

Fecal metabolites effectively discriminate inflammatory bowel disease (IBD) and show differential associations with diet. Metabolomics and AI-based models, including explainable AI (XAI), play crucial roles in understanding IBD. Using datasets from the UK Biobank and the Human Microbiome Project Phase II IBD Multi'omics Database (HMP2 IBDMDB), this study uses multiple machine learning (ML) classifiers and Shapley additive explanations (SHAP)-based XAI to prioritize plasma and fecal metabolites and analyze their diet correlations. Key findings include the identification of discriminative metabolites like glycoprotein acetyl and albumin in plasma, as well as nicotinic acid metabolites andurobilin in feces. Fecal metabolites provided a more robust disease predictor model (AUC [95%]: 0.93 [0.87–0.99]) compared to plasma metabolites (AUC [95%]: 0.74 [0.69–0.79]), with stronger and more group-differential diet-metabolite associations in feces. The study validates known metabolite associations and highlights the impact of IBD on the interplay between gut microbial metabolites and diet.

INTRODUCTION

Inflammatory bowel disease (IBD), which primarily includes ulcerative colitis (UC) and Crohn's disease (CD), is characterized by chronic gastrointestinal conditions that collectively affect approximately 5 million individuals as of 2019.[1] Despite the rising global prevalence rates of IBD,[2] its etiology remains elusive, with the main regulator of IBD pathogenesis believed to be the adaptive immune system as the main mediator of gut inflammation.[3–6] This immune response is inherently linked to the genetic makeup of an individual; however, studies show varying proportions of heritability.[7–10] Studies have also strongly linked IBD development with factors such as the gut microbiome,[11] the use of antibiotics,[12] and diet.[12] Unraveling such biological complexity necessitates targeted-omics studies, with metabolomics recently helping to identify distinct disease-related patterns[13–17] and key differences between individuals with IBD and those without.[18–20] These differences have been observed as alterations in fecal short-chain fatty acids[21] and serological lipids, such as cholesterol levels and its lipoprotein levels,[22,23] and as changes in amino acid profiles, generally increased in feces[21,24] and decreased in serum or plasma,[25,26] as well as energy-related metabolites.[27,28]

As metabolomics is used to unravel these intricacies of IBD, it is increasingly being assessed in clinical practice, with AI-based modeling of metabolomics data integration helping identify potential metabolic markers that could be leveraged for therapeutic purposes. Promising studies have shown the power of machine learning (ML) in predicting IBD diagnosis, remission responses, and surgery risk,[29–32] but in order for these approaches to be properly translated into clinical practice, issues regarding model interpretability and explainability need to be tackled. The better the performance of a model, the greater tendency for it to be increasingly complex, like in the case of ensembles and deep learning models, and not intrinsically interpretable, as in decision trees models. Therefore, balancing model performance with complexity is essential,[33,34] with post-hoc explainability techniques,[35] also known as explainable AI (XAI),[36] emerging as key resources. The application of XAI in biomedical research particularly saw a surge in 2020, correlating to the rise of COVID-19 globally.[35] However, the utilization of XAI in ML-based studies on IBD pathogenesis is an area that has received limited exploration and investigation.

Further, AI-identified important metabolites in IBD are also likely intertwined with diet, as the nutrients consumed by an individual play a crucial role in modulating numerous metabolic processes within the organism,[37] and numerous studies have explored the relationship between diet and IBD. Diets rich in fiber have been linked to a lower risk of either the development of IBD or recurrence of symptoms after remission,[38–40] while western-style diets characterized by high consumption of refined carbohydrates, red meat, high-fat foods, and ultra-processed foods have been found to elevate the risk of IBD.[41–43] Additionally, dietary interventions for remission, including enteral nutrition for induction[44–46] and specific carbohydrate diets[47–50] for maintenance, have been explored as strategies for managing IBD. However, despite

[1]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK
[2]Centre for Health Data Research, University of Birmingham, Birmingham, UK
[3]Institute of Translational Medicine, University of Birmingham, Birmingham, UK
[4]Lead contact
*Correspondence: a.acharjee@bham.ac.uk
https://doi.org/10.1016/j.isci.2024.110298

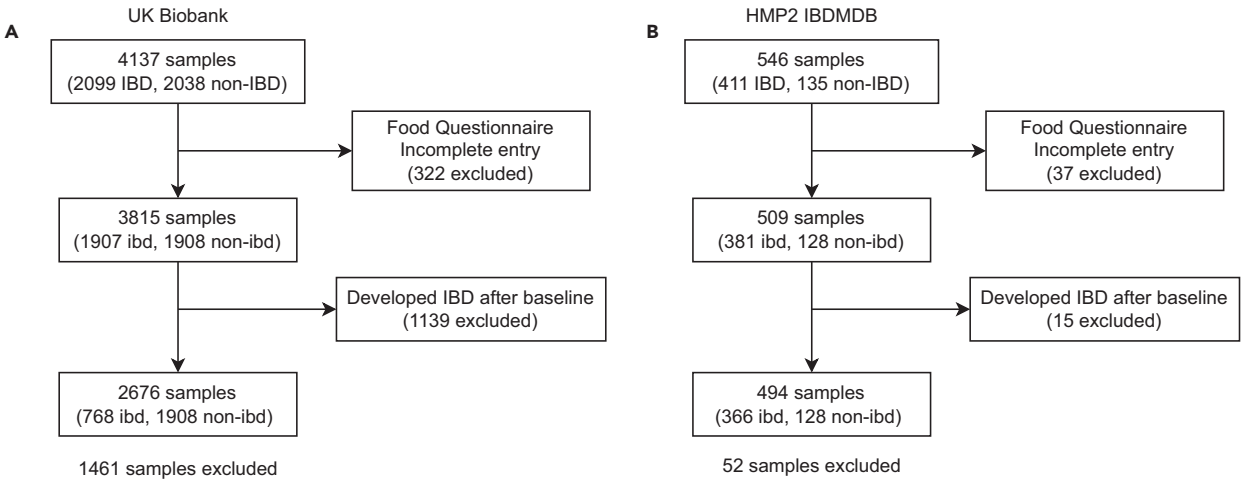


Figure 1. Sample exclusion flowchart

The filtering process of samples of (A) the UKBB cohort and (B) the HMP2 cohort, based on their respective exclusion criteria.

this wealth of research on diet associations with IBD, there exists a notable gap in understanding the interactions of diet directly with metabolic processes in IBD.

This study aims to fill these gaps by leveraging two publicly available datasets: the UK Biobank (UKBB)[51] and the Human Microbiome Project Phase II IBD Multi'omics Database (HMP2)[52] for metabolomics-based IBD prediction, applying XAI, and exploring the relationships between key metabolic profiles and dietary intake in IBD and non-IBD individuals. Results will potentially reveal important biomarkers for IBD, improve understanding of complex model predictions in IBD, and offer a nuanced understanding of the intricate relationship between metabolism, diet, and IBD pathogenesis. Ultimately, these findings may pave the way for the development of more targeted and effective interventions for individuals affected by IBD.

RESULTS

Baseline data characteristics

A total of 1,461 and 52 samples were excluded for the UKBB and HMP2 cohort, respectively (Figure 1), with an average of 5.1 (SD = 1.4) samples per individual in the latter cohort. Baseline demographic characteristics of the resulting 2,676 samples of the UKBB, and 494 samples of the HMP2, are shown in Table 1. About one-third of the UKBB samples (768 samples) belonged to the IBD class, while an equivalent proportion of the HMP2 cohort consisted of non-IBD samples (128 samples). The median ages were 59 and 21 for the UKBB and HMP2 cohorts, respectively. They both had a gender distribution split of approximately 50% (52.1% females in the UKBB plasma dataset; 49.0% females in the HMP2 feces dataset), with over 90% belonging to the white race in both cohorts, and within each class. In the HMP2 cohort, more than half of the IBD samples consisted primarily of individuals who had received a diagnosis within the past year. Conversely, an equivalent proportion of the UKBB IBD samples consisted of individuals who had been diagnosed at least 6.7 years ago.

Model performance

The UKBB metabolome data of 37 features were used to train the four classifiers used in this study: extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), random forest (RF), and least absolute shrinkage and selection operator (LASSO) regularization. The plasma metabolites of the UKBB cohort did not generally yield good model performances. LASSO emerged the top among the four classifiers, achieving an area under the curve (AUC) test score of 0.74 (95% CI: 0.69–0.79), while the other ML methods scored about 0.67 (Figure 2). With an optimal grid alpha value of 0.001, LASSO performed excellently in identifying individuals without IBD, having a specificity of 0.97 (Table 2). However, it was not effective in capturing all true IBD cases, as only about 14% of the IBD samples were correctly predicted (recall = 0.143). However, when it did predict IBD, there was an approximate 69% likelihood that it was truly IBD (precision = 0.688).

Among the HMP2 data, all the classifiers trained on the 160 fecal metabolites performed well. However, LASSO notably performed the least (AUC = 0.86), with all ensemble methods achieving about the same AUC score of 0.93 (Figure 2). Considering the performance of the ensembles across the other metrics, LGBM emerged as the preferred classifier, yielding an AUC test score of 0.93 (95% CI: 0.87–0.99) with the following optimal hyperparameters: 'colsample_bytree' = 0.8, 'learning_rate' = 0.1, 'max_depth' = 20, 'num_leaves' = 31, and 'subsample' = 0.8. This indicates that the model was effective at distinguishing between IBD and non-IBD cases, with an F1 score of 0.94 reflecting a balance between precision and recall (Table 2). Out of all positive predictions, up to 89% were truly IBD, with the model exceptional in capturing true IBD cases with a 99% recall rate. Nevertheless, the model's specificity score of 0.65 suggests that its ability to accurately identify non-IBD cases was limited. Essentially, the model had more errors in correctly identifying non-IBD cases than in missing IBD cases.

Table 1. Baseline demographic characteristics

	UKBB[a] Cohort Characteristics				HMP2[b] Cohort Characteristics			
	Total	Non-IBD	IBD	*p* value	Total	Non-IBD	IBD	*p* value
Participants	2676	1908	768		494	128	366	
Median age (years)	59.0 (52.0–64.0)	59.0 (52.0–64.0)	60.0 (51.0–64.0)	0.42	23.0 (14.0–43.0)	23.0 (13.0–55.0)	22.0 (15.0–41.0)	0.83
Female participants (%)	1394 (52.1)	985 (51.6)	409 (53.3)	0.47	242 (49.0)	56 (43.8)	186 (50.8)	0.20
White[c] race (%)	2569 (96.0)	1827 (95.8)	742 (96.6)	0.36	445 (90.1)	123 (96.1)	322 (88.0)	0.01
Median years since diagnosis (IQR, years)			6.7 (3.0–11.3)				0.0 (0.0–12.0)	

[a]UKBB: UK Biobank.
[b]HMP2: Human Microbiome Project Phase II.
[c]For UKBB, this includes "British," "any other white ethnicity," and "Irish;" For the HMP2, this includes only "White."

Feature selection and interpretation

Following model predictions, Shapley additive explanations (SHAP) was applied to interpret such predictions, focusing on the contributions of the top 20 features at local and global levels. In the UKBB model, the top 20 out of the 37 metabolites contributed 99.6% of the cumulative absolute mean SHAP values, resulting in a top-to-bottom ratio of 44.51. SHAP local impact and global importance plots were generated for these features (Figure 3). Overall, these top metabolites included markers of inflammation (GlycA), lipoprotein subclasses (S-LDL-FC, XL-HDL-FC, and XXL-VLDL-TG), fatty acids (omega-3 and LA), amino acids (glycine and valine), energy metabolism (glucose and acetone), and waste product creatinine. Notably, GlycA emerged as the most discriminative metabolite with a substantial lead over the next two closely ranked features, S-LDL-FC and albumin, as observed in the global importance plot (Figure 3B). These results are complimented by the LASSO analysis in R that show GlycA, albumin, S-LDL, as well as omega-3 and glycine, to be the top features of the 400 bootstrapped models (Figure S1). Analyzing the local model impact plot (Figure 3A), higher values of GlycA, glycine, glucose, and creatinine were observed to drive the model toward prediction of the positive class, IBD. Contrarily, IBD prediction was driven by lower values of S-LDL-FC, omega-3, and albumin.

