Diet and gut microbiota dominate in explaining the human serum metabolome

Noam Bar\*, Tal Korem\*, Omer Weissbrod, David Zeevi, Daphna Rothschild, Sigal Peled-Leviatan, Noa Kosower, Maya Lotan-Pompan, Adina Weinberger, Eran Segal

# Abstract

The collection of metabolites circulating in the human blood, termed the serum metabolome, contains a plethora of biomarkers and causative agents. Although the origin of specific compounds is known, we have a poor understanding of the key determinants of most metabolites. Here, we measured the levels of 1251 circulating metabolites in 521 serum samples from a healthy cohort, and devised machine learning algorithms to predict their levels in held-out subjects based on a comprehensive profile consisting of gut microbiome, clinical parameters, diet, lifestyle, anthropometric measurements and medication data. Notably, and despite out-of-sample prediction being a lower bound for explained variance, we obtained significant predictions for over 92% of the profiled metabolites, with diet and microbiome each explaining hundreds of metabolites, and with 64% of the variance of some metabolites explained using only gut microbiome data. We further uncover dense networks of interactions between diet, gut microbiome, drug intake, and lifestyle in explaining the levels of some metabolites. We demonstrate that some of these interactions are causal, as some metabolites we predicted to be positively associated with bread increased in level following a randomized clinical trial of bread intervention. Overall, our results unravel the potential determinants of over 1000 metabolites, paving the way towards mechanistic understanding of the alterations in metabolites under different conditions and to designing interventions for manipulating metabolite levels.

# Introduction

Blood serves as a liquid conveyor for molecules inside the body by delivering necessary substances to the cells and transporting metabolic waste products[1](http://f1000.com/work/citation?ids=42922&pre=&suf=&sa=0). Of particular importance are the thousands of circulating small molecules termed the serum metabolome, which are either naturally produced by our body or uptaken from the environment. While the connection of most of these metabolites to human health is yet to be elucidated, some are known to be predictive diagnostic biomarkers or even causal agents in the development of disease. For example, high blood cholesterol leads to buildup of plaque in the blood vessels, termed atherosclerosis, which in turn increases the risk for a major cardiovascular event such as heart attack, stroke, and peripheral artery disease[2](http://f1000.com/work/citation?ids=5135104&pre=&suf=&sa=0). As a result, blood cholesterol level serves as both a diagnostic biomarker and a therapeutic target for drugs such as statins[3](http://f1000.com/work/citation?ids=1799635&pre=&suf=&sa=0). As another example, type II diabetes which impacts around 10% of the population[4](http://f1000.com/work/citation?ids=6221451&pre=&suf=&sa=0), is diagnosed in part by measurements of blood glucose levels, with a recent study suggesting that a new set of metabolites significantly improves diagnosis[5](http://f1000.com/work/citation?ids=2868716&pre=&suf=&sa=0). These are only examples for the wealth of potential biomarkers and therapeutic targets that could be found in the blood, making blood an attractive source in which to search for novel biomarkers for early detection and treatment of disease.

Mass spectrometry can accurately identify thousands of metabolites from different biofluids. While some of its identified compounds are well studied and characterized, the determinants of most serum metabolites are still unknown. Studies focusing on human genetics estimated a median heritability of 6.9% for serum metabolites[6](http://f1000.com/work/citation?ids=494477&pre=&suf=&sa=0), thereby leaving much of the variation in metabolite levels unaccounted for and suggesting major contributions from environmental factors. Other studies have suggested that the gut microbiome is actively involved in the metabolism of many metabolites which are detectable in human serum[7–10](http://f1000.com/work/citation?ids=160444,5640679,3810489,5258039&pre=&pre=&pre=&pre=&suf=&suf=&suf=&suf=&sa=0,0,0,0), including a diverse set of biochemicals such as short-chain fatty acids[11,12](http://f1000.com/work/citation?ids=6134359,6221441&pre=&pre=&suf=&suf=&sa=0,0), branched-chain[13](http://f1000.com/work/citation?ids=1877576&pre=&suf=&sa=0) and aromatic[14](http://f1000.com/work/citation?ids=6133511&pre=&suf=&sa=0) amino acids. A notable example is the metabolite trimethylamine N-oxide (TMAO), which is derived from gut microbial metabolism of choline and carnitine, and was reported to act as a marker for cardiovascular disease in humans[15–17](http://f1000.com/work/citation?ids=73885,5547809,123979&pre=&pre=&pre=&suf=&suf=&suf=&sa=0,0,0), with further evidence indicating proatherogenicity[18](http://f1000.com/work/citation?ids=6134395&pre=&suf=&sa=0) and prothromboticity[19](http://f1000.com/work/citation?ids=1305084&pre=&suf=&sa=0) in mouse models. The effect of nutrition on serum metabolites was long established as dietary patterns such as the intake of red meat, whole-grain bread, tea and coffee were linked to changes in a wide range of compounds[20,21](http://f1000.com/work/citation?ids=6133597,926361&pre=&pre=&suf=&suf=&sa=0,0). Smoking was suggested as impacting serum metabolites, with some of these smoking-related changes in human serum metabolites being reversible after smoking cessation[22](http://f1000.com/work/citation?ids=6133626&pre=&suf=&sa=0). However, no study to date incorporated all of the above potential determinants within a single human cohort and quantified their relative contribution in explaining serum metabolites.

Here, we measured 1251 serum metabolites in 521 samples from healthy individuals, for whom we also obtained gut microbiome data, cardiometabolic parameters, diet, lifestyle, anthropometrics measurements, and medication data, resulting in a unique and deeply phenotyped cohort. We show that the levels of over 92% of all profiled metabolites can be significantly predicted in held-out subjects. Of the environmental factors measured, diet and gut microbiome features have the strongest predictive power, the latter independently explaining up to 64% of the variance for some metabolites. We further show that diet significantly predicts the levels of xenobiotics, while the gut microbiome significantly predicts the levels of many unknown compounds. Using feature attribution analysis, we uncover specific dietary and bacterial attributes that explain the levels of many blood metabolites, and demonstrate that the diurnal cycle might affect the levels of multiple metabolites. Finally, we provide evidence for the causal nature of some of the interactions we uncovered, by showing that the levels of some metabolites that we predicted to be positively associated with sourdough bread consumption increased following a randomized clinical trial of increased sourdough bread consumption.

Overall, our results uncover dense networks of interactions between diet, gut microbiome, drug intake, and lifestyle in explaining the levels of serum metabolites, providing a comprehensive source for the potential determinants of many circulating metabolites in human serum.

# Results

**Untargeted serum metabolomics and deep phenotyping of a large cohort**

We used mass spectrometry to profile 521 serum samples of 491 healthy individuals for whom we previously collected extensive clinical data, anthropometrics measurements, cardiometabolic parameters, medication data, lifestyle, gut microbiome, dietary logging and questionnaires[23](http://f1000.com/work/citation?ids=1005395&pre=&suf=&sa=0) (**Fig. 1A**; Methods). Our untargeted metabolomics measured the levels of 1251 metabolites, covering a wide range of biochemicals including lipids, amino acids, xenobiotics, carbohydrates, peptides, nucleotides and approximately 30% unknown compounds (**Fig. 1B**, Methods). Most measured metabolites were prevalent across the cohort, including 498 metabolites that were detected in all samples, and 1104 metabolites detected in at least 50% of the samples (**Fig. 1C**).

Demonstrating the reproducibility of our metabolomic measurements, we find that samples taken a week apart for 20 participants were significantly correlated (median Spearman R=0.68), in contrast to samples of different participants that are not correlated (median Spearman R=0.05; Methods; **Fig. 1D**). To test whether our measurements accurately report metabolite levels, we compared the metabolomic levels of creatinine and cholesterol to measurements of these compounds using standardized lab tests (Methods) performed separately on different blood samples taken from these same participants at the same visit, and here too found excellent agreement (R=0.88, creatinine; R=0.79, cholesterol, **Fig. 1E-F**). In addition to validating the reproducibility and accuracy of our data, these results further suggest that the metabolite profile of a person is a unique ‘fingerprint-like’ signature that is relatively stable at least over short time intervals.



**Figure 1 | Untargeted metabolomics and deep phenotyping of a large-scale cohort.** **(A)** Illustration of the measurements we obtained from our cohort. **(B)** Breakdown of the 1251 identified metabolites by pathways. **(C)** Number of samples (y-axis) in which each metabolite (x-axis) was identified, for a sorting of metabolites by their prevalence. **(D)** Spearman correlations (y-axis) between standardized metabolomic profiles (Methods) of different individuals (n=475; median Spearman 0.05) stratified by sex, and between standardized metabolomic profiles of the same participant (n=20; median Spearman 0.68) taken one week apart. **(E-F)** Mass-spectrometry measurements (y-axis) vs. standardized lab tests results (x-axis; Methods) for creatinine (E; Pearson R=0.88, p<10-20) and cholesterol (F; R=0.79, p<10-20). C&V, Cofactors and vitamins; a.u. Arbitrary units.

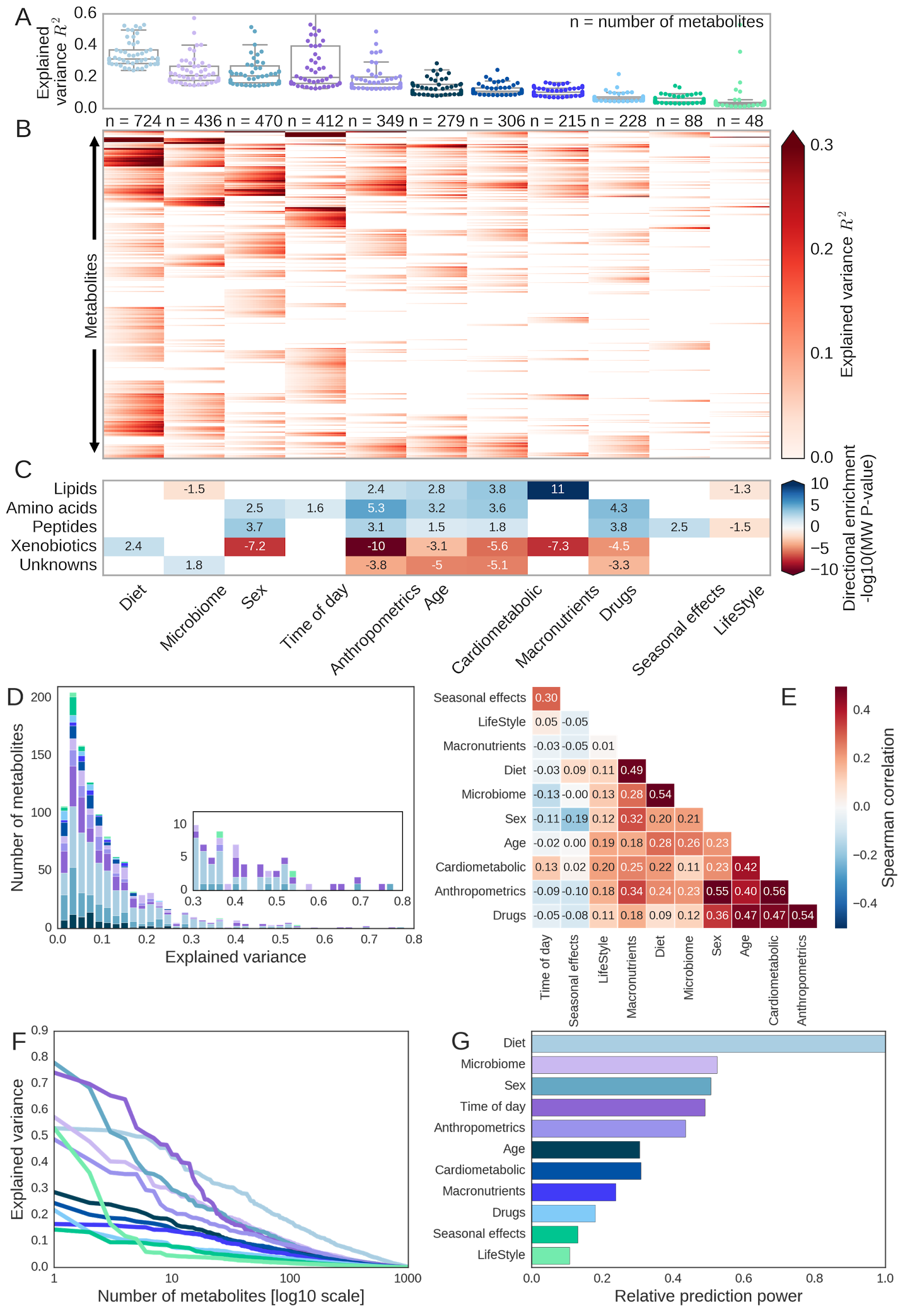
**Diet, microbiome, and clinical data** **provide statistically significant predictions of most serum metabolites**

To estimate the extent to which each metabolite can be predicted by the wealth of data we collected, we devised machine learning algorithms that predict the level of each metabolite in held-out subjects (out-of-sample prediction). We used gradient boosting decision trees[24](http://f1000.com/work/citation?ids=6157367&pre=&suf=&sa=0) (Methods) as these models can capture nonlinear interactions which are likely to be present in such a heterogeneous feature space and within the high dimensionality of the diet and microbiome data. Notably, over 92% of the metabolites are significantly predicted (FDR<10%; Methods) using at least one of the feature groups, with diet significantly explaining the largest number of metabolites (724), and gut microbiome explaining 436 metabolites (**Fig. 2A,B**). Together, our predictive models explained over 10% of the variance for 435 metabolites (**Fig. 2D**), and in some cases our models explained over 60% of the variance, as in the case of gut microbiome features explaining 64% of the variance of X-16124 (**Fig 2A**).

To understand whether specific feature groups better predict particular metabolite group, for each feature group we sorted its predictions across all metabolites and asked whether better predictions are enriched for any metabolite group (**Fig. 2C**; Methods). We found that age, sex, cardiometabolic parameters and anthropometric measurements better predicted blood lipids, amino acids and peptides but not xenobiotics and unknown compounds (**Fig. 2C**). As another example, daily average consumption of macronutrients (lipids, proteins, carbohydrates), calories and water better predicted blood lipids (p<10-11) but not xenobiotics (p<10-7). In contrast, other dietary features obtained from daily logging of food types and food frequency questionnaires (FFQ; Methods) better predicted the levels of xenobiotics (p<0.004) and were in fact the only feature group to do so, suggesting that information about specific foods is required in order to explain the levels of multiple molecules, and xenobiotics in particular. Finally, gut microbiome data predominantly explained unknown compounds (p<0.02), highlighting the potential of the microbiome for discovering microbiome-derived metabolites and explaining the origin of the large number of unknown compounds.

We next asked whether different feature groups predict similar metabolites, by computing the correlation between the metabolite predictions of every pair of input feature groups (**Fig. 2E**). We found that predictions based on anthropometrics were highly correlated with those of sex (Spearman R=0.55) and cardiometabolic parameters (R=0.56), suggesting that much of the information captured by anthropometrics is shared with these other feature groups, while in comparison to the lower yet significant correlation between predictions made by sex and cardiometabolic parameters (R=0.23), it implies that each capture unique information about metabolites. Notably, predictions based on gut microbiome data had the highest correlation to predictions based on diet (R=0.54), showing that this two feature groups were most similar in terms of predicting the levels of metabolite groups. Intriguingly, predictions based on age showed a significant correlation with those based on sex (R=0.22) although age and sex were not correlated in this cohort (p>0.1, t-test), implying that many metabolites distribute differently according to both age and sex. These correlations point to possible interactions between these feature groups in explaining the levels of many serum metabolites, an aspect that we further explore below.

Taken together, our results show that we can produce statistically significant predictions for most serum metabolites using diet, gut microbiome, or the other lifestyle and clinical parameters that we measured, with each feature group having several metabolites for which it drives most of the predictive power. Summing over the explained variance (EV) of all metabolites for each feature group, we found that diet had the biggest predictive power, while lifestyle parameters were the least predictive, carrying only 10% of the predictive power of diet (**Fig. 2G**). Notably, gut microbiome data ranked second, having 52% of the predictive power of diet, suggesting that the gut microbiota may be an important determinant of the levels of many serum metabolites.