In the HMP2 model, the top 20 metabolites out of the total 160 contributed 60% of the cumulative absolute mean SHAP values, with a top-to-bottom ratio of 81.09. Among many other classes, these metabolites included vitamin B3 compounds (nicotinuric acid [NUA], N1-methyl-2-pyridone-5-carboxamide [NMPC], 1-methylnicotinamide [1-MNA], and nicotinamide [NAM]), and lipids (a lysophosphatidylcholine [C18:1 LPC-P], a sphingomyelin [C16:0 SM], and a phosphatidylcholine [C36:2 PC]). Remarkably, NUA emerged as the most discriminatory by an extreme margin (Figure 3D), with elevated levels driving IBD prediction (Figure 3C). Interestingly, no non-IBD sample contributed to IBD prediction, while a mix of non-IBD and IBD samples having low NUA values contributed to non-IBD prediction (Figure S2). Similar to NUA, most other top metabolites like pyridoxine, and n-acetylputrescine (N-AcPut) had higher values associated with IBD. In contrast, elevated levels of features like urobilin and hydroxycotinine drove prediction of the negative class, non-IBD. Complementary to these results, six of the top 20 SHAP-ranked metabolites, NUA, pyridoxine, N-AcPut, urobilin, C16:0 SM, and hydroxycotinine, were among the features of the 400 bootstrapped LASSO models in R that appeared the most, based on a threshold (Figure S1).

Diet-metabolite correlations

After the top SHAP-ranked features were correlated with their corresponding dietary components for each case-control group, the diet-metabolite correlations were found to be similar across both groups in the UKBB cohort. As depicted in Figures 4A and 4B, individuals with IBD had 36 correlations, while those without had 32, all being significant. Moreover, 90% of the correlations fell within the range of (−0.091, 0.109) for the IBD group, and (−0.094, 0.094) for non-IBD group. Dietary wise, omega-3 fatty acid had the strongest correlation across both groups, displaying significantly moderate correlation levels with oily fish, with levels up to three times higher than average in each group (in IBD: r = 0.365, and in non-IBD: r = 0.367; false discovery rates [FDR] < 0.001; refer to Table S4 for all correlation and FDR values). Similarly, across both groups, glycine was significantly negatively correlated with red meat intake (processed, pork, lamb, and beef), although weak. However, there was a considerable difference; certain food-metabolite correlation patterns observed were stronger among those without disease. Particularly, among non-IBD individuals, omega-3 and XL-HDL-FC exhibited a pattern opposite to that of creatinine and XXL-VLDL-TG across processed meat, bread, fruit and vegetable intake. For instance, XL-HDL-FC in the non-IBD group was negatively correlated with bread (r = −0.156, FDR < 0.001) while creatinine showed a positive correlation (r = 0.182, FDR < 0.001). However, among individuals with IBD, these correlations were notably diminished, with only creatinine showing significant associations across the aforementioned food groups.

In contrast to the minor variations observed across both IBD and non-IBD groups in plasma metabolite and diet interactions, the differences in fecal-diet interactions in the HMP2 cohort were much more pronounced. In this cohort, the non-IBD group had more correlations (R = 159) compared to the IBD group (R = 111). Individuals without IBD exhibited correlations stronger than those with IBD (Figures 4C and 4D), although only 22 out of the 156 correlations were significant, while in just over half of the 111 IBD correlations, 65 were significant. However, this is likely due to IBD samples being about three times the amount of non-IBD samples. Nonetheless, most correlations within this cohort

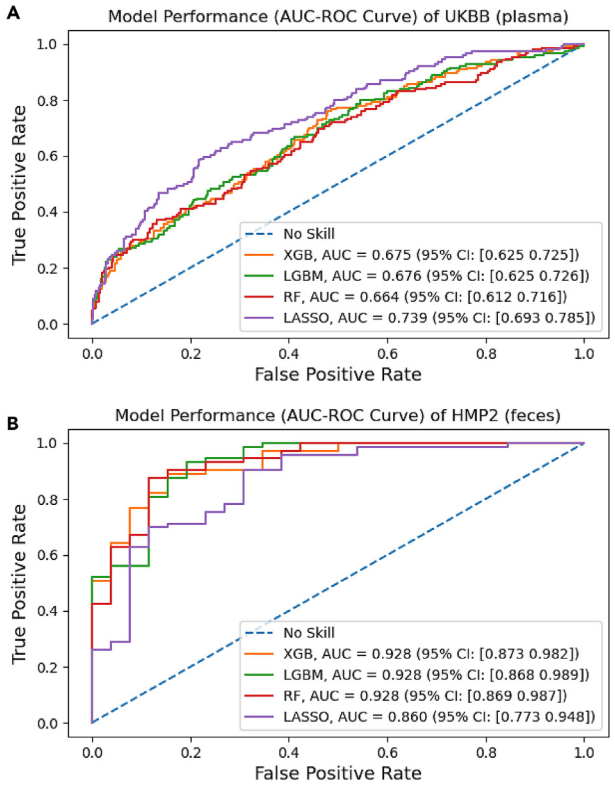


Figure 2. Model performance comparison

The AUC_{test}-ROC curves of the optimized classifiers for (A) the UKBB data and (B) the HMP2 data, with the blue dashed line representing a model performance that had no skill at all for comparison, are illustrated here (ROC, receiver operating characteristic; XGB, extreme gradient boosting; LGBM, light gradient boosting machine; RF, random forest; LASSO, least absolute shrinkage and selection operator). Data are represented as "mean (confidence interval)." See also Figure S1.

were stronger than the correlations observed in the UKBB cohort, with 90% of the HMP2 correlations falling within the range of (−0.139, 0.149) for the IBD group, and (−0.224, 0.191) for the non-IBD group. Another notable difference was metabolites expressing opposing or non-existent associations for the same food group between IBD and non-IBD. For instance, while fish was significantly positively correlated with 1-MNA in the IBD group ($r = -0.285$, FDR < 0.001), it showed non-significant negative association in the non-IBD group ($r = 0.192$; refer to Table S5 for all correlation and FDR values). Further, water with pyridoxine ($r = 0.371$, FDR < 0.001) and cadaverine with vegetables ($r = -0.371$, FDR < 0.001), which tied for strongest correlation among the non-IBD samples, exhibited either weak or invalid correlation values among the IBD samples (both $r = 0.065$). Beyond these overview distinctions, there were a few diet-metabolite associations that remained similar across both groups, albeit with varying strengths. For instance, NMPC was significantly inversely correlated with soft drink consumption in both groups (IBD: $r = -0.183$ and non-IBD: $r = -0.323$, 0.001 < FDR ≤ 0.01), and hydroxycotinine demonstrated negative correlations with beans intake (IBD: $r = -0.136$, 0.01 ≤ FDR < 0.05 and non-IBD: $r = -0.188$). Further, some metabolites clustered in both groups, like lipid-related metabolites, C18:1 LPC-P and C16:1 SM and some vitamin B3 metabolites, NMPC and 1-MNA, indicating that they undergo similar metabolic patterns in association with diet, regardless of how different the associations may be with disease or without.

DISCUSSION

Model performance

In the UKBB data, LASSO outperformed the other ML methods, which were all ensembles. This is likely due to a number of possibly redundant features included in the ensemble models, which were assigned zero weights in the LASSO model, which automatically filters out highly correlated variables. On the other hand, all the ensemble methods each outperformed LASSO in HMP2, highlighting the possible non-linear nature of the associations between outcome and features.

Discriminatory plasma metabolites

SHAP XAI successfully explained the prediction of both cohort models and revealed the highly influential features of these models. The top features of the UKBB model included GlycA, albumin, sub-classes of high and low-density lipoprotein cholesterol forms, and omega-3 fatty

Table 2. Performance metric scores of the UKBB and the HMP2

Dataset (matrix)	Classifier	Test AUC (%)	Specificity (%)	Recall (%)	Precision (%)	F1 (%)
UKBB (plasma)	XGB	0.675	0.969	0.175	0.692	0.280
	LGBM	0.676	0.963	0.227	0.714	0.345
	RF	0.664	0.974	0.188	0.744	0.301
	LASSO	0.739	0.974	0.143	0.688	0.237
HMP2 (feces)	XGB	0.928	0.654	0.959	0.886	0.921
	LGBM	0.928	0.654	0.986	0.889	0.935
	RF	0.928	0.615	0.973	0.877	0.922
	LASSO	0.858	0.692	0.890	0.890	0.890

acid. However, the model's average AUC score (0.74) implies that these features are not very effective for classifying IBD within this cohort, which could indicate an insufficiency of these NMR-identified plasma metabolites to distinguish properly between IBD and healthy individuals, particularly in older adults. However, as the results are not as low as chance (0.50), it is reasonable to suggest that these highlighted key metabolites are involved in IBD pathogenesis, with findings validated by the existing literature in the following text.

For instance, a prior study also identified significantly elevated levels of GlycA, a well-established stable inflammatory marker,[53–55] in individuals with active IBD compared to controls.[56] Similarly, decreased levels of albumin, which has been stated to be an inverse biomarker for systemic inflammation,[57] align with the UKBB model's prediction of IBD. However, it is important to note that these two inflammatory markers are inclined to predict as IBD, only individuals with active IBD or individuals without IBD but are undergoing systemic inflammation at the time of testing.

Similar misclassifications may have occurred with other influential metabolites, such as LDL-C (S-LDL-FC), also known as "bad" cholesterol, where varying study findings suggest mixed LDL-C profiles in individuals with IBD, with some showing low levels,[58,59] and others indicating high levels.[60,61] However, the lower levels of HDL-C (XL-HDL-FC), or "good" cholesterol, observed in IBD predictions is common.[58–61] Although for both lipoprotein cholesterols, the changes between the healthy and diseased were noticed more prominently with CD than with UC.[58–61] Nonetheless, lipoprotein cholesterol levels being associated with IBD, and cholesterols having been associated with cardiovascular risk,[62–64] add to the increasing number of findings linking IBD to cardiovascular disease risk.[65–67] This suggests some shared lipid metabolisms underlying both disease types.

Further, in line with previous studies,[68,69] lower levels of omega-3 fatty acid were found to be associated with IBD. However, these results are inconclusive concerning the role of omega-3 in IBD individuals. While it is known that omega-3 boosts one's immunity,[70] it has been potentially linked to a decreased risk of developing IBD[71,72] and among IBD patients, has been linked to reduced intestinal inflammation[71]; it has also been speculated that the ratio of omega-3 to omega-6 is more crucial than the level of omega-3 itself.[68] Further, it is unclear if depleted levels of omega-3 are causal, elevated levels provide protection, or supplementation is truly helpful for all individuals with IBD or only a subset of them.[71,73]

Discriminatory fecal metabolites

While the plasma metabolites were not as effective in distinguishing between both healthy and diseased, the fecal metabolites in the HMP2 cohort showed much greater capabilities. Although achieving an AUC score of 0.93, higher than commonly published metabolomics-based predictions,[74–76] may be inflated due to there being much more IBD than non-IBD samples, the high performance still underscores the potential importance of fecal metabolites in the pathogenesis or diagnostic assessment of IBD. This high discriminatory score prompts further investigation into the specific metabolites and underlying biological mechanisms driving this predictive power. Insights from the SHAP algorithm unveiled several metabolites that exhibited higher discriminatory abilities where variations in their levels—both high and low—were generally associated with class predictions. These key metabolites included metabolites of vitamin B3, phospholipid metabolites, and urobilin.

Vitamin B3, mainly represented as NA, with NAM as its main metabolite, is present in the body mainly through diet, and to a lesser extent, synthesized *de novo*. Fecal levels of NA have been found to be reduced in both CD and UC patients, compared to healthy controls, although more pronounced in CD patients.[77] However, diminished levels of its metabolites (NUA, NMPC, 1-MNA, and NAM) were notably influential in predicting non-IBD cases. While low values contributed to IBD prediction, elevated levels were also involved, but increasingly so in the case of NUA.