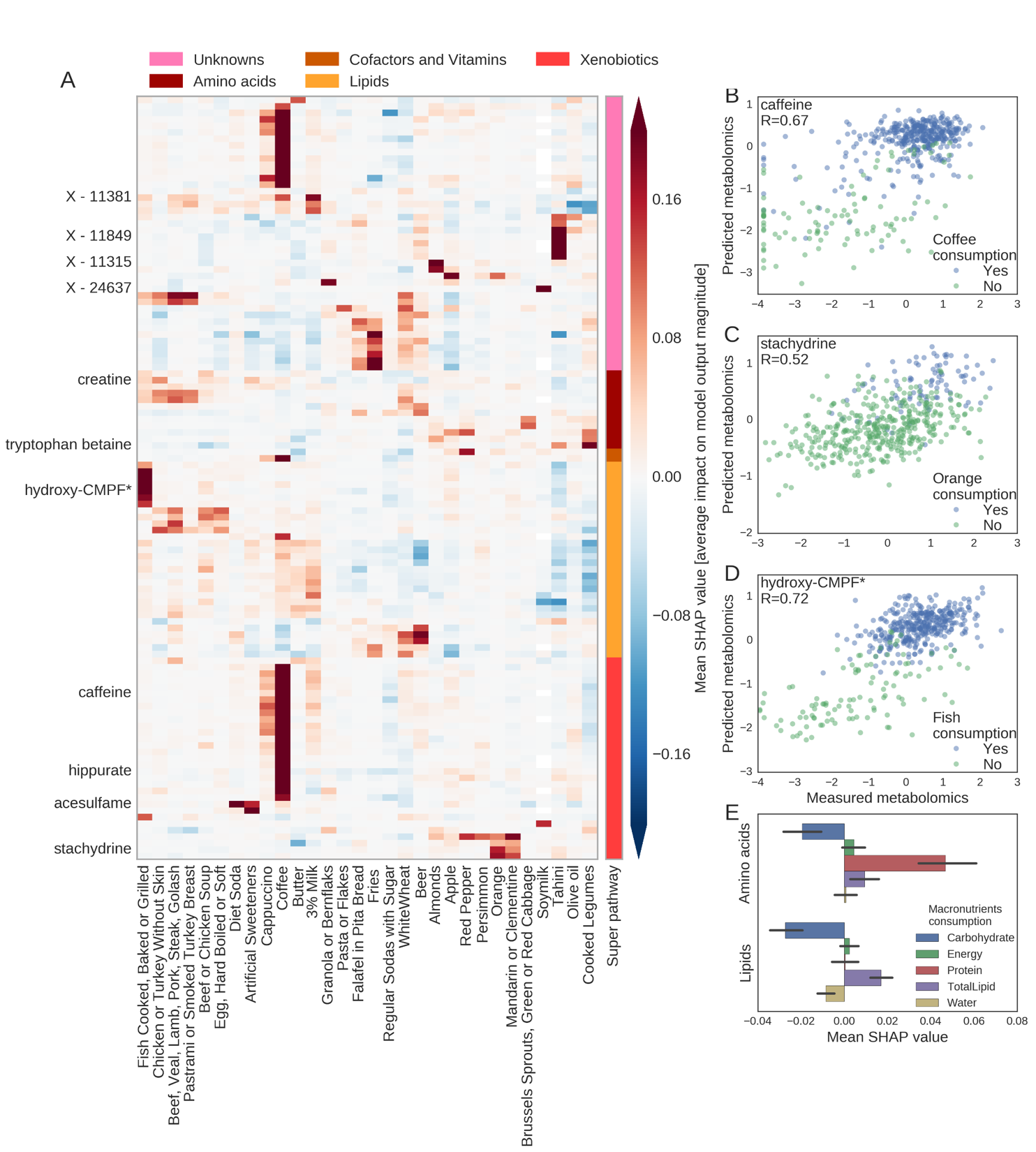


**Figure 2 | Diet, gut microbiome and clinical data predict the levels of most serum metabolites.** Figure panels show results of 10-fold cross validation predictions of the levels of every metabolite derived separately for each feature group. **(A)** Box plots (box, IQR; whiskers, 1.5\*IQR) showing the explained variance (R2 ; EV) of the top 50 predicted metabolites of each feature group, with the feature groups sorted by their median EV across these 50 metabolites. **(B)** Heatmap showing the 95% confidence interval (CI) for EV (color gradient from left to right corresponds to lower and higher CI bounds) predicted for each metabolite (y-axis) by every feature group (a-axis). Only metabolites with significant predictions (p<0.05 with 10% FDR) are shown. The number of metabolites significantly predicted by each feature group is indicated above each column. P-values and CIs were estimated using bootstrapping (Methods). **(C)** Enrichment of pathway categories in the metabolites predicted by each feature group (Mann-Whitney *U* test; Methods). Enrichment values are shown in -log10 P-value multiplied by the sign of the test statistic. Positive values indicate that metabolites from some pathway are better predicted (higher R2) than metabolites from all other pathway categories by the corresponding feature group, while negative values indicate that metabolites from the pathway category are significantly worse predicted using the corresponding feature group. Only significant enrichments (p<0.05 after 10% FDR correction are shown). **(D)** A stacked histogram of the number of metabolites (y-axis) with any value of EV. For each metabolite the EV is taken as the maximum EV across all feature groups, and its color is according to the feature group that best predicted it. Inset shows the metabolites with EV in the range 0.3-0.8. **(E)** Spearman correlation matrix computed between the held-out metabolite prediction P-values of every pair of feature groups. The matrix is clustered using a hierarchical clustering on euclidean distances of the Spearman correlation. **(F)** For each feature group, shown is the number of metabolites (x-axis) that it predicts with an EV above a certain value (y-axis), for all possible EV values. **(G)** The fraction of total EV, which is the sum of the EV of all metabolites (y-axis), for each feature group (x-axis), relative to the total EV of the most predictive feature group (diet, top).

**Diet explains a large portion of the serum metabolome**

As diet had the largest predictive power over the serum metabolome, we next sought to interpret the diet-based models and ask which dietary features drive the predictions of each metabolite. Our diet data consists of both food frequency questionnaires and one week of dietary logging via an App we devised[23](http://f1000.com/work/citation?ids=1005395&pre=&suf=&sa=0), and thus allows us to address the predictive power of both long term and short term nutritional patterns. In order to explain the output of our machine learning models and find specific associations between features and metabolite levels we used SHAP (SHapley Additive exPlanations)[25](http://f1000.com/work/citation?ids=5534754&pre=&suf=&sa=0), a feature attribution analysis tool which assigns each feature an importance value (SHAP value) for a particular prediction[26](http://f1000.com/work/citation?ids=5535395&pre=&suf=&sa=0) (Methods). We found multiple diet features that were strongly predictive of blood metabolites in our models (**Fig. 3A**). Notably, the reported consumption of coffee (both long- and short-term) had the most predictive power over a large number of xenobiotics and unknown compounds (p<0.001). As previously reported[27](http://f1000.com/work/citation?ids=6174860&pre=&suf=&sa=0), metabolites from the xanthine metabolism pathway such as paraxanthine (R=0.64, p<10-20) and caffeine (R=0.67, p<10-20, **Fig. 3B**) were significantly predicted using coffee consumption (p<10-10). Another strong predictor was the reported consumption of fish, which accurately predicted the levels of several blood lipids such as 3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid (CMPF) (R=0.72, p<10-20, **Fig. 3D**), a potent uremic toxin known to accumulate in the serum of CKD patients[28](http://f1000.com/work/citation?ids=6174931&pre=&suf=&sa=0) and which was also suggested to prevent and reverses steatosis[29](http://f1000.com/work/citation?ids=6174932&pre=&suf=&sa=0). Other examples included saccharin (R=0.58, p<10-20) and acesulfame (R=0.47, p<10-20), two artificial sweeteners which were mainly predicted by the reported consumption of artificial sweeteners and diet soda. Lactose (R=0.3, p<10-7) was also significantly predicted by the reported consumption of 3% milk and coffee.

Next we investigated another global aspect of diet, by assessing the predictive power that our models ascribes to daily average consumption of macronutrients (lipids, proteins and carbohydrates), calories and water based on the nutritional loggings. We found that high levels amino acid metabolites were significantly predicted by consumption of proteins (p<10-10), while levels of blood lipids were mainly predicted by consumption of dietary lipids (p<10-5, **Fig. 3E**). The predicted power of macronutrients was significantly lower than that of the full dietary profiling even when comparing to the predictive power of only the detailed types and amounts of food that are the part of that sum, (**Extended Data Fig. 1**), suggestive of the fact that accurately predicting the levels of many metabolites (especially xenobiotics) requires the detailed food components.



**Figure 3 | Diet explain a large portion of the serum metabolome.** **(A)** Heatmap showing the directional mean absolute SHAP values (Methods) of various features (x-axis) computed from 10-fold cross validation models that predict metabolite levels (y-axis) using only diet data. Positive values indicate positive correlation between a dietary feature and the levels of a metabolite, while negative values indicate negative correlation between the two. Metabolites are sorted by their super pathways (colored bar) and clustered within each group. Shown are the top 120 predicted metabolites using diet, and the top 30 dietary features by maximum mean absolute SHAP value. **(B-D)** Predicted (y-axis) vs. measured (x-axis) levels of of caffeine (B; Pearson R=0.67, p<10-20), stachydrine (C; R=0.52, p<10-20) and hydroxy-CMPF (D; R=0.72, p<10-20) based only on dietary data. Dots are colored by the reported consumption of items having the highest mean absolute SHAP value for each metabolite. P-values estimated via bootstrapping (Methods). **(E)** The directional mean absolute SHAP value (x-axis) of different macronutrients (y-axis; bar, average; error bars, std), categorized by lipids and amino acids.

**Gut microbiome explains the levels of a wide range of metabolites**

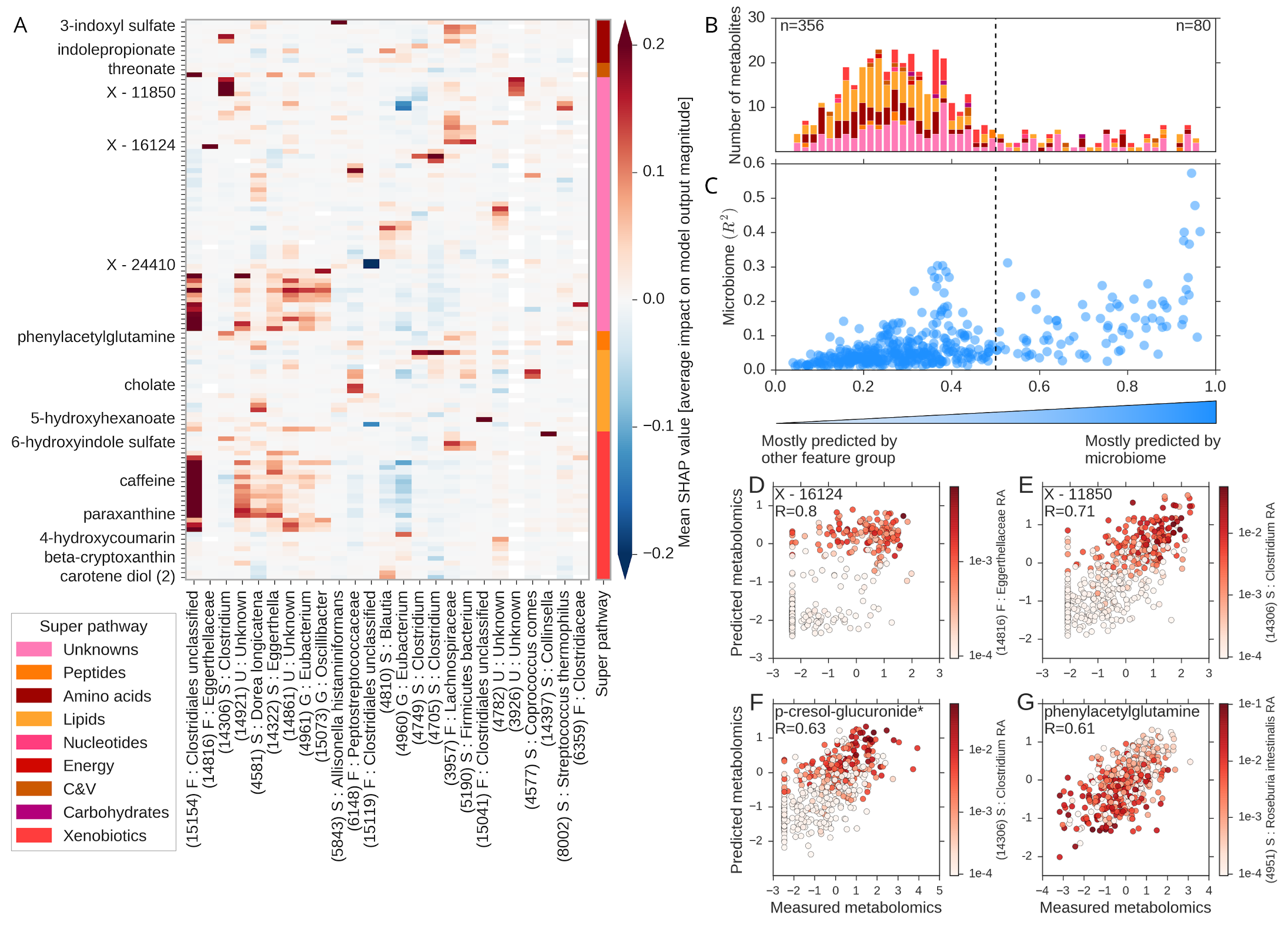
As with the dietary data, we used SHAP to unravel the interactions that our models learned and utilized for its predictions between metabolites and different bacterial taxa from the metagenomic gut microbiome data (**Fig. 4A**). Notably, the microbiome data on its own predicted the levels of many metabolites with high accuracy, as in the case of X-16124 (Pearson R=0.8, p<10-20; **Fig. 4D**), an unknown metabolite whose prediction is mainly driven by the relative abundance of a bacteria from the *Eggerthellaceae* family, and X-11850 (R=0.71, p<10-20; **Fig. 4E**), another unknown compound whose predictions are mainly driven by a species of *Clostridium*. The microbiome data was also highly predictive of two uremic toxins (phenylacetylglutamine, R=0.61, p<10-20, **Fig. 4G,** and indoxyl sulfate, R=0.36, p<10-8) previously reported in association with CKD[30](http://f1000.com/work/citation?ids=6175475&pre=&suf=&sa=0) and several other comorbidities[31,32](http://f1000.com/work/citation?ids=4328555,6151107&pre=&pre=&suf=&suf=&sa=0,0), and these predictions were negatively driven by *Roseburia intestinalis* and positively by another bacteria from the *Lachnospiraceae* family.

As a more global view, we next asked whether the predictive power of the microbiome depends on a few bacteria, or whether the abundance of multiple bacterial features are needed for accurate predictions. To this end, we defined the bacterial taxa with the maximal mean absolute SHAP value for each metabolite as the main driver of the predictions of that metabolite. We found that 24 bacterial features were the main drivers for the top 50 predicted metabolites (R>0.4, p<10-18; **Supplementary Table 1**), and one bacterial feature from the *Clostridiales* family (15154: leftmost column in **Fig. 4A**) was the main driver of 18 metabolites, including 12 xenobiotics mostly from the xanthine metabolism pathway (e.g., caffeine, R=0.48, p<10-20, and paraxanthine, R=0.49, p<10-20). These 18 metabolites were also significantly associated with the consumption of coffee. Another species of *Clostridium* (14306) was the main driver of 5 metabolites, including 3 unknown compounds, phenylacetylcarnitine (R=0.45, p<10-20) and p-cresol-glucuronide (R=0.63, p<10-20; **Fig. 4F**) which was previously reported to be metabolised by *Clostridium*[*33*](http://f1000.com/work/citation?ids=6081401&pre=&suf=&sa=0). Furthermore, 5 bacterial features were the main drivers of 2 metabolites each, and each of the other 17 bacterial features was a main driver of a single metabolite. These results suggest that in most cases many specific bacteria are required in order to accurately predict the levels of distinct metabolites, but in some cases a single bacteria might underlie the predictions of a broad metabolic pathway involving dozens of metabolites.

In terms of higher bacterial taxonomy levels, among the bacterial features that best predicting the top 120 metabolites, 106 were *Firmicutes*, 6 were *Actinobacteria* and 8 belonged to an unknown phyla, showing the domination of *Firmicutes* in explaining the levels of many metabolites. Furthermore, as *Bacteroidetes* is the second most abundant phyla in our cohort (**Extended Data Fig. 2**), it is surprising that I-urobilinogen (R=0.23, p<10-5) which is the first metabolite to be best explained by a bacteria of this phyla is only the 157-th best predicted metabolite using microbiome data.

As some metabolites could be significantly predicted by more than one feature group, we next studied those metabolites that were best explained by gut microbiome data. For each of the 436 metabolite groups which were significantly predicted using gut microbiome (**Fig. 2B**) we computed a score between 0 and 1, representing the fraction of variance that the microbiome data model explains out of that explained by the sum of the microbiome model and the next best model from the feature groups except microbiome (**Fig. 4C**). For 80 of the 436 microbiome predicted metabolite groups, the score was higher than 0.5, indicating that microbiome had the highest predictive power among all feature groups tested (**Fig. 4B**; **Supplementary Table 2**).

Using enrichment analysis of the scores by pathway categories we revealed that the scores of unknown compounds (Mann-Whitney *U* p<0.009) and xenobiotics (p<0.003) were higher, while those of lipids (p<0.005) and amino acids (p<0.004) were significantly lower. The highest scoring metabolite was X-12261 (R2=0.42, p<10-20) which scored 0.96, as the next best model explained only 1% of its variance. Other notable examples are taurodeoxycholate (0.86; R2=0.096, p<10-6), glycoursodeoxycholate (0.85; R2=0.14, p<10-12) and isoursodeoxycholate (0.926; R2=0.2, p<10-16), three secondary bile acids, which are molecules shown to be formed by bacterial 7α-dehydroxylation of the primary bile acids[34](http://f1000.com/work/citation?ids=955941&pre=&suf=&sa=0),[35](http://f1000.com/work/citation?ids=974995&pre=&suf=&sa=0). As many metabolites were exclusively predicted using gut microbiome, these results highlight the potential of the gut microbiome in producing many metabolites.



**Figure 4 | Gut microbiome predicts the levels of a wide range of metabolites. (A)** Same as **Fig. 3A**for the directional mean absolute SHAP values of top 25 bacterial features (by maximum mean absolute SHAP values) and the top 120 predicted metabolites using only microbiome data. X-axis ticks represent the ID of the bacterial feature followed by the lowest taxonomy level identified. F, Family; G, Genus; S, Species; U, Unknown; **(B)** Histogram of scores per metabolite computed as the explained variance (EV) using only microbiome features divided by the sum of the EV using microbiome features and the EV of the most predictive feature group excluding microbiome. Score ranges between 0 (EV of microbiome is zero) and 1 (EV of second best predicting group is zero). Metabolites for which the score is higher than 0.5 are best predicted using only microbiome data. Numbers of metabolites in each half of the plot are indicated in the top corners. Shown are the 436 metabolites which are significantly predicted using only microbiome data (FDR=0.1, **Fig. 2B**). Stacked colors correspond to super pathways of the metabolites. **(C)** Dot plot showing the EV (R2, y-axis) of metabolites using microbiome vs the score (x-axis) described in (B). **(D-G)** Predicted (y-axis) vs measured (x-axis) levels of X-16124 (D; Pearson R=0.8, p<10-20), X-11850 (E; R=0.71, p<10-20), p-cresol-glucuronide (F; R=0.63, p<10-20) and phenylacetylglutamine (G; R=0.61, p<10-20) based only on microbiome data, and colored by the relative abundance of the bacterial taxa having the highest mean absolute SHAP value for each metabolite. P-values for prediction were estimated via bootstrapping.