The elevated NUA levels that strongly drove IBD prediction has been previously discovered to be due to a confounding drug effect.[78] Only individuals with IBD displayed increased NUA levels (Figure S2). Further literature research then revealed that the NUA levels in the HMP2 cohort were confounded by intake of the 5-aminosalicylic acid (5-ASA) drug, the common first line of treatment for IBD patients,[79,80] as only individuals that took it displayed increased NUA levels.[78] Knowing this, the classifiers were re-run on the HMP2 data without considering NUA, and the ensemble models actually performed better on average, with the top classifier achieving an AUC score of 0.95 (Figure S5). This suggests that the potential bias introduced by NUA led to a slight decrease in model performance. However, the analysis without NUA was

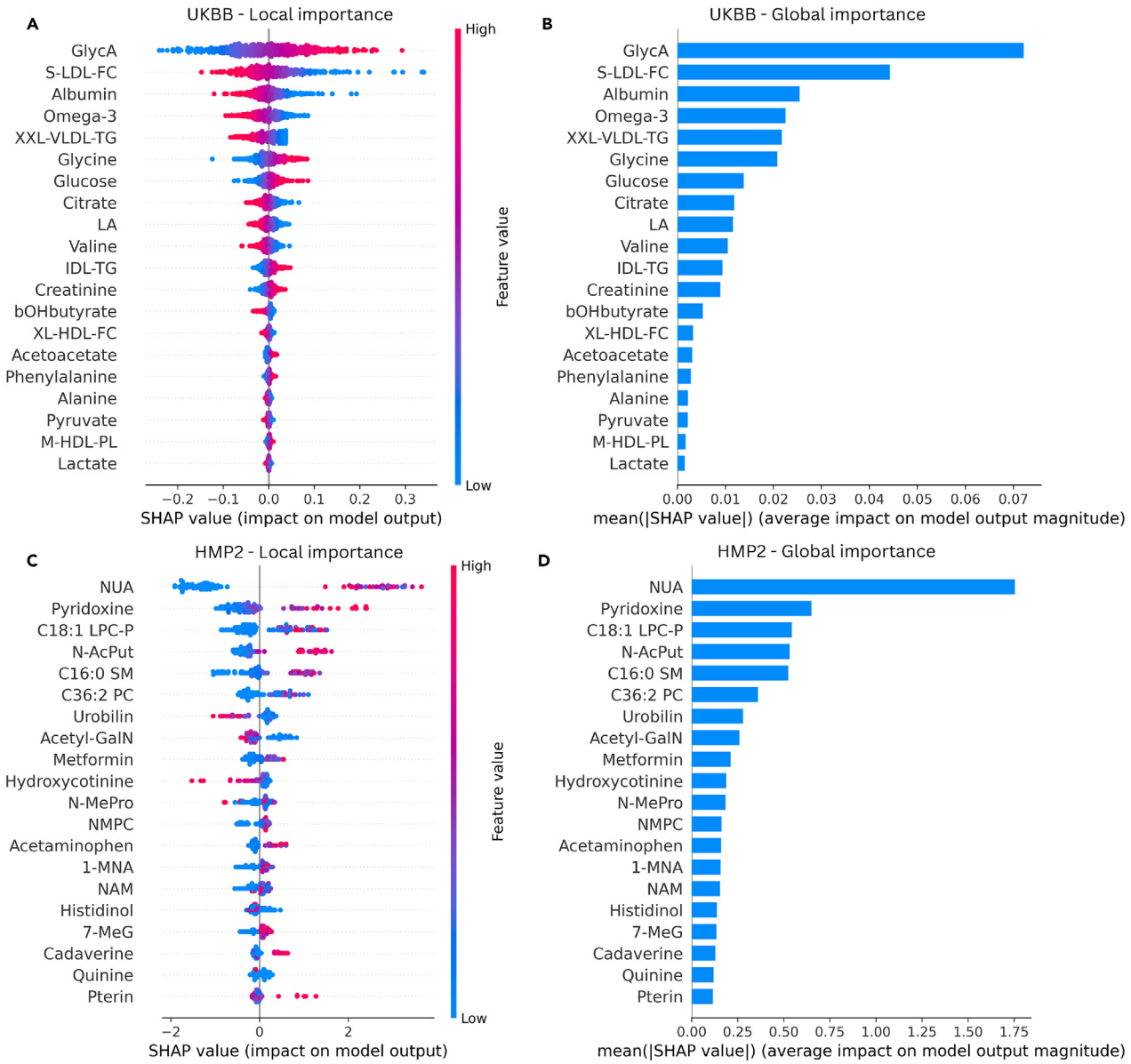


Figure 3. SHAP-based feature selection and model interpretation

This figure showcases ranked SHAP summary plots of model predictions. (A) and (B) depicts LASSO-based plots of the UKBB cohort, and (C) and (D) depicts LGBM-based plots of the HMP2 cohort. The features shown are the top 20 metabolites of each model based on SHAP values. In the local importance summary plot for SHAP values (left), the samples are represented as the colored dots, with the color determined by the value it has for that feature. A positive SHAP value corresponds to a positive impact on the model, driving the algorithm toward prediction of the positive class, and vice versa. In the global importance summary plot for mean absolute SHAP values (right), features higher in rank correspond to a greater number of samples with SHAP values significantly deviating from zero, either positively or negatively (LASSO, least absolute shrinkage and selection operator; LGBM, light gradient boosting machine). Note: (A) and (B) illustrate summary plots for prediction non-specific to class, as linear models typically output a single set of SHAP values, while (C) and (D) represent a tree model, which produce class-specific SHAP values and thus, are specific to the positive class, IBD. See also Figures S1 and S2.

kept supplementary as model performance was not significantly impacted. Further, in this way, post-hoc measures are minimized, prioritizing realism and generalizability.

Nonetheless, higher levels of fecal NMPC and 1-MNA being associated with IBD prediction could be due to increased levels of NAM,[81,82] a degradation product of nicotinamide adenine dinucleotide (NAD).[83,84] NA and NAM metabolism, which involves NAD turnover that is increased in IBD,[85] is currently a therapeutic target of manipulation for IBD patients,[86] as multiple studies show the involvement of these metabolic processes and pathways with IBD.[81,87–89]

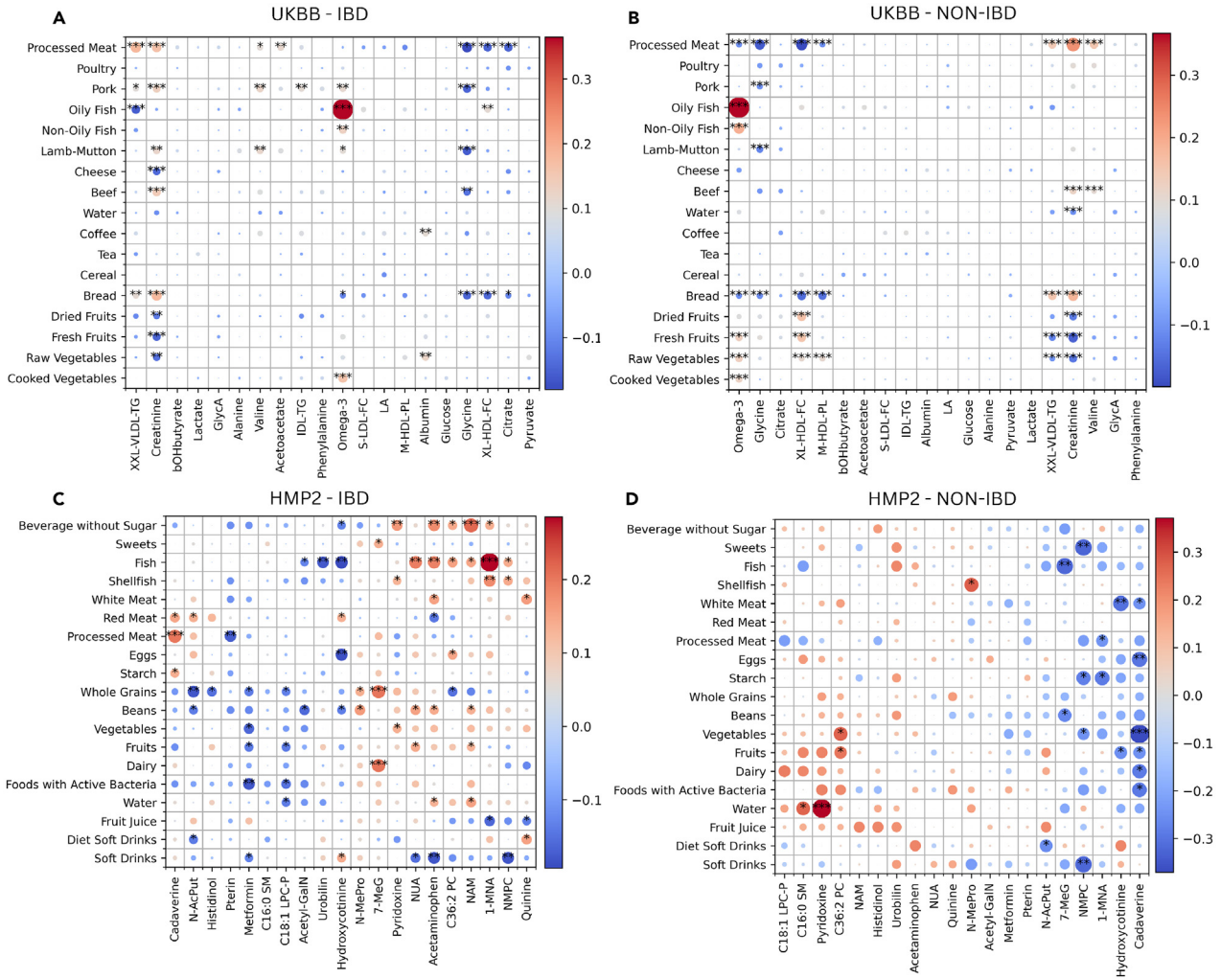


Figure 4. Diet-metabolite associations

This figure shows diet-metabolite heatmaps of the SHAP-calculated top 20 metabolites for the (A) IBD class of the UKBB, (B) non-IBD class of the UKBB, (C) IBD class of the HMP2, and (D) non-IBD class of the HMP2. Circles are color-coded to represent Spearman correlation values, with the circle size indicating the strength of correlation. Significance levels are denoted by asterisks (***: FDR < 0.001, ** 0.001 ≤ FDR < 0.01, *: 0.01 ≤ FDR < 0.05). The metabolites on the x-axis are "ward" clustered. See also Figures S4 and S5.

Phospholipid metabolites and urobilin were key discriminators as well. In line with existing studies,[75,90] higher LPC (C18:1 LPC-P) was associated with IBD classification, with higher concentrations consistently shown to promote inflammation, injuring endothelial cells[91] and damaging the epithelial barrier.[92] Furthermore, elevated fecal levels of SM (C16:1 SM) were predictive of IBD, consistent with prior research showing increased SMs in IBD patients.[19,93] Additionally, high urobilin levels were associated with non-IBD prediction, which is line with a previous study that found L-urobilin to be the most discriminative metabolite for the colitis phenotype in rats, with much higher concentrations being associated with the healthy, while non-existent in the colitis rats.[94] Moreover, urobilin emerged the top discriminator when NUA was not considered (Figure S5). Nonetheless, with a recent study suggesting an elevated ratio of fecal sphingolipids to L-urobilin as an IBD-associated marker warranting further investigation,[74] it may be the ratio to SMs and not the concentration in and of itself that is discriminatory.

Reasons for the disparity between the plasma and fecal metabolites

There was notable discrepancy in performance between fecal metabolites (AUC = 0.93) and plasma metabolites (AUC = 0.74). Insights obtained from the LASSO modeling in R (Figures S1 and S5) also complement the ensemble results from python; the stronger association with IBD was found to be with the fecal metabolites of HMP2 as opposed to the plasma metabolites of the UKBB, evidenced by more features selected above the bootstrap threshold, as well as higher AUC values (mean: 0.896 ± 0.025 vs. 0.652 ± 0.017) in the 400 different bootstraps.