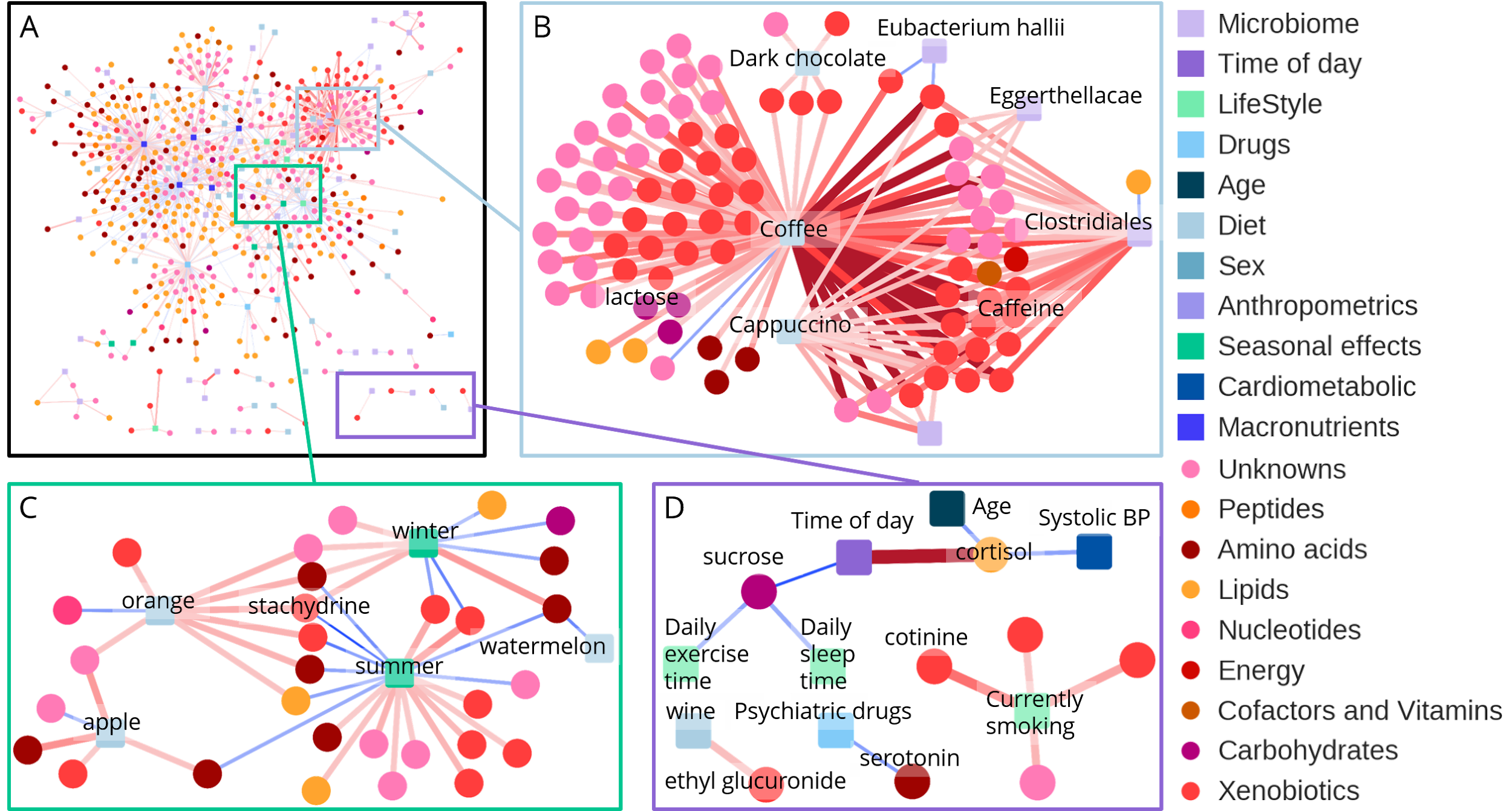
**Networks of interactions between features explain diverse metabolites**

As multiple metabolites were significantly predicted using more than one feature group, we next examined how different feature groups interact in explaining the levels of these metabolites. By building separate predictive models each based on a different feature group and using SHAP in order to estimate the impact of each specific feature on the output of the models, we uncovered a dense network of interactions between feature groups in explaining the levels of metabolites (**Fig. 5A**; Methods).

As briefly mentioned above, we found that the reported consumption of coffee was linked to a large number of metabolites, most of which are unknown compounds and xenobiotics from the xanthine metabolism pathway. Notably, we found that a specific bacterial feature from the *Clostridiales* family was linked to a large number of these metabolites (**Fig. 5B**), pointing at an interaction between coffee consumption and the presence of this bacteria in explaining the levels of these metabolites. Being the highest impacting features among their feature categories, coffee consumption and this *Clostridiales* species may be targets for validation using interventional studies.

We next focused on metabolites which were significantly explained using seasonal effects, and examined which dietary features interact with them (**Fig. 5C**). The consumption of citrus fruits such as oranges was positively associated with several metabolites such as stachydrine (Pearson R=0.52, p<10-20), which in turn had higher values in samples taken in winter months compared to samples taken during the summer, consistent with the fact that oranges are seasonal fruits available in Israel mostly during winter. Another example is N-methyltaurine (R=0.3, p<10-8), an amino acid which has higher levels in samples taken during winter, and is also negatively driven by the consumption of watermelon, a summer seasonal fruit.

Finally, we explored some known examples of associations between metabolites and features to further validate the quality of data in our cohort (**Fig. 5D**). The diurnal cycle is known to regulate the levels of multiple circulating metabolites. We found that the levels of cortisol (R=0.63, p<10-20) were significantly lower in samples taken during the second half of the day, consistent with previous studies showing that cortisol levels peak early in the morning[36](http://f1000.com/work/citation?ids=5264118&pre=&suf=&sa=0). The levels of sucrose were lower in samples taken during the first half of the day and were also lower in participants who reported higher rates of daily exercise and longer sleeping times. We also found that the levels of tobacco-related metabolites such as cotinine (R=0.72, p<10-20) were higher in samples of active smokers, and that no other feature could significantly explain their levels. Finally, we found that blood levels of serotonin (R=0.46, p<10-6) were lower in samples of participants who reported to take psychiatric drugs, despite serotonin being a therapeutic target for selective serotonin reuptake inhibitors (SSRI)[37](http://f1000.com/work/citation?ids=6210849&pre=&suf=&sa=0) which are prescribed in order to increase the levels of serotonin in the brain.



**Figure 5 | Networks of interactions between phenotypes explain diverse metabolites.** Interactions between features from different feature groups predictive of similar metabolites are presented in a graphical layout, in which nodes are either metabolites or features, and edges are the directional mean absolute SHAP values computed from models trained only on features from the respective feature group. Circular nodes - metabolites; predictive feature nodes - squares; both colored by relevant categories. Shown are only edges with a mean absolute SHAP value greater than 0.12. **(A)** Network of associations between metabolites and features. Shown are the following feature groups: macronutrients, diet, microbiome, lifestyle, drugs and seasonal effects. **(B)** A large group of metabolites which are mainly driven by the reported consumption of coffee, and the relative abundance of a bacteria from the *Clostridiales* family. **(C)** Metabolites explained by seasonal fruit consumption. **(D)** Examples of associations and interactions between metabolites and features in predictive models.

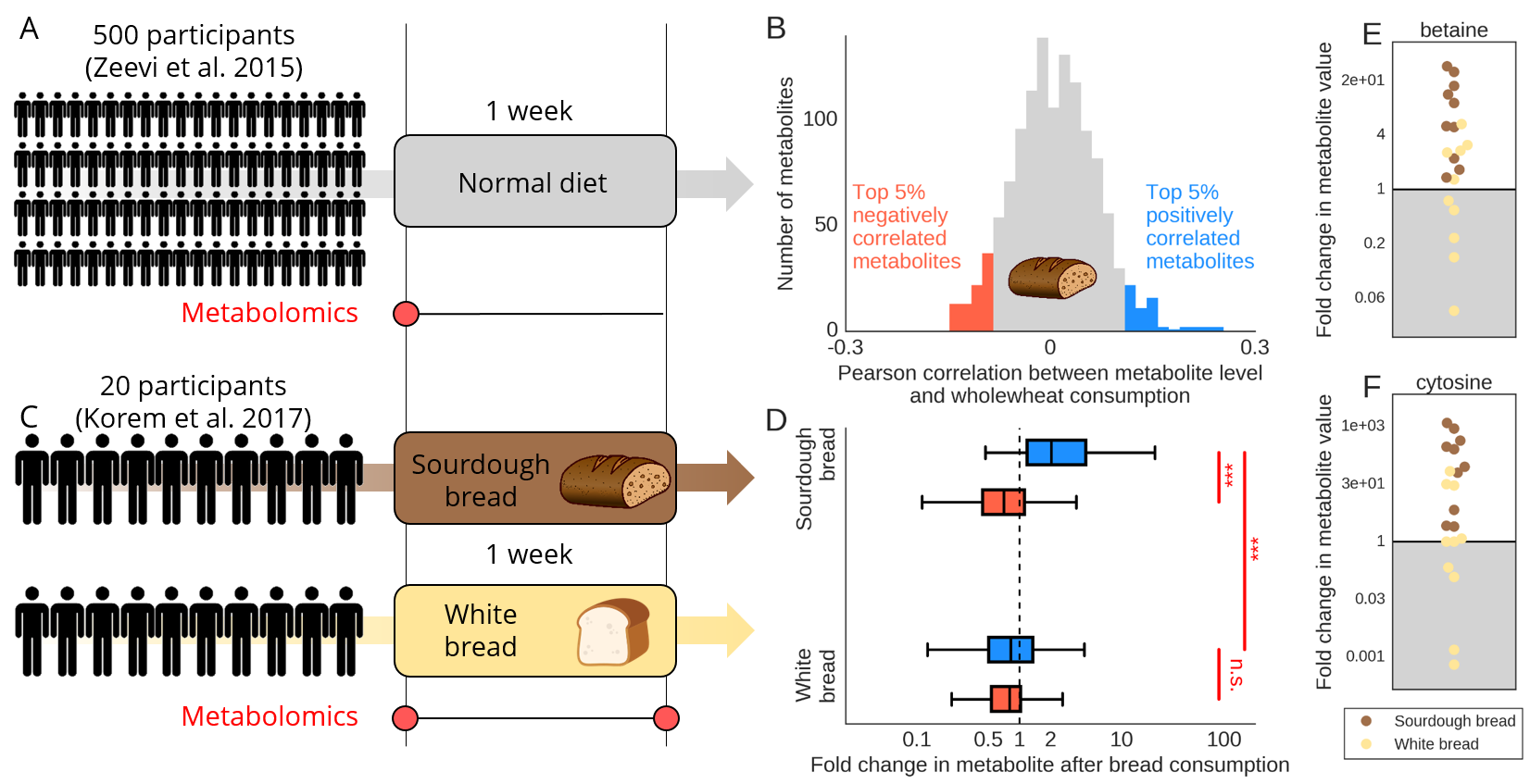
**Metabolites explained by bread increase following a bread consumption intervention**

As proof of concept to examine whether some of the feature-metabolite interactions that we uncovered may be causal, we profiled the serum metabolome of samples from a randomized cross-over trial that we previously conducted[38](http://f1000.com/work/citation?ids=3778778&pre=&suf=&sa=0), in which we compared the effects of consuming artisanal whole-grain sourdough bread (hereinafter, “sourdough bread”) to those of industrial white bread made from refined wheat (“white bread”). 20 healthy subjects were randomly divided to two groups of 10, which then underwent a 1-week-long dietary intervention of increased bread consumption, where each group received a different type of bread. Following two weeks of washout, the intervention was performed again, switching bread types between the groups. (**Fig. 6C**; Methods). In the present study, we performed metabolomic profiling of blood samples that were taken at both the beginning and the end of the first week of intervention, in order to estimate the effect of the dietary intervention on serum metabolites.

We used the healthy cohort of 458 participants for which we had one week of logged normal diet, without any intervention (**Fig. 6A**) in order to identify potential associations between the reported consumption of white and whole-grain wheat and the levels of metabolites (**Fig. 6B**; Methods). We ranked the metabolites according to their correlation with the consumption of both breads in the 458 participants, and selected the top 5% positively and negatively correlated metabolites for further analysis (**Fig. 6B**). Notably, analyzing the metabolomic samples of subjects who received the sourdough bread intervention, we found that metabolites that were positively correlated with the consumption of whole-grain wheat in our cohort increased significantly more (median fold-change 2.04) than metabolites that were negatively correlated with the consumption of whole-grain wheat in the 458-participants cohort (median fold-change 0.7, p<10-12, Mann-Whitney *U*; **Fig. 6D**; Methods). Moreover, we found no statistically significant differences when comparing the mean fold-change of these metabolites in the group which received the white bread intervention (p>0.2, Mann-Whitney *U*; **Fig. 6D**).

Some of the metabolites which increased in levels following the sourdough bread intervention were previously reported to be linked to the consumption of whole-grain wheat flour. A notable example is betaine, an amino acid which has been shown to protect internal organs, improve vascular risk factors[39](http://f1000.com/work/citation?ids=3054440&pre=&suf=&sa=0) and is also known to be highly abundant in a wide variety of foods, where wheat bran and wheat germ are the highest sources of naturally occurring betaine[40,41](http://f1000.com/work/citation?ids=2005315,6222128&pre=&pre=&suf=&suf=&sa=0,0). We found that in the group that received sourdough bread the mean fold-change in betaine levels was 6.16, while the mean fold-change in the group that received white bread was 0.82 (Mann-Whitney *U* p<0.004; **Fig. 6E**; Methods), consistent with the correlation between betaine levels and the consumption of whole-grain wheat in the larger cohort (Spearman R=0.14, p<0.003). Another example is cytosine, for which the mean fold-change was far greater at 78.5 in the sourdough bread group in comparison to a mean fold-change of 0.53 in the white bread group (Mann-Whitney *U* p<0.002; **Fig. 6F**). Unlike betaine, the levels of cytosine were not previously linked to the rate or type of bread consumption.

We also performed a similar analysis using metabolites that were correlated with white wheat consumption in our cohort, but did not find significant changes in these metabolites in the bread intervention study. Overall, these results suggest that some of the associations that we found between the consumption of whole-grain wheat and the levels of metabolites in our larger cohort might be causal, as their levels increase following a dietary intervention that increased the consumption of whole-wheat bread.



**Figure 6 | Metabolites explained by bread increase following an intervention that increases bread consumption. (A)** Measuring associations between dietary features and metabolite levels using samples from this study. **(B)** Histogram of Pearson correlations between metabolite levels and the reported consumption of whole-grain wheat across our cohort. The top 5% (n=62; blue) positively correlated metabolites and the top 5% (n=62; red) negatively correlated metabolites are marked and used for further analysis. **(C)** A randomized controlled trial[42](http://f1000.com/work/citation?ids=5180192&pre=&suf=&sa=0) with 20 healthy subjects comparing the effect of consuming traditionally milled and prepared whole-grain sourdough bread to that of consuming industrial white bread made from refined wheat[38](http://f1000.com/work/citation?ids=3778778&pre=&suf=&sa=0). We analyzed sample for the first week of the trial, in which 10 subjects increased consumption of sourdough bread and 10 others increase consumption of white bread. **(D)** Box plots (box, IQR; whiskers, 1.5\*IQR) showing the mean fold-change (FC) of the top 5% positively (blue) and negatively (red) correlated metabolites, separated by intervention group. Among the group which recieved the sourdough bread intervention the mean FC of the top 5% positively correlated metabolites was significantly higher than the mean FC of the top 5% negatively correlated metabolites (p<10-12, Mann-Whitney *U*). \*\*\* Mann-Whitney *U* p<0.001; n.s. Not significant. **(E-F)** FC (y-axis) of two metabolites separated by intervention groups. In the sourdough bread group the FC of both betaine (E; Mann-Whitney *U* p<0.004) and cytosine (F; Mann-Whitney *U* p<0.002) were higher compared to the same FC in the group having white bread.

# Methods

All statistical and machine learning analyses were performed using Python (version 2.7.8).

**Description of cohorts**

In this study we used a previously described cohort of 491 Israeli individuals[23](http://f1000.com/work/citation?ids=1005395&pre=&suf=&sa=0). Study participants were healthy individuals aged between 18 and 70 (for full inclusion and exclusion criteria, see previous description[23](http://f1000.com/work/citation?ids=1005395&pre=&suf=&sa=0)). Before the study, participants answered medical, lifestyle and nutritional questionnaires. During a period of one week, participants were instructed to record all daily activities, including standardized and real-life meals, in real-time using their smartphones.

**Metadata**

The diet data included detailed FFQs aimed at capturing long term dietary habits, and the daily mean consumption of foods computed based on the one-week-long recordings. Regarding both the logging of foods and the FFQs, we kept only items which were reported to be consumed at least once by at least 5% of our participants, resulting with 257 different food from logging, and 137 different item in the questionnaires. We also used the recordings in order to compute the daily mean consumption of macronutrients (lipids, proteins, carbohydrates), calories and water. Anthropometrics includes weight, BMI, waist and hips cire circumference, and waist to hips ratio (WHR). Height was excluded from the anthropometrics as it is a strong proxy for sex. Cardiometabolic parameters include systolic and diastolic blood pressure, heart rate in beats per minute and a glycaemic status as previously described[43](http://f1000.com/work/citation?ids=4908266&pre=&suf=&sa=0). Lifestyle parameters includes smoking status (current, past) from questionnaires, the daily mean sleeping time, exercise time and midday sleep time based on the one-week-long recordings. Medication data was taken from questionnaires. As the time of day we created a binary feature indicating whether the sample was taken during the first half of the day. We considered the season feature as the month in which the sample was taken, and in some analyses we also grouped months by season (Winter: December - February; Spring: March - May; Summer: June - August; Fall: September - November).

**Metabolomics profiling**

Metabolite concentrations were measured from serum samples by Metabolon, Inc., Durham, North Carolina, USA, by using an untargeted LC/MS platform as previously described[6,44,45](http://f1000.com/work/citation?ids=5667141,494477,5350347&pre=&pre=&pre=&suf=&suf=&suf=&sa=0,0,0).

**Metabolomics preprocessing**

A total of 540 serum samples were ran in 15 batches of 36 samples each. In 11 batches we included one pooled sample as control (technical replicate), while in each of the other 4 batches we had two such samples. The other 521 serum samples belonged to 491 participants. In total, 1251 different metabolites were identified in the 521 serum samples. 498 metabolites were measured in all 521 samples, while 1104 metabolites where measured in at least 50% of samples.