The difference between the predictive capabilities of the plasma and fecal metabolites can be attributed to a number of biological factors, associated with cohort differences. One could be age differences. Elderly individuals, of which the UKBB comprises, regardless of disease may exhibit similar metabolic patterns in the blood, particularly with cell membrane related lipids such as phospholipids,[95] decreasing the discriminatory potential of the plasma metabolites. Moreover, the gut microbiome's pivotal role in IBD,[96] with its functions in inflammation regulation concerning the gastrointestinal tract[97] and overall gut health, render fecal metabolites as the more potent markers for IBD diagnosis. To buttress this point, just recently, Raygoza et al. show that the composition of the gut microbiome is linked to the future development of IBD, particularly CD.[98]

Diet-metabolome associations

The interplay between the metabolome and diet has been extensively studied,[99] and also between diet and IBD as a whole,[100] as mentioned in the introduction. However, there are gaps in understanding how the presence of IBD alters the metabolism of ingested food.

In the UKBB cohort's plasma analysis, increased consumption of oily fish, notably rich in long-chain omega-3 polyunsaturated fatty acids, was associated with increased omega-3 levels in both groups. Moreover, the consumption of red meat has also been associated with reduced levels of glycine in the plasma.[101–103] Studies also confirm the intake of fibers (fruit and vegetable), being positively correlated with omega-3 and HDL-C[104–106] and negatively with creatinine[107] observed among the non-IBD group. Considering that both groups had similar food intake distributions (Figure S3), and the overall metabolite profiles between the two groups did not generally exhibit visibly significant differences (Figure S4), a potential explanation for the slightly stronger correlation observed among those without IBD, could be the heterogeneity within the IBD cohort in terms of disease activity. A median time since diagnosis of 7 years, with the interquartile range going from 3 to 11 (Table 1), suggests varying disease activity levels. The subset of individuals with a longer disease history, potentially experiencing higher disease activity, could contribute to the observed metabolic profiles. During active IBD, or flare-ups in the case of inactivity, the body may respond to dietary intake in a distinct manner compared to those without active disease or without IBD. As this only possibly applies to a subset or multiple subsets within the whole IBD cohort, this variation could explain the weaker correlations observed within the IBD group.

In the HMP2 cohort however, the correlations were generally stronger than the ones observed in the UKBB, which is not surprising considering the well-established bidirectional relationship between the gut microbiome and diet.[108–111] Further, there were only a few consistent diet-metabolite associations shared between the diseased and non-diseased groups of the HMP2 cohort among the top 20 metabolites, and the most prominent are currently not confirmed in literature (e.g., NMPC with soft drinks, and hydroxycotinine with beans). Further, differences across both groups was much more numerous among the fecal metabolites. This divergence extended beyond slightly diminished correlation strengths to near-zero or opposite correlation strengths. For instance, the relationship between fish intake and urobilin varied significantly, being positively associated among those with IBD but negatively among non-IBD samples, with similar strengths observed (0.216 vs. −0.177; refer to Table S5 for all correlation and FDR values). This underscores the notion that disease, in this case, IBD, exerts a biological influence on how individuals respond to diet. Just as explained with the plasma metabolites, the reduced strengths of correlation when compared to the healthy could also be due to differential responses to dietary intake due to different activity states of IBD. However, as IBD is also an immune-mediated disease, its presence may diminish diet-metabolite correlations more significantly than if it was absent in an individual, as the gut microbiome also regulates the immune system and affects systemic immune responses.[112–114] Overall, considering the plasma and fecal dietary associations, the differential responses to diet when compared to the healthy, and even within IBD, possibly due to differing disease activity states, buttresses the need for more personalized approaches to dietary therapies for IBD management.

Limitations of the study

This study is not without limitations. The inclusive diagnosis of individuals using ICD-9 and ICD-10 codes in the UKBB, without accounting for comorbidities, introduces potential confounding factors that may have impacted the model's predictive accuracy, as these additional conditions might have strongly affected the metabolisms of some participants. The large UKBB sample size, while beneficial, may not have fully counteracted the effects of these confounding comorbidities. Another challenge arises from sample imbalance in both the UKBB and HMP2 cohorts, potentially introducing bias in model training and evaluation, particularly for the smaller HMP2 dataset, where the total number of samples was relatively small. This could be why the AUC score, which assesses a model's ability to predict the positive class, IBD, was higher than expected (AUC = 0.93) considering that IBD is a multifactorial disease. Nonetheless, variations in the timelines of dietary information between the two datasets, an average of the past year for the UKBB compared to the past week for the HMP2 cohort, may have impacted the representativeness of subjects' dietary habits during metabolic profiling, contributing to weaker correlations in plasma metabolites compared to fecal metabolites. Finally, the use of food questionnaires introduces a potential source of human error, as data accuracy relies on participants' recall and honesty.

Conclusion

Overall, our study addresses gaps in IBD research and lays a foundation for future studies by advancing our understanding of IBD pathogenesis through several key avenues. By leveraging the largely sampled UKBB data of over 2,500 individuals, to the best of our knowledge, this study represents the first-ever published comprehensive machine learning analysis of the plasma metabolome of IBD patients in the UKBB, offering unique insights. Further, by analyzing the well-documented fecal metabolites of the IBD samples of another major cohort, the HMP2, valuable contrasts are made available. Additionally, contributing to the work of developing diagnostic ML models drives us closer to developing an effective algorithm that predicts IBD before its onset, presenting a promising avenue for transforming patient care as more diverse risk factors such as smoking habits, and the presence of anti-saccharomyces cerevisiae antibodies in the blood are incorporated. Moreover,

the use of XAI in the predictive models offers transparency, facilitating translation into clinical practice. Finally, the exploration of dietary associations illuminates the complex interplay between gut microbial metabolites and dietary factors in IBD, enhancing our understanding of disease mechanisms, and facilitating the development of targeted interventions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Study design and participants
  - Metabolite data pre-processing
  - Diet data pre-processing
  - ML classification of IBD and non-IBD
  - Running the explainable AI on the best classifier
  - Statistical analysis of the top metabolites
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110298.

ACKNOWLEDGMENTS

The authors acknowledge support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, HYPERMARKER (grant agreement ID 101095480), and the MRC Health Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

AUTHOR CONTRIBUTIONS

A.A. conceptualized, developed methodology and designed the project; S.O. contributed to the data analysis and literature review; L.B.-M. contributed to data analysis and extracted data from Biobank; A.A., S.O., L.B.-M., and G.V.G. were involved in writing – review and editing first draft; and A.A. supervised the project. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, ChatGPT was used in order to paraphrase some sentences for better clarity and improved grammatical structure. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Received: February 12, 2024
Revised: April 29, 2024
Accepted: June 14, 2024
Published: June 17, 2024

REFERENCES

1. Wang, R., Li, Z., Liu, S., and Zhang, D. (2023). Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019. BMJ Open *13*, e065186. https://doi.org/10.1136/bmjopen-2022-065186.
2. Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I., Panaccione, R., Ghosh, S., Wu, J.C.Y., Chan, F.K.L., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of

population-based studies. Lancet *390*, 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0.