We used two separate normalization schemes, one for single metabolites which we further used in feature attribution analysis, and the second for metabolite groups which we used for global and enrichment analyses.

In the first scheme, metabolites for which we had less than 10 measurements were excluded from further analysis, resulting with 1224 metabolites. The normalization steps of the metabolomics data included a robust standardization (subtracting the median and dividing by the standard deviation) over the log10 transformed metabolite levels, followed by clipping outlier samples which were farther than 5 standard deviations. Next we regressed out the number of days passed from the first sample taken, and finally we imputed missing values as the minimum value per metabolite, as these represent samples for which the compound’s concentration was lower than the detection threshold.

In the second scheme, in order to handle possible bias resulting from uncertainty of metabolite assignment and a high rate of extremely correlated mass spectrometry peaks we grouped the metabolites by correlation to form 1116 metabolite groups (Spearman correlation >0.85), 1025 of which are singletons. Prior to grouping, we performed a robust standardization (subtracting the median and dividing by the standard deviation) over the log10 transformed metabolite levels, followed by clipping outlier samples which were farther than 5 standard deviations. For each metabolite group we considered the mean of the metabolites in the group for each participant. After grouping, we regressed out the number of days passed from the first sample taken (only for metabolite groups with over than 50 non missing samples), and finally we imputed missing values as the minimum value per metabolomic group. The category of each metabolite group was assigned based on majority vote, where unknown compounds were excluded from the vote, unless all metabolites in the group were unknown.

**Microbiome preprocessing**

Sample collection, DNA extraction, and sequencing of the samples in this study has been described previously[23,38,43](http://f1000.com/work/citation?ids=4908266,1005395,3778778&pre=&pre=&pre=&suf=&suf=&suf=&sa=0,0,0). Briefly, we used only samples which were collected using swab, and filtered metagenomic reads containing Illumina adapters, filtered low-quality reads and trimmed low-quality read edges. We detected host DNA by mapping with GEM[46](http://f1000.com/work/citation?ids=396578&pre=&suf=&sa=0) to the human genome (hg19) with inclusive parameters, and removed human reads. We subsampled all samples to have at most 10 million reads.

Bacterial relative abundance estimation was performed by mapping bacterial reads to SGB representative genomes (**Supplementary Table 3**)REF. We selected all SGB representatives with at least 5 genomes in group, and for these representatives genomes kept only unique regions as a reference data set. Mapping was performed using bowtie2[47](http://f1000.com/work/citation?ids=48791&pre=&suf=&sa=0) and abundance was estimated by calculating the mean coverage of unique genome regions across the 50 percent most dense covered areas as previously described[48](http://f1000.com/work/citation?ids=1074184&pre=&suf=&sa=0). Feature names include the lowest taxonomy level identified.

**Comparing metabolomics to lab tests**

We compared the levels of both creatinine and cholesterol which we previously obtained via standard lab tests[23](http://f1000.com/work/citation?ids=1005395&pre=&suf=&sa=0) with their metabolomic levels. Since the lab tests were performed by two different labs, we centered (by reducing the mean) both lab tests by the lab in which the test was performed, and then standardize their levels. The metabolomic profiling and the lab tests were performed on separate blood samples which were taken at the same day.

**Correlation of metabolic profiles within and between individuals**

We compared the Spearman correlations between standardized metabolomic profiles of the same participant taken one week apart (n=20), to correlations between standardized metabolomic profiles of different individuals (n=475). Each pair of samples taken from the same participant was ran in the same metabolomic batch. In the group of different individuals, only pairs of individuals from the same batch were included (resulting in a total of 3835 such pairs), and were further stratified by sex.

**Predictive models of metabolite groups**

We used LightGBM (version 2.1.2)[24](http://f1000.com/work/citation?ids=6157367&pre=&suf=&sa=0), a gradient boosting decision trees model in order to predict the levels of 1116 metabolite groups based on 11 feature groups in held-out subjects (**Fig. 2**). In order to estimate the explained variance of each metabolite group we ran a 10-fold cross validation model using each feature group as input, and then evaluated the results over the full set of samples using Pearson correlation. For all prediction results we computed 95% confidence intervals and p-values via 1000 iterations of bootstrapping[49](http://f1000.com/work/citation?ids=324259&pre=&suf=&sa=0). Standard error of correlations were estimated from bootstrapping and p-values were then estimated via the normal CDF using the Wald test[50](http://f1000.com/work/citation?ids=6239496&pre=&suf=&sa=0). We corrected p-values of all predictions (1116\*11) for multiple hypotheses using the false discovery rate (FDR) procedure[51](http://f1000.com/work/citation?ids=6239504&pre=&suf=&sa=0), and used a cutoff of FDR<0.1. In all cross validation and bootstrapping runs we used a fixed and predetermined set of hyperparameters (**Supplementary Table 4**).

For each feature group we performed an enrichment analysis of the pathway categories over the levels of metabolite groups. For each category we used a rank sum test (Mann-Whitney *U*) comparing the prediction accuracy of metabolites from that category compared to prediction accuracy of metabolites from other categories. P-values were log10 transformed and multiplied by the sign of the test statistic, hence positive values indicate that metabolites from some pathway are better predicted than metabolites from all other pathway categories by the corresponding feature group, while negative values indicate that metabolites from the pathway category are significantly depleted in the predictions of the corresponding feature group. We considered only metabolite groups for which at least one feature group had a significant prediction (after correcting for multiple hypothesis), resulting with 1030 metabolite groups.

In predicting metabolite groups, microbiome features includes the relative abundance of SGB representative genomes as described above, as well as relative abundance of bacterial species as computed by MetaPhlAn2[52](http://f1000.com/work/citation?ids=465803&pre=&suf=&sa=0), and the first 10 PCs computed over the log transformed relative abundance of a bacterial gene catalog[53](http://f1000.com/work/citation?ids=433565&pre=&suf=&sa=0) as previously described[43](http://f1000.com/work/citation?ids=4908266&pre=&suf=&sa=0).

**Feature attribution analysis**

In order to explain the output of our machine learning models and point at specific associations between features and metabolite levels we used SHAP (SHapley Additive exPlanations)[25](http://f1000.com/work/citation?ids=5534754&pre=&suf=&sa=0), a feature attribution analysis tool which assigns each feature an importance value (SHAP value) for a particular prediction. Individual SHAP values were computed for held-out subjects in 10-fold cross validation using the module TreeExplainer (version 0.24.0)[26,54](http://f1000.com/work/citation?ids=5535395,6243802&pre=&pre=&suf=&suf=&sa=0,0), based on models trained only on features from the respective feature group. In each fold we ran a random hyperparameter search consistent of 20 iterations using the module RandomizedSearchCV from sklearn (version 0.18.2)[55](http://f1000.com/work/citation?ids=6243884&pre=&suf=&sa=0), and chose the best model for predicting the held out subjects and computing the SHAP values. In all feature attribution analyses we used the ungrouped list of 1224 metabolites.

For every feature, we computed the mean absolute SHAP value across all instances in a specific model, this value reflects the mean impact of each feature on the predictions and may serve as a feature importance measure. We further used these values in order to compute a directional mean absolute SHAP values, by multiplying them with the sign of the Spearman correlation between the population feature and the target. Here, positive values indicate positive correlation between a feature and the levels of a metabolite, while negative values indicate negative correlation between the two.

In order for SHAP values of features computed from different models (different target/metabolite) to be comparable, we standardized the levels of metabolites, as SHAP values are measured in the same units as the target.

When performing feature attribution analysis with microbiome as input, we only included the relative abundance of SGB representative genomes as features, taking only features which were present in over 5% of the samples.

**Interaction networks**

We used a graphical layout in order to visualise the associations of features with the levels of metabolites. The nodes are either metabolites or features, and the edges are the directional mean absolute SHAP values computed from models trained only on features from the respective feature group as described above. All networks were constructed using Cytoscape[56](http://f1000.com/work/citation?ids=121985&pre=&suf=&sa=0).

**Analysis of bread intervention**

We previously conducted a randomized cross-over trial[38](http://f1000.com/work/citation?ids=3778778&pre=&suf=&sa=0) in 20 healthy subjects, in which we compared the effect of consumption of artisanal whole-grain sourdough bread to that of industrial white bread made from refined wheat. The 20 participants were randomly divided into two groups of 10 and they each received two 1-week-long dietary interventions of each type of bread but in different order. In this study, we performed metabolomic profiling of blood samples that were taken at both the beginning and ending of the first week of intervention, in order to estimate the effect of the dietary intervention on serum metabolites.

In order to find the associations between metabolite levels and the consumption of both types of bread in the study cohort we computed a Spearman correlation between the levels of each metabolite and the reported consumption of wholemeal wheat and white wheat. We used the top 5% positively and negatively correlated metabolites for further analysis. The correlations were computed using 458 samples of distinct individuals, a subset of our cohort from which we excluded all samples of individuals which participated in the intervention study.