3. Geremia, A., Biancheri, P., Allan, P., Corazza, G.R., and Di Sabatino, A. (2014). Innate and adaptive immunity in inflammatory bowel disease. Autoimmun. Rev. *13*, 3–10. https://doi.org/10.1016/j.autrev.2013.06.004.
4. Neurath, M.F. (2014). Cytokines in inflammatory bowel disease. Nat. Rev. Immunol. *14*, 329–342. https://doi.org/10.1038/nri3661.
5. Neurath, M.F., and Schürmann, G. (2000). Immunopathogenesis of inflammatory bowel diseases. Chirurg *71*, 30–40. https://doi.org/10.1007/s001040050005.
6. de Mattos, B.R., Garcia, M.P., Nogueira, J.B., Paiatto, L.N., Albuquerque, C.G., Souza, C.L., Fernandes, L.G., Tamashiro, W.M., and Simioni, P.U. (2015). Inflammatory Bowel Disease: An Overview of Immune Mechanisms and Biological Treatments. Mediat. Inflamm. *2015*, 493012. https://doi.org/10.1155/2015/493012.
7. El Hadad, J., Schreiner, P., Vavricka, S.R., and Greuter, T. (2024). The Genetics of Inflammatory Bowel Disease. Mol. Diagn. Ther. *28*, 27–35. https://doi.org/10.1007/s40291-023-00678-7.
8. Turpin, W., Goethel, A., Bedrani, L., and Croitoru Mdcm, K. (2018). Determinants of IBD Heritability: Genes, Bugs, and More. Inflamm. Bowel Dis. *24*, 1133–1148. https://doi.org/10.1093/ibd/izy085.
9. Thompson, N.P., Driscoll, R., Pounder, R.E., and Wakefield, A.J. (1996). Genetics versus environment in inflammatory bowel disease: results of a British twin study. BMJ *312*, 95–96. https://doi.org/10.1136/bmj.312.7023.95.
10. Orholm, M., Binder, V., Sørensen, T.I., Rasmussen, L.P., and Kyvik, K.O. (2000). Concordance of Inflammatory Bowel Disease among Danish Twins: Results of a Nationwide Study. Scand. J. Gastroenterol. *35*, 1075–1081. https://doi.org/10.1080/003655200451207.
11. Lees, C.W., and Satsangi, J. (2009). Genetics of inflammatory bowel disease: implications for disease pathogenesis and natural history. Expet Rev. Gastroenterol. Hepatol. *3*, 513–534. https://doi.org/10.1586/egh.09.45.
12. Frolkis, A., Dieleman, L.A., Barkema, H.W., Panaccione, R., Ghosh, S., Fedorak, R.N., Madsen, K., and Kaplan, G.G.; Alberta IBD Consortium (2013). Environment and the Inflammatory Bowel Diseases. Can. J. Gastroenterol. *27*, e18–e24. https://doi.org/10.1155/2013/102859.
13. De Preter, V., and Verbeke, K. (2013). Metabolomics as a diagnostic tool in gastroenterology. World J. Gastrointest. Pharmacol. Therapeut *4*, 97–107. https://doi.org/10.4292/wjgpt.v4.i4.97.
14. Ippolito, J.E., Xu, J., Jain, S., Moulder, K., Mennerick, S., Crowley, J.R., Townsend, R.R., and Gordon, J.I. (2005). An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. Proc. Natl. Acad. Sci. USA *102*, 9901–9906. https://doi.org/10.1073/pnas.0500756102.
15. Lee, H.-S., Park, T.-J., Kim, J.-M., Yun, J.H., Yu, H.-Y., Kim, Y.-J., and Kim, B.-J. (2020). Identification of metabolic markers predictive of prediabetes in a Korean population. Sci. Rep. *10*, 22009. https://doi.org/10.1038/s41598-020-78961-4.
16. Ottosson, F., Smith, E., Ericson, U., Brunkwall, L., Orho-Melander, M., Di Somma, S., Antonini, P., Nilsson, P.M., Fernandez, C., and Melander, O. (2022). Metabolome-Defined Obesity and the Risk of Future Type 2 Diabetes and Mortality. Diabetes Care *45*, 1260–1267. https://doi.org/10.2337/dc21-2402.
17. Osadchiy, V., Bal, R., Mayer, E.A., Kunapuli, R., Dong, T., Vora, P., Petrasek, D., Liu, C., Stains, J., and Gupta, A. (2023). Machine learning model to predict obesity using gut metabolite and brain microstructure data. Sci. Rep. *13*, 5488. https://doi.org/10.1038/s41598-023-32713-2.
18. Aldars-García, L., Gisbert, J.P., and Chaparro, M. (2021). Metabolomics Insights into Inflammatory Bowel Disease: A Comprehensive Review. Pharmaceuticals *14*, 1190. https://doi.org/10.3390/ph14111190.
19. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. *4*, 293–305. https://doi.org/10.1038/s41564-018-0306-4.
20. Yang, Z.-H., Liu, F., Zhu, X.-R., Suo, F.-Y., Jia, Z.J., and Yao, S.-K. (2021). Altered profiles of fecal bile acids correlate with gut microbiota and inflammatory responses in patients with ulcerative colitis. World J. Gastroenterol. *27*, 3609–3629. https://doi.org/10.3748/wjg.v27.i24.3609.
21. Sun, Q., Jia, Q., Song, L., and Duan, L. (2019). Alterations in fecal short-chain fatty acids in patients with irritable bowel syndrome: A systematic review and meta-analysis. Medicine *98*, e14513. https://doi.org/10.1097/MD.0000000000014513.
22. Zhang, Y., Lin, L., Xu, Y., Lin, Y., Jin, Y., and Zheng, C. (2013). 1H NMR-based spectroscopy detects metabolic alterations in serum of patients with early-stage ulcerative colitis. Biochem. Biophys. Res. Commun. *433*, 547–551. https://doi.org/10.1016/j.bbrc.2013.03.012.
23. Chen, H., Li, W., Hu, J., Xu, F., Lu, Y., Zhu, L., and Shen, H. (2023). Association of serum lipids with inflammatory bowel disease: a systematic review and meta-analysis. Front. Med. *10*, 1198988. https://doi.org/10.3389/fmed.2023.1198988.
24. Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., and Schmitt-Kopplin, P. (2009). Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease. PLoS One *4*, e6386. https://doi.org/10.1371/journal.pone.0006386.
25. Murgia, A., Hinz, C., Liggi, S., Denes, J., Hall, Z., West, J., Santoru, M.L., Piras, C., Manis, C., Usai, P., et al. (2018). Italian cohort of patients affected by inflammatory bowel disease is characterised by variation in glycerophospholipid, free fatty acids and amino acid levels. Metabolomics *14*, 140. https://doi.org/10.1007/s11306-018-1439-4.
26. Notararigo, S., Martín-Pastor, M., Viñuela-Roldán, J.E., Quiroga, A., Dominguez-Munoz, J.E., and Barreiro-de Acosta, M. (2021). Targeted 1H NMR metabolomics and immunological phenotyping of human fresh blood and serum samples discriminate between healthy individuals and inflammatory bowel disease patients treated with anti-TNF. J. Mol. Med. *99*, 1251–1264. https://doi.org/10.1007/s00109-021-02094-y.
27. Schicho, R., Shaykhutdinov, R., Ngo, J., Nazyrova, A., Schneider, C., Panaccione, R., Kaplan, G.G., Vogel, H.J., and Storr, M. (2012). Quantitative Metabolomic Profiling of Serum, Plasma, and Urine by ^{1}H NMR Spectroscopy Discriminates between Patients with Inflammatory Bowel Disease and Healthy Individuals. J. Proteome Res. *11*, 3344–3357. https://doi.org/10.1021/pr300139q.
28. Santoru, M.L., Piras, C., Murgia, F., Leoni, V.P., Spada, M., Murgia, A., Liggi, S., Lai, M.A., Usai, P., Caboni, P., et al. (2021). Metabolic Alteration in Plasma and Biopsies From Patients With IBD. Inflamm. Bowel Dis. *27*, 1335–1345. https://doi.org/10.1093/ibd/izab012.
29. Kraszewski, S., Szczurek, W., Szymczak, J., Reguła, M., and Neubauer, K. (2021). Machine Learning Prediction Model for Inflammatory Bowel Disease Based on Laboratory Markers. Working Model in a Discovery Cohort Study. J. Clin. Med. *10*, 4745. https://doi.org/10.3390/jcm10204745.
30. Waljee, A.K., Wallace, B.I., Cohen-Mekelburg, S., Liu, Y., Liu, B., Sauder, K., Stidham, R.W., Zhu, J., and Higgins, P.D.R. (2019). Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease. JAMA Netw. Open *2*, e193721. https://doi.org/10.1001/jamanetworkopen.2019.3721.
31. Dong, Y., Xu, L., Fan, Y., Xiang, P., Gao, X., Chen, Y., Zhang, W., and Ge, Q. (2019). A novel surgical predictive model for Chinese Crohn's disease patients. Medicine *98*, e17510. https://doi.org/10.1097/MD.0000000000017510.
32. Wang, L., Fan, R., Zhang, C., Hong, L., Zhang, T., Chen, Y., Liu, K., Wang, Z., and Zhong, J. (2020). Applying Machine Learning Models to Predict Medication Nonadherence in Crohn's Disease Maintenance Therapy. Patient Prefer. Adherence *14*, 917–926. https://doi.org/10.2147/PPA.S253732.
33. Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy *23*, 18. https://doi.org/10.3390/e23010018.
34. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
35. Malinverno, L., Barros, V., Ghisoni, F., Visonà, G., Kern, R., Nickel, P.J., Ventura, B.E., Šimić, I., Stryeck, S., Manni, F., et al. (2023). A historical perspective of biomedical explainable AI research. Patterns *4*, 100830. https://doi.org/10.1016/j.patter.2023.100830.
36. Gunning, D., and Aha, D.W. (2019). DARPA's Explainable Artificial Intelligence Program. AI Mag. *40*, 44–58. https://doi.org/10.1609/aimag.v40i2.2850.
37. Gaundal, L., Myhrstad, M.C.W., Rud, I., Gjøvaag, T., Byfuglien, M.G., Retterstøl, K., Holven, K.B., Ulven, S.M., and Telle-Hansen, V.H. (2022). Gut microbiota is associated with dietary intake and metabolic markers in

healthy individuals. Food Nutr. Res. *66*. https://doi.org/10.29219/fnr.v66.8580.
38. Deng, M., Dan, L., Ye, S., Chen, X., Fu, T., Wang, X., and Chen, J. (2023). Higher dietary fibre intake is associated with lower risk of inflammatory bowel disease: prospective cohort study. Aliment. Pharmacol. Ther. *58*, 516–525. https://doi.org/10.1111/apt.17649.
39. Serrano Fernandez, V., Seldas Palomino, M., Laredo-Aguilera, J.A., Pozuelo-Carrascosa, D.P., and Carmona-Torres, J.M. (2023). High-Fiber Diet and Crohn's Disease: Systematic Review and Meta-Analysis. Nutrients *15*, 3114. https://doi.org/10.3390/nu15143114.
40. Ananthakrishnan, A.N., Khalili, H., Konijeti, G.G., Higuchi, L.M., de Silva, P., Korzenik, J.R., Fuchs, C.S., Willett, W.C., Richter, J.M., and Chan, A.T. (2013). A Prospective Study of Long-term Intake of Dietary Fiber and Risk of Crohn's Disease and Ulcerative Colitis. Gastroenterology *145*, 970–977. https://doi.org/10.1053/j.gastro.2013.07.050.
41. Li, T., Qiu, Y., Yang, H.S., Li, M.Y., Zhuang, X.J., Zhang, S.H., Feng, R., Chen, B.L., He, Y., Zeng, Z.R., and Chen, M.H. (2020). Systematic review and meta-analysis: Association of a pre-illness Western dietary pattern with the risk of developing inflammatory bowel disease. J. Dig. Dis. *21*, 362–371. https://doi.org/10.1111/1751-2980.12910.
42. Rizzello, F., Spisni, E., Giovanardi, E., Imbesi, V., Salice, M., Alvisi, P., Valerii, M.C., and Gionchetti, P. (2019). Implications of the Westernized Diet in the Onset and Progression of IBD. Nutrients *11*, 1033. https://doi.org/10.3390/nu11051033.
43. Brown, K., DeCoffe, D., Molcan, E., and Gibson, D.L. (2012). Diet-Induced Dysbiosis of the Intestinal Microbiota and the Effects on Immunity and Disease. Nutrients *4*, 1095–1119. https://doi.org/10.3390/nu4081095.
44. González-Torres, L., Moreno-Álvarez, A., Fernández-Lorenzo, A.E., Leis, R., and Solar-Boga, A. (2022). The Role of Partial Enteral Nutrition for Induction of Remission in Crohn's Disease: A Systematic Review of Controlled Trials. Nutrients *14*, 5263. https://doi.org/10.3390/nu14245263.
45. Buchanan, E., Gaunt, W.W., Cardigan, T., Garrick, V., Mcgrogan, P., and Russell, R.K. (2009). The use of exclusive enteral nutrition for induction of remission in children with Crohn's disease demonstrates that disease phenotype does not influence clinical remission. Aliment. Pharmacol. Ther. *30*, 501–507. https://doi.org/10.1111/j.1365-2036.2009.04067.x.
46. Yang, Q., Gao, X., Chen, H., Li, M., Wu, X., Zhi, M., Lan, P., and Hu, P. (2017). Efficacy of exclusive enteral nutrition in complicated Crohn's disease. Scand. J. Gastroenterol. *52*, 995–1001. https://doi.org/10.1080/00365521.2017.1335770.
47. Obih, C., Wahbeh, G., Lee, D., Braly, K., Giefer, M., Shaffer, M.L., Nielson, H., and Suskind, D.L. (2016). Specific carbohydrate diet for pediatric inflammatory bowel disease in clinical practice within an academic IBD center. Nutrition *32*, 418–425. https://doi.org/10.1016/j.nut.2015.08.025.
48. Suskind, D.L., Cohen, S.A., Brittnacher, M.J., Wahbeh, G., Lee, D., Shaffer, M.L., Braly, K., Hayden, H.S., Klein, J., Gold, B., et al. (2018). Clinical and Fecal Microbial Changes With Diet Therapy in Active Inflammatory Bowel Disease. J. Clin. Gastroenterol. *52*, 155–163. https://doi.org/10.1097/MCG.0000000000000772.
49. Suskind, D.L., Lee, D., Kim, Y.-M., Wahbeh, G., Singh, N., Braly, K., Nuding, M., Nicora, C.D., Purvine, S.O., Lipton, M.S., et al. (2020). The Specific Carbohydrate Diet and Diet Modification as Induction Therapy for Pediatric Crohn's Disease: A Randomized Diet Controlled Trial. Nutrients *12*, 3749. https://doi.org/10.3390/nu12123749.
50. Dixon, L.J., Kabi, A., Nickerson, K.P., and McDonald, C. (2015). Combinatorial Effects of Diet and Genetics on Inflammatory Bowel Disease Pathogenesis. Inflamm. Bowel Dis. *21*, 912–922. https://doi.org/10.1097/MIB.0000000000000289.
51. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.
52. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9.
53. Mehta, N.N., Dey, A.K., Maddineni, R., Kraus, W.E., and Huffman, K.M. (2020). GlycA measured by NMR spectroscopy is associated with disease activity and cardiovascular disease risk in chronic inflammatory diseases. Am. J. Prev. Cardiol. *4*, 100120. https://doi.org/10.1016/j.ajpc.2020.100120.
54. Chiesa, S.T., Charakida, M., Georgiopoulos, G., Roberts, J.D., Stafford, S.J., Park, C., Mykkänen, J., Kähönen, M., Lehtimäki, T., Ala-Korpela, M., et al. (2022). Glycoprotein Acetyls: A Novel Inflammatory Biomarker of Early Cardiovascular Risk in the Young. J. Am. Heart Assoc. *11*, e024380. https://doi.org/10.1161/JAHA.121.024380.
55. Ritchie, S.C., Würtz, P., Nath, A.P., Abraham, G., Havulinna, A.S., Fearnley, L.G., Sarin, A.-P., Kangas, A.J., Soininen, P., Aalto, K., et al. (2015). The Biomarker GlycA Is Associated with Chronic Inflammation and Predicts Long-Term Risk of Severe Infection. Cell Syst. *1*, 293–301. https://doi.org/10.1016/j.cels.2015.09.007.
56. Dierckx, T., Verstockt, B., Vermeire, S., and van Weyenbergh, J. (2019). GlycA, a Nuclear Magnetic Resonance Spectroscopy Measure for Protein Glycosylation, is a Viable Biomarker for Disease Activity in IBD. J. Crohns Colitis *13*, 389–394. https://doi.org/10.1093/ecco-jcc/jjy162.
57. Vermeire, S., Van Assche, G., and Rutgeerts, P. (2006). Laboratory markers in IBD: useful, magic, or unnecessary toys? Gut *55*, 426–431. https://doi.org/10.1136/gut.2005.069476.
58. Hrabovský, V., Zadák, Z., Bláha, V., Hyšpler, R., Karlík, T., Martínek, A., and Mendlová, A. (2009). Cholesterol metabolism in active Crohn's disease. Wien Klin. Wochenschr. *121*, 270–275. https://doi.org/10.1007/s00508-009-1150-6.
59. Soh, H., Im, J.P., Han, K., Park, S., Hong, S.W., Moon, J.M., Kang, E.A., Chun, J., Lee, H.J., and Kim, J.S. (2020). Crohn's disease and ulcerative colitis are associated with different lipid profile disorders: a nationwide population-based study. Aliment. Pharmacol. Ther. *51*, 446–456. https://doi.org/10.1111/apt.15562.
60. Sappati Biyyani, R.S.R., Putka, B.S., and Mullen, K.D. (2010). Dyslipidemia and lipoprotein profiles in patients with inflammatory bowel disease. J. Clin. Lipidol. *4*, 478–482. https://doi.org/10.1016/j.jacl.2010.08.021.
61. Koutroumpakis, E., Ramos-Rivers, C., Regueiro, M., Hashash, J.G., Barrie, A., Swoger, J., Baidoo, L., Schwartz, M., Dunn, M.A., Koutroubakis, I.E., and Binion, D.G. (2016). Association Between Long-Term Lipid Profiles and Disease Severity in a Large Cohort of Patients with Inflammatory Bowel Disease. Dig. Dis. Sci. *61*, 865–871. https://doi.org/10.1007/s10620-015-3932-1.
62. Mooradian, A.D. (2009). Dyslipidemia in type 2 diabetes mellitus. Nat. Rev. Endocrinol. *5*, 150–159. https://doi.org/10.1038/ncpendmet1066.
63. Arsenault, B.J., Boekholdt, S.M., and Kastelein, J.J.P. (2011). Lipid parameters for measuring risk of cardiovascular disease. Nat. Rev. Cardiol. *8*, 197–206. https://doi.org/10.1038/nrcardio.2010.223.
64. Chen, Q.-J., Lai, H.-M., Chen, B.-D., Li, X.-M., Zhai, H., He, C.-H., Pan, S., Luo, J.-Y., Gao, J., Liu, F., et al. (2016). Appropriate LDL-C-to-HDL-C Ratio Cutoffs for Categorization of Cardiovascular Disease Risk Factors among Uygur Adults in Xinjiang, China. Int. J. Environ. Res. Publ. Health *13*, 235. https://doi.org/10.3390/ijerph13020235.
65. Feng, W., Chen, G., Cai, D., Zhao, S., Cheng, J., and Shen, H. (2017). Inflammatory Bowel Disease and Risk of Ischemic Heart Disease: An Updated Meta-Analysis of Cohort Studies. J. Am. Heart Assoc. *6*, e005892. https://doi.org/10.1161/JAHA.117.005892.
66. Lee, M.T., Mahtta, D., Chen, L., Hussain, A., Al Rifai, M., Sinh, P., Khalid, U., Nasir, K., Ballantyne, C.M., Petersen, L.A., and Virani, S.S. (2021). Premature Atherosclerotic Cardiovascular Disease Risk Among Patients with Inflammatory Bowel Disease. Am. J. Med. *134*, 1047–1051.e2. https://doi.org/10.1016/j.amjmed.2021.02.029.
67. Chen, B., Collen, L.V., Mowat, C., Isaacs, K.L., Singh, S., Kane, S.V., Farraye, F.A., Snapper, S., Jneid, H., Lavie, C.J., and Krittanawong, C. (2022). Inflammatory Bowel Disease and Cardiovascular Diseases. Am. J. Med. *135*, 1453–1460. https://doi.org/10.1016/j.amjmed.2022.08.012.
68. Scaioli, E., Liverani, E., and Belluzzi, A. (2017). The Imbalance between n-6/n-3 Polyunsaturated Fatty Acids and Inflammatory Bowel Disease: A Comprehensive Review and Future Therapeutic Perspectives. Int. J. Mol. Sci. *18*, 2619. https://doi.org/10.3390/ijms18122619.
69. Bugajska, J., Berska, J., Zwolińska-Wcisło, M., and Sztefko, K. (2022). The risk of essential fatty acid insufficiency in patients with inflammatory bowel diseases: fatty acid profile of phospholipids in serum and in colon biopsy specimen. Arch. Med. Sci. *18*, 1103–1107. https://doi.org/10.5114/aoms/150041.
70. Gutiérrez, S., Svahn, S.L., and Johansson, M.E. (2019). Effects of Omega-3 Fatty Acids on Immune Cells. Int. J. Mol. Sci. *20*, 5028. https://doi.org/10.3390/ijms20205028.

71. Marton, L.T., Goulart, R.d.A., Carvalho, A.C.A.d., and Barbalho, S.M. (2019). Omega fatty acids and inflammatory bowel diseases: An overview. Int. J. Mol. Sci. *20*, 4851. https://doi.org/10.3390/ijms20194851.
72. Huang, X., Li, Y., Zhuang, P., Liu, X., Zhang, Y., Zhang, P., and Jiao, J. (2022). Habitual Fish Oil Supplementation and Risk of Incident Inflammatory Bowel Diseases: A Prospective Population-Based Study. Front. Nutr. *9*, 905162. https://doi.org/10.3389/fnut.2022.905162.
73. Barbalho, S.M., Goulart, R.d.A., Quesada, K., Bechara, M.D., and de Carvalho, A.d.C.A. (2016). Inflammatory bowel disease: can omega-3 fatty acids really help? Ann. Gastroenterol. *29*, 37–43.
74. Vich Vila, A., Hu, S., Andreu-Sánchez, S., Collij, V., Jansen, B.H., Augustijn, H.E., Bolte, L.A., Ruigrok, R.A.A.A., Abu-Ali, G., Giallourakis, C., et al. (2023). Faecal metabolome and its determinants in inflammatory bowel disease. Gut *72*, 1472–1485. https://doi.org/10.1136/gutjnl-2022-328048.
75. Wu, X., Liu, K., Wu, Q., Wang, M., Chen, X., Li, Y., Qian, L., Li, C., Dai, G., Zhang, Q., et al. (2022). Biomarkers of Metabolomics in Inflammatory Bowel Disease and Damp-Heat Syndrome: A Preliminary Study. Evid. Based. Complement. Alternat. Med. *2022*, 3319646. https://doi.org/10.1155/2022/3319646.
76. Levhar, N., Hadar, R., Braun, T., Efroni, G., Naamneh, R., Agranovich, B., Talan Asher, A., Selinger, L., Picard, O., Lahat, A., et al. (2024). DOP09 Models for predicting Crohn Disease (CD) exacerbation using serum and fecal metabolomics. J. Crohns Colitis *18*, i88–i89. https://doi.org/10.1093/ecco-jcc/jjad212.0049.
77. Santoru, M.L., Piras, C., Murgia, A., Palmas, V., Camboni, T., Liggi, S., Ibba, I., Lai, M.A., Orrù, S., Blois, S., et al. (2017). Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. Sci. Rep. *7*, 9523. https://doi.org/10.1038/s41598-017-10034-5.
78. Mehta, R.S., Mayers, J.R., Zhang, Y., Bhosle, A., Glasser, N.R., Nguyen, L.H., Ma, W., Bae, S., Branck, T., Song, K., et al. (2023). Gut microbial metabolism of 5-ASA diminishes its clinical efficacy in inflammatory bowel disease. Nat. Med. *29*, 700–709. https://doi.org/10.1038/s41591-023-02217-7.
79. Ungaro, R., Mehandru, S., Allen, P.B., Peyrin-Biroulet, L., and Colombel, J.-F. (2017). Ulcerative colitis. Lancet *389*, 1756–1770. https://doi.org/10.1016/S0140-6736(16)32126-2.
80. Laharie, D. (2017). Towards therapeutic choices in ulcerative colitis. Lancet *390*, 98–99. https://doi.org/10.1016/S0140-6736(17)31263-1.
81. Kang, Y.H., Tucker, S.A., Quevedo, S.F., Inal, A., Korzenik, J.R., and Haigis, M.C. (2022). Metabolic analyses reveal dysregulated NAD+ metabolism and altered mitochondrial state in ulcerative colitis. PLoS One *17*, e0273080. https://doi.org/10.1371/journal.pone.0273080.
82. Diab, J., Hansen, T., Goll, R., Stenlund, H., Jensen, E., Moritz, T., Florholmen, J., and Forsdahl, G. (2019). Mucosal Metabolomic Profiling and Pathway Analysis Reveal the Metabolic Signature of Ulcerative Colitis. Metabolites *9*, 291. https://doi.org/10.3390/metabo9120291.
83. Niño-Narvión, J., Rojo-López, M.I., Martinez-Santos, P., Rossell, J., Ruiz-Alcaraz, A.J., Alonso, N., Ramos-Molina, B., Mauricio, D., and Julve, J. (2023). NAD+ Precursors and Intestinal Inflammation: Therapeutic Insights Involving Gut Microbiota. Nutrients *15*, 2992. https://doi.org/10.3390/nu15132992.
84. Lenglet, A., Liabeuf, S., Bodeau, S., Louvet, L., Mary, A., Boullier, A., Lemaire-Hurtel, A.S., Jonet, A., Sonnet, P., Kamel, S., and Massy, Z.A. (2016). N-methyl-2-pyridone-5-carboxamide (2PY)—Major Metabolite of Nicotinamide: An Update on an Old Uremic Toxin. Toxins *8*, 339. https://doi.org/10.3390/toxins8110339.
85. Xue, X., Miao, Y., and Wei, Z. (2023). Nicotinamide adenine dinucleotide metabolism: driving or counterbalancing inflammatory bowel disease? FEBS Lett. *597*, 1179–1192. https://doi.org/10.1002/1873-3468.14528.
86. Chen, C., Yan, W., Tao, M., and Fu, Y. (2023). NAD+ Metabolism and Immune Regulation: New Approaches to Inflammatory Bowel Disease Therapies. Antioxidants *12*, 1230. https://doi.org/10.3390/antiox12061230.
87. Ning, L., Shan, G., Sun, Z., Zhang, F., Xu, C., Lou, X., Li, S., Du, H., Chen, H., and Xu, G. (2019). Quantitative Proteomic Analysis Reveals the Deregulation of Nicotinamide Adenine Dinucleotide Metabolism and CD38 in Inflammatory Bowel Disease. BioMed Res. Int. *2019*, 3950628. https://doi.org/10.1155/2019/3950628.
88. Schneider, M., Schumacher, V., Lischke, T., Lücke, K., Meyer-Schwesinger, C., Velden, J., Koch-Nolte, F., and Mittrücker, H.-W. (2015). CD38 Is Expressed on Inflammatory Cells of the Intestine and Promotes Intestinal Inflammation. PLoS One *10*, e0126007. https://doi.org/10.1371/journal.pone.0126007.
89. Gerner, R.R., Klepsch, V., Macheiner, S., Arnhard, K., Adolph, T.E., Grander, C., Wieser, V., Pfister, A., Moser, P., Hermann-Kleiter, N., et al. (2018). NAD metabolism fuels human and mouse intestinal inflammation. Gut *67*, 1813–1823. https://doi.org/10.1136/gutjnl-2017-314241.
90. Tefas, C., Ciobanu, L., Tanțău, M., Moraru, C., and Socaciu, C. (2019). The potential of metabolic and lipid profiling in inflammatory bowel diseases: a pilot study. Bosn. J. Basic Med. Sci. *20*, 262. https://doi.org/10.17305/bjbms.2019.4235.
91. Chang, M.-C., Lee, J.-J., Chen, Y.-J., Lin, S.-I., Lin, L.-D., Jein-Wen Liou, E., Huang, W.-L., Chan, C.-P., Huang, C.-C., and Jeng, J.-H. (2017). Lysophosphatidylcholine induces cytotoxicity/apoptosis and IL-8 production of human endothelial cells: Related mechanisms. Oncotarget *8*, 106177–106189. https://doi.org/10.18632/oncotarget.22425.
92. Tang, X., Wang, W., Hong, G., Duan, C., Zhu, S., Tian, Y., Han, C., Qian, W., Lin, R., and Hou, X. (2021). Gut microbiota-mediated lysophosphatidylcholine generation promotes colitis in intestinal epithelium-specific Fut2 deficiency. J. Biomed. Sci. *28*, 20. https://doi.org/10.1186/s12929-021-00711-z.
93. Braun, A., Treede, I., Gotthardt, D., Tietje, A., Zahn, A., Ruhwald, R., Schoenfeld, U., Welsch, T., Kienle, P., Erben, G., et al. (2009). Alterations of phospholipid concentration and species composition of the intestinal mucus barrier in ulcerative colitis: A clue to pathogenesis. Inflamm. Bowel Dis. *15*, 1705–1720. https://doi.org/10.1002/ibd.20993.
94. Liu, F., Liu, J., Wang, T.T.Y., Liu, Z., Xue, C., Mao, X., Tang, Q., and Li, R.W. (2020). Molecular and Microbial Signatures Predictive of Prebiotic Action of Neoagarotetraose in a Dextran Sulfate Sodium-Induced Murine Colitis Model. Microorganisms *8*, 995. https://doi.org/10.3390/microorganisms8070995.
95. Yu, Z., Zhai, G., Singmann, P., He, Y., Xu, T., Prehn, C., Römisch-Margl, W., Lattka, E., Gieger, C., Soranzo, N., et al. (2012). Human serum metabolic profiles are age dependent. Aging Cell *11*, 960–967. https://doi.org/10.1111/j.1474-9726.2012.00865.x.
96. Qiu, P., Ishimoto, T., Fu, L., Zhang, J., Zhang, Z., and Liu, Y. (2022). The Gut Microbiota in Inflammatory Bowel Disease. Front. Cell. Infect. Microbiol. *12*, 733992. https://doi.org/10.3389/fcimb.2022.733992.
97. Al Bander, Z., Nitert, M.D., Mousa, A., and Naderpoor, N. (2020). The Gut Microbiota and Inflammation: An Overview. Int. J. Environ. Res. Publ. Health *17*, 7618. https://doi.org/10.3390/ijerph17207618.
98. Raygoza Garay, J.A., Turpin, W., Lee, S.-H., Smith, M.I., Goethel, A., Griffiths, A.M., Moayyedi, P., Espin-Garcia, O., Abreu, M., Aumais, G.L., et al. (2023). Gut Microbiome Composition Is Associated With Future Onset of Crohn's Disease in Healthy First-Degree Relatives. Gastroenterology *165*, 670–681. https://doi.org/10.1053/j.gastro.2023.05.032.
99. Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L.O., Draper, J., Rappaport, S.M., van der Hooft, J.J.J., and Wishart, D.S. (2014). The food metabolome: a window over dietary exposure. Am. J. Clin. Nutr. *99*, 1286–1308. https://doi.org/10.3945/ajcn.113.076133.
100. Khalili, H., Chan, S.S.M., Lochhead, P., Ananthakrishnan, A.N., Hart, A.R., and Chan, A.T. (2018). The role of diet in the aetiopathogenesis of inflammatory bowel disease. Nat. Rev. Gastroenterol. Hepatol. *15*, 525–535. https://doi.org/10.1038/s41575-018-0022-9.
101. Schmidt, J.A., Rinaldi, S., Scalbert, A., Ferrari, P., Achaintre, D., Gunter, M.J., Appleby, P.N., Key, T.J., and Travis, R.C. (2016). Plasma concentrations and intakes of amino acids in male meat-eaters, fish-eaters, vegetarians and vegans: a cross-sectional analysis in the EPIC-Oxford cohort. Eur. J. Clin. Nutr. *70*, 306–312. https://doi.org/10.1038/ejcn.2015.144.
102. Wittenbecher, C., Mühlenbruch, K., Kröger, J., Jacobs, S., Kuxhaus, O., Floegel, A., Fritsche, A., Pischon, T., Prehn, C., Adamski, J., et al. (2015). Amino acids, lipid metabolites, and ferritin as potential mediators linking red meat consumption to type 2 diabetes. Am. J. Clin. Nutr. *101*, 1241–1250. https://doi.org/10.3945/ajcn.114.099150.
103. Altorf-van der Kuil, W., Brink, E.J., Boetje, M., Siebelink, E., Bijlsma, S., Engberink, M.F., van't Veer, P., Tomé, D., Bakker, S.J.L., van Baak, M.A., and Geleijnse, J.M. (2013). Identification of biomarkers for intake of protein from meat, dairy products and grains: a controlled dietary intervention study. Br. J. Nutr. *110*, 810–822. https://doi.org/10.1017/S0007114512005788.
104. Rondanelli, M., Giacosa, A., Morazzoni, P., Guido, D., Grassi, M., Morandi, G., Bologna, C., Riva, A., Allegrini, P., and Perna, S. (2016).

MediterrAsian Diet Products That Could Raise HDL-Cholesterol: A Systematic Review. BioMed Res. Int. *2016*, 2025687. https://doi.org/10.1155/2016/2025687.

105. Keung, V., Lo, K., Cheung, C., Tam, W., and Lee, A. (2019). Changes in dietary habits and prevalence of cardiovascular risk factors among school students in Macao, China. Obes. Res. Clin. Pract. *13*, 541–547. https://doi.org/10.1016/j.orcp.2019.10.007.
106. Liu, J., Li, Y., Wang, X., Gao, D., Chen, L., Chen, M., Ma, T., Ma, Q., Ma, Y., Zhang, Y., et al. (2021). Association between Fruit Consumption and Lipid Profile among Children and Adolescents: A National Cross-Sectional Study in China. Nutrients *14*, 63. https://doi.org/10.3390/nu14010063.
107. Nakano, T., Tanaka, S., Tsuruya, K., and Kitazono, T. (2022). Low intake of β carotene and dietary fiber from vegetables and fruits in patients with chronic kidney disease. Sci. Rep. *12*, 19953. https://doi.org/10.1038/s41598-022-24471-4.
108. Rinninella, E., Cintoni, M., Raoul, P., Lopetuso, L.R., Scaldaferri, F., Pulcini, G., Miggiano, G.A.D., Gasbarrini, A., and Mele, M.C. (2019). Food Components and Dietary Habits: Keys for a Healthy Gut Microbiota Composition. Nutrients *11*, 2393. https://doi.org/10.3390/nu11102393.
109. Deehan, E.C., Yang, C., Perez-Muñoz, M.E., Nguyen, N.K., Cheng, C.C., Triador, L., Zhang, Z., Bakal, J.A., and Walter, J. (2020). Precision Microbiome Modulation with Discrete Dietary Fiber Structures Directs Short-Chain Fatty Acid Production. Cell Host Microbe *27*, 389–404.e6. https://doi.org/10.1016/j.chom.2020.01.006.
110. Wastyk, H.C., Fragiadakis, G.K., Perelman, D., Dahan, D., Merrill, B.D., Yu, F.B., Topf, M., Gonzalez, C.G., Van Treuren, W., Han, S., et al. (2021). Gut-microbiota-targeted diets modulate human immune status. Cell *184*, 4137–4153.e14. https://doi.org/10.1016/j.cell.2021.06.019.
111. Lakshmanan, A.P., Mingione, A., Pivari, F., Dogliotti, E., Brasacchio, C., Murugesan, S., Cusi, D., Lazzaroni, M., Soldati, L., and Terranegra, A. (2022). Modulation of gut microbiota: The effects of a fruits and vegetables supplement. Front. Nutr. *9*, 930883. https://doi.org/10.3389/fnut.2022.930883.
112. Honda, K., and Littman, D.R. (2016). The microbiota in adaptive immune homeostasis and disease. Nature *535*, 75–84. https://doi.org/10.1038/nature18848.
113. Blander, J.M., Longman, R.S., Iliev, I.D., Sonnenberg, G.F., and Artis, D. (2017). Regulation of inflammation by microbiota interactions with the host. Nat. Immunol. *18*, 851–860. https://doi.org/10.1038/ni.3780.
114. Wiertsema, S.P., van Bergenhenegouwen, J., Garssen, J., and Knippels, L.M.J. (2021). The Interplay between the Gut Microbiome and the Immune System in the Context of Infectious Diseases throughout Life and the Role of Nutrition in Optimizing Treatment Strategies. Nutrients *13*, 886. https://doi.org/10.3390/nu13030886.
115. Mahalanobis, P.C. (1936). On the generalized distance in statistics. Proc. Natl. Inst. Sci. (Calcutta) *2*, 49–55.
116. R Core Team (2023). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
117. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Software *42*, 1–28. https://doi.org/10.18637/jss.v042.i08.
118. Walter, R.I. (1959). Nuclear magnetic resonance. J. Chem. Educ. *36*, 531. https://doi.org/10.1021/ed036p531.1.
119. Zhou, B., Xiao, J.F., Tuli, L., and Ressom, H.W. (2012). LC-MS-based metabolomics. Mol. Biosyst. *8*, 470–481. https://doi.org/10.1039/C1MB05350G.
120. Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V., Nokso-Koivisto, J., Kristiansson, K., Perola, M., Salomaa, V., et al. (2023). Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nat. Commun. *14*, 604. https://doi.org/10.1038/s41467-023-36231-7.
121. Stekhoven, D.J., and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics *28*, 112–118. https://doi.org/10.1093/bioinformatics/btr597.
122. Grace, S.C., and Hudson, D.A. (2016). Processing and Visualization of Metabolomics Data Using R. In Metabolomics - Fundamentals and Applications (InTech) (IntechOpen). Chapter 4. https://doi.org/10.5772/65405.
123. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom. *7*, 142. https://doi.org/10.1186/1471-2164-7-142.
124. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Software *28*, 1–24. https://doi.org/10.18637/jss.v028.i05.
125. Dong, X., Zhuang, Z., Zhao, Y., Song, Z., Xiao, W., Wang, W., Li, Y., Huang, N., Jia, J., Liu, Z., et al. (2023). Unprocessed Red Meat and Processed Meat Consumption, Plasma Metabolome, and Risk of Ischemic Heart Disease: A Prospective Cohort Study of UK Biobank. J. Am. Heart Assoc. *12*, e027934. https://doi.org/10.1161/JAHA.122.027934.
126. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
127. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.).
128. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
129. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Stat. Soc. B *58*, 267–288.
130. Van Rossum, G., and Drake Jr, F.L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam *111*, 1–152.
131. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195.
132. Bravo-Merodio, L., Acharjee, A., Hazeldine, J., Bentley, C., Foster, M., Gkoutos, G.V., and Lord, J.M. (2019). Machine learning for the detection of early immunological markers as predictors of multi-organ dysfunction. Sci. Data *6*, 328. https://doi.org/10.1038/s41597-019-0337-6.
133. Bravo-Merodio, L., Williams, J.A., Gkoutos, G.V., and Acharjee, A. (2019). -Omics biomarker identification pipeline for translational medicine. J. Transl. Med. *17*, 155. https://doi.org/10.1186/s12967-019-1912-5.
134. Lundberg, S., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions.
135. Shapley, L.S. (1953). 17. A Value for n-Person Games. In Contributions to the Theory of Games (AM-28), *II*Contributions to the Theory of Games (AM-28) (Princeton University Press), pp. 307–318. https://doi.org/10.1515/9781400881970-018.
136. Vallat, R. (2018). Pingouin: statistics in Python. J. Open Source Softw. *3*, 1026. https://doi.org/10.21105/joss.01026.
137. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.
138. Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021. https://doi.org/10.21105/joss.03021.
139. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.
140. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
141. Shapiro, S.S., and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika *52*, 591. https://doi.org/10.2307/2333709.
142. Onwuka, S., and Bravo Merodio, L. (2024). XAIMetabolomeDiet: v1.0. Zenodo. https://doi.org/10.5281/zenodo.11411432.
143. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software *33*, 1–22. https://doi.org/10.18637/jss.v033.i01.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Plasma metabolomics data	UK Biobank[51]	UK Biobank: 31224;
Fecal metabolomics data	HMP2 IBDMDB[52]	https://ibdmdb.org/results
Software and algorithms		
Codes generated for this study	This paper[142]	Zenodo: https://doi.org/10.5281/zenodo.11411432;
R	R Core Team[116]	https://cran.r-project.org/
tidyverse	Wickham et al.[139]	https://github.com/tidyverse/tidyverse
missForest	Stekhoven & Bühlmann[121]	https://github.com/stekhoven/missForest
caret	Kuhn[124]	https://github.com/topepo/caret
glmnet	Friedman, Hastie & Tibshirani[143]	https://github.com/cran/glmnet
Python	Van Rossum et al.[130]	https://www.python.org/downloads/
xgboost	Chen & Guestrin[126]	https://github.com/dmlc/xgboost
lightgbm	Ke, Meng, Finley, Wang, Chen, Ma, Ye & Liu T[127]	https://github.com/microsoft/LightGBM
sklearn.ensemble.RandomForestClassifier	Pedregosa et al.[131]	https://github.com/scikit-learn/scikit-learn
sklearn.linear_model.Lasso	Pedregosa et al.[131]	https://github.com/scikit-learn/scikit-learn
shap	Lunderg & Lee[134]	https://github.com/shap/shap
pingouin	Vallat[136]	https://github.com/raphaelvallat/pingouin
matplotlib	Hunter[137]	https://github.com/matplotlib/matplotlib
seaborn	Waskom[138]	https://github.com/mwaskom/seaborn

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr Animesh Acharjee (a.acharjee@bham.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Accessibility details for these datasets are listed in the key resources table.
- This paper does not report original code. Derivative codes generated for this study have been deposited at Zenodo and is publicly available as of the date of publication. The DOI is listed as a citation in the key resources table.
- Additional information needed to reanalyze the data presented in this paper can be obtained from the lead contact upon request.

METHOD DETAILS

Study design and participants

The blood-based metabolomics data in this study was retrieved from the UKBB. The UKBB is a large-scale cohort study conducted between 2006 and 2010, involving 500,000 consenting participants aged 40 to 69 from across the UK who provided detailed health information, and is approved by the North West Multi-centre Research Ethics Committee amongst others. At the data retrieval stage, 4,137 plasma samples (2,099 IBD and 2,038 healthy controls—no reported ICD-10 diagnosis—matched based on age and sex) were extracted from the UKBB. Matching was done using the "nearest" method which utilizes a greedy search to match each sample with their nearest neighbour. The distance was calculated using the Mahalanobis distance, which estimates the distribution closest for each point.[115] This procedure was

performed in R (v4.2)[116] using the MatchIt package.[117] IBD diagnosis was defined by corresponding International Disease codes (ICD-9: 555, 556; ICD-10: K50, K51) which included both Crohn's disease (CD) and ulcerative colitis (UC). Self-reported cases of IBD (1461, 1462, 1463) were also considered.

The fecal-based metabolomics data in this study was gotten from the HMP2 IBDMDB. As part of the Integrative Human Microbiome Project, which is carried out under the National Institute of Health (NIH), the IBDMDB followed 132 consenting subjects over a year to generate comprehensive longitudinal molecular profiles of host and microbial activity during IBD. In this cohort, 546 fecal samples (411 IBD samples encompassing both CD and UC, and 135 control samples) were retrieved from the stools of 106 individuals, with an average of 5.6 samples per person (SD = 1.2), having been followed longitudinally for up to one year each.

The datasets of both cohorts were filtered to retain only individuals with complete data for their corresponding food questionnaire. Individuals who developed IBD after baseline were excluded in both cohorts, considering 'Age at recruitment' (Code: 21022) for the UKBB and 'age at consent' for the HMP2 as the baseline ages. These included IBD participants that had a history of disease of varying years ('Median years since diagnosis' in Table 1). Following filtering, multiple samples per participant within the HMP2 cohort of fecal samples were retained where applicable. This approach allowed for data to be maximally utilized while avoiding disproportionate emphasis on specific features, as temporal changes in the microbiome are more frequent and pronounced in IBD.[52]

Metabolite data pre-processing

Prior to any data processing, the UKBB dataset consisted of 168 metabolites including lipids and lipoproteins, fatty acids, and small molecules such as amino acids and metabolites related to fluid balance, inflammation, and glycolysis (refer to Table S1 for detailed information). The HMP2 dataset contained 176 metabolites which consisted of phospholipids, amino acids and derivatives, carnitines, and more (refer to Table S2 for detailed information). While the UKBB metabolomics data was generated using nuclear magnetic resonance (NMR),[118] the HMP2 metabolomics data was generated using liquid chromatography mass spectrometry (LC-MS).[119] Among the four LC-MS methods, metabolites derived from the HILIC-pos method were specifically selected. The data generation and quantification processes of the UKBB and HMP2 metabolomics data have been detailed elsewhere.[52,120]

All pre-processing steps were performed using R (v4.3.0).[117] Imputation of the missing values, which accounted for less than 1% of the data, was employed on the UKBB datasets using "missForest" R package (v1.5).[121] The 8.69% of the HMP2 data that were missing were imputed with half of the minimum positive value for each column according to common practices.[122] As per common metabolomics procedures,[122] pareto scaling was applied to both datasets to mitigate the influence of larger features while retaining cross-feature variance. Subsequently, a log2 transformation was conducted to address heteroscedasticity from the data and rectify skewed data distribution.[123] To avoid errors in the log transformation for zero values, a pseudo-count of 1 was added to all values, since there were only a few zeros present in the UKBB dataset (< 1%) and none in the HMP2. Prior to removing the highly correlated features, features that represented sums of other cells and particle sizes were removed to enhance the efficiency of correlation analysis. The exclusion of sum-related features and particle sizes ensured that well-established features, such as sub-classes of HDL-C, were retained if possible, when found to be highly correlated with these metabolites. The highly correlated features of the remaining 156 features were then filtered out at a threshold of 0.9 based on the Spearman correlation method performed using the "caret" R package (v6.0.94).[124] The resulting pre-processed datasets, derived from both UKBB (2676 samples x 37 features) and HMP2 (494 samples x 161 features) sources, were utilized for subsequent statistical analysis and machine learning tasks.

Diet data pre-processing

The UKBB dietary data encompassed a food frequency questionnaire (FFQ) about average diet intake in the past 12 months. The UKBB codes and corresponding names of these food groups are contained in Table S3. Two types of features were observed. The values of the numerical features (fruit, vegetables, coffee, tea, water, bread, and cereal intake) were used as is. Similar to a method applied by another study in handling the FFQ data in the UKBB,[125] the frequency of the categorical features (processed and non-processed meat, and cheese intake) were assigned weights: never (0), less than once a week (0.07), once a week (0.14), two to four times a week (0.43), five to six times a week (0.79), once or more daily (1). This yielded 18 final diet features: 9 numerical and 9 weighted categorical. For the HMP2, the diet data consisted of a food questionnaire assessing food consumption frequency in the past week (Table S3). Weights were assigned as follows: no consumption (0), consumed {within the past [four to seven days (0.2), two to three days (0.56)], yesterday, one to two times (0.9), yesterday, three or more times (1)}. Refer to Figure S3 for the boxplots of food intake frequency distributions in each group for the UKBB and HMP2 cohorts.

ML classification of IBD and non-IBD

This study tested four machine learning methods, XGBoost,[126] LGBM,[127] RF,[128] and LASSO[129] in classifying disease and non-disease. XGBoost, LGBM, and RF make use of an ensemble of classification trees and combine the predictions from multiple individual decision trees to make more accurate and robust predictions, hence making them suitable for disease classification tasks. LASSO, on the other hand, is a popular regularization algorithm for logistic regression that helps reduce the feature space and highlight key associations.

All machine learning analysis carried out in python (v3.9.13)[130] was done using the "scikit-learn" module (v1.0.2).[131] The metabolomics datasets underwent an initial 80-20 train-test split. The optimization of training sets was performed using grid search ("GridSearchCV") over predefined parameter grids of the various ML models to be tested. Stratified k-fold cross-validation with 5 folds was employed within the grid search ensuring robust model assessment and hyper-parameter tuning. The optimal grid model of each classifier was the model with the

highest average validation AUC score. The best classifier of each dataset, based on the highest test AUC score metric, was passed into the SHAP explainer.

Using the same test and train split data produced from python, LASSO regularization[129] was implemented with bootstrapping and out-of-bag sample assessment of the models[132,133] in R to investigate its robustness and assess the stability of the top features. With the training set, 400 bootstraps were run with the 'glmnet' package in R, with LASSO (alpha = 1). Lambda was optimized using 10-fold cross-validation, with lambda chosen to be that which produces the highest AUC by 1 standard deviation. Evaluating the model at this lambda, the top features, which were chosen as those appearing more times than a threshold (average between the fourth and fifth quantile), and their coefficients were extracted and visualized (see Figures S1 and S5).

Running the explainable AI on the best classifier

Following training and testing, the best classifier and test set were introduced to the SHAP (SHapley Additive exPlanations) tool in Python ("shap" v0.41.0).[134] Metabolites with the highest impact were identified using SHAP global importance plots. SHAP local impact plots that illustrate the contribution of each top metabolite to sample predictions were also generated. SHAP elucidates the contribution of each feature to the model's predictions, employing concepts from cooperative game theory to quantify feature importance.[135] This approach not only enhances model interpretability but also provides insights into the decision-making process, thereby increasing transparency in the model's output.

Statistical analysis of the top metabolites

The top 20 metabolites based on SHAP-based ranking were then spearman-correlated with diet features using the "pingouin" Python package (v0.5.3).[136] The results were visualized on a circle-style heatmap generated using "matplotlib" in Python (v3.5.2).[137] Metabolites on the x-axis were clustered using the 'ward' method in the "seaborn" Python package (v0.11.2).[138] Correlation values with absolute values above 0.1 were counted as valid. Only valid correlation values were considered significant, that is, their false discovery rates (FDR) being less than 0.05. Further, to visualize how well each top feature predicted the samples, faceted boxplot distributions of the SHAP values of IBD and non-IBD samples for the top metabolites of both cohorts were generated (Figure S2). Additionally, in order to also visualize the differences in metabolite profiles between the IBD and non-IBD classes, faceted boxplot distributions of both IBD and non-IBD groups of these top metabolites were generated using R packages, with differential metabolites calculated using the Wilcoxon rank sum test (Figure S4). All boxplots were created using the "ggplot2" R package (v3.4.2), located in the "tidyverse" library,[139] and transformed into publishable-ready plots using "ggpubr" R package (v0.6.0.999).

QUANTIFICATION AND STATISTICAL ANALYSIS

In this paper, all pre-processing analysis on the metabolomics and diet data was done in R (v4.3.0), while the machine learning and explainable AI tasks were performed in Python (v3.9.13).

Significance was determined by P-values adjusted for FDRs according to the Benjamini-Hochberg principle[140] because it is less strict; values falling below the critical FDR of 0.05 were considered significant. Significance stars were displayed on plots, when applicable, according to the significance level (***: FDR < 0.001, ** $0.001 \leq$ FDR < 0.01, *: $0.01 \leq$ FDR < 0.05).

In the ML tasks, the datasets were stratified by the target class when splitting between test and train, and across folds. The performances of the models in python were represented as the mean AUC score across the folds with the 95% confidence interval, while the LASSO regression task in R was represented as mean AUC score with SD and ages of the participants represented as median and SD (Table 1). Only individuals with complete diet data were included, as well as only those that already had IBD at baseline.

Differences between the IBD and non-IBD groups were calculated using either the Wilcoxon rank sum test or the Chi-square test, depending on the nature of the data (refer to Table 1; Figure S4). Although both transformed datasets had a normal distribution, confirmed by the shapiro-wilk,[141] non-parametric tests like the Wilcoxon rank sum test and spearman-based correlation method were used due to their robustness to outliers and independence from specific distributional assumptions.