For each metabolite in every individual, we computed the fold-change of metabolite levels between the samples taken at the end of the first week of intervention and the start of that week. Prior to computing fold-changes we imputed missing values with the minimum per metabolite and standardized their log transformed levels. Furthermore, for each intervention group, we computed the mean fold-change of every metabolite based on the 10 samples from that group. Then we compared the mean fold-changes of the top 5% positively and negatively correlated metabolites mentioned above within each intervention group by performing a rank sum test (Mann-Whitney *U*) over the mean fold-changes.

For comparing the fold-changes of betaine and cytosine between the two intervention groups, we used a rank sum test (Mann-Whitney *U* test).

# References

Bibliography

[1. Psychogios, N. *et al.* The human serum metabolome. *PLoS ONE* **6,** e16957 (2011).](http://f1000.com/work/bibliography/42922)

[2. Ridker, P. M., Stampfer, M. J. & Rifai, N. Novel risk factors for systemic atherosclerosis. *JAMA* **285,** 2481 (2001).](http://f1000.com/work/bibliography/5135104)

[3. Baigent, C. *et al.* Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366,** 1267–1278 (2005).](http://f1000.com/work/bibliography/1799635)

[4. National Diabetes Statistics Report | Data & Statistics | Diabetes | CDC. at <https://www.cdc.gov/diabetes/data/statistics/statistics-report.html>](http://f1000.com/work/bibliography/6221451)

[5. Floegel, A. *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62,** 639–648 (2013).](http://f1000.com/work/bibliography/2868716)

[6. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46,** 543–550 (2014).](http://f1000.com/work/bibliography/494477)

[7. Wikoff, W. R. *et al.* Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc Natl Acad Sci USA* **106,** 3698–3703 (2009).](http://f1000.com/work/bibliography/160444)

[8. Fischbach, M. A. Microbiome: focus on causation and mechanism. *Cell* **174,** 785–790 (2018).](http://f1000.com/work/bibliography/5640679)

[9. Liu, R. *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat. Med.* **23,** 859–868 (2017).](http://f1000.com/work/bibliography/3810489)

[10. Fujisaka, S. *et al.* Diet, genetics, and the gut microbiome drive dynamic changes in plasma metabolites. *Cell Rep.* **22,** 3072–3086 (2018).](http://f1000.com/work/bibliography/5258039)

[11. Wilson, M. *Microbial Inhabitants of Humans: Their ecology and role in health and disease*. (Cambridge University Press, 2004). doi:10.1017/CBO9780511735080](http://f1000.com/work/bibliography/6134359)

[12. Topping, D. L. Short-chain fatty acids produced by intestinal bacteria. *Asia Pac. J. Clin. Nutr.* **5,** 15–19 (1996).](http://f1000.com/work/bibliography/6221441)

[13. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535,** 376–381 (2016).](http://f1000.com/work/bibliography/1877576)

[14. Patel, K. P., Luo, F. J.-G., Plummer, N. S., Hostetter, T. H. & Meyer, T. W. The production of p-cresol sulfate and indoxyl sulfate in vegetarians versus omnivores. *Clin. J. Am. Soc. Nephrol.* **7,** 982–988 (2012).](http://f1000.com/work/bibliography/6133511)

[15. Tang, W. H. W. *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368,** 1575–1584 (2013).](http://f1000.com/work/bibliography/73885)

[16. Li, X. S. *et al.* Gut microbiota-dependent trimethylamine N-oxide in acute coronary syndromes: a prognostic marker for incident cardiovascular events beyond traditional risk factors. *Eur. Heart J.* **38,** 814–824 (2017).](http://f1000.com/work/bibliography/5547809)

[17. Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19,** 576–585 (2013).](http://f1000.com/work/bibliography/123979)

[18. Brown, J. M. & Hazen, S. L. Metaorganismal nutrient metabolism as a basis of cardiovascular disease. *Curr. Opin. Lipidol.* **25,** 48–53 (2014).](http://f1000.com/work/bibliography/6134395)

[19. Zhu, W. *et al.* Gut microbial metabolite TMAO enhances platelet hyperreactivity and thrombosis risk. *Cell* **165,** 111–124 (2016).](http://f1000.com/work/bibliography/1305084)

[20. Floegel, A. *et al.* Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. *Eur. J. Clin. Nutr.* **67,** 1100–1108 (2013).](http://f1000.com/work/bibliography/6133597)

[21. Thorburn, A. N., Macia, L. & Mackay, C. R. Diet, metabolites, and “western-lifestyle” inflammatory diseases. *Immunity* **40,** 833–842 (2014).](http://f1000.com/work/bibliography/926361)

[22. Xu, T. *et al.* Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Med.* **11,** 60 (2013).](http://f1000.com/work/bibliography/6133626)

[23. Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163,** 1079–1094 (2015).](http://f1000.com/work/bibliography/1005395)

[24. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. (2017).](http://f1000.com/work/bibliography/6157367)

[25. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* (2017).](http://f1000.com/work/bibliography/5534754)

[26. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* (2018).](http://f1000.com/work/bibliography/5535395)

[27. Ashihara, H., Monteiro, A. M., Gillies, F. M. & Crozier, A. Biosynthesis of caffeine in leaves of coffee. *Plant Physiol.* **111,** 747–753 (1996).](http://f1000.com/work/bibliography/6174860)

[28. Tsutsumi, Y. *et al.* Renal disposition of a furan dicarboxylic acid and other uremic toxins in the rat. *J. Pharmacol. Exp. Ther.* **303,** 880–887 (2002).](http://f1000.com/work/bibliography/6174931)

[29. Prentice, K. J. *et al.* CMPF, a Metabolite Formed Upon Prescription Omega-3-Acid Ethyl Ester Supplementation, Prevents and Reverses Steatosis. *EBioMedicine* **27,** 200–213 (2018).](http://f1000.com/work/bibliography/6174932)

[30. Hung, S.-C., Kuo, K.-L., Wu, C.-C. & Tarng, D.-C. Indoxyl sulfate: A novel cardiovascular risk factor in chronic kidney disease. *J. Am. Heart Assoc.* **6,** (2017).](http://f1000.com/work/bibliography/6175475)

[31. Barrios, C. *et al.* Gut-Microbiota-Metabolite Axis in Early Renal Function Decline. *PLoS ONE* **10,** e0134311 (2015).](http://f1000.com/work/bibliography/4328555)

[32. Poesen, R. *et al.* Microbiota-Derived Phenylacetylglutamine Associates with Overall Mortality and Cardiovascular Disease in Patients with CKD. *J. Am. Soc. Nephrol.* **27,** 3479–3487 (2016).](http://f1000.com/work/bibliography/6151107)

[33. Evenepoel, P., Meijers, B. K. I., Bammens, B. R. M. & Verbeke, K. Uremic toxins originating from colonic microbial metabolism. *Kidney Int. Suppl.* S12-9 (2009). doi:10.1038/ki.2009.402](http://f1000.com/work/bibliography/6081401)

[34. Kakiyama, G. *et al.* Modulation of the fecal bile acid profile by gut microbiota in cirrhosis. *J. Hepatol.* **58,** 949–955 (2013).](http://f1000.com/work/bibliography/955941)

[35. Ridlon, J. M., Kang, D. J., Hylemon, P. B. & Bajaj, J. S. Bile acids and the gut microbiome. *Curr. Opin. Gastroenterol.* **30,** 332–338 (2014).](http://f1000.com/work/bibliography/974995)

[36. Smyth, J. M. *et al.* Individual differences in the diurnal cycle of cortisol. *Psychoneuroendocrinology* **22,** 89–105 (1997).](http://f1000.com/work/bibliography/5264118)

[37. Hyttel, J. Pharmacological characterization of selective serotonin reuptake inhibitors (SSRIs). *Int. Clin. Psychopharmacol.* **9,** 19–26 (1994).](http://f1000.com/work/bibliography/6210849)

[38. Korem, T. *et al.* Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses. *Cell Metab.* **25,** 1243–1253.e5 (2017).](http://f1000.com/work/bibliography/3778778)

[39. Olthof, M. R., van Vliet, T., Boelsma, E. & Verhoef, P. Low dose betaine supplementation leads to immediate and long term lowering of plasma homocysteine in healthy men and women. *J. Nutr.* **133,** 4135–4138 (2003).](http://f1000.com/work/bibliography/3054440)

[40. Craig, S. A. S. Betaine in human nutrition. *Am. J. Clin. Nutr.* **80,** 539–549 (2004).](http://f1000.com/work/bibliography/2005315)

[41. Fardet, A. *et al.* Whole-grain and refined wheat flours show distinct metabolic profiles in rats as assessed by a 1H NMR-based metabonomic approach. *J. Nutr.* **137,** 923–929 (2007).](http://f1000.com/work/bibliography/6222128)

[42. Chalmers, T. C. *et al.* A method for assessing the quality of a randomized control trial. *Control. Clin. Trials* **2,** 31–49 (1981).](http://f1000.com/work/bibliography/5180192)

[43. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555,** 210–215 (2018).](http://f1000.com/work/bibliography/4908266)

[44. Bridgewater BR, E. A. High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *Metabolomics* **04,** (2014).](http://f1000.com/work/bibliography/5667141)

[45. Zierer, J. *et al.* The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **50,** 790–795 (2018).](http://f1000.com/work/bibliography/5350347)

[46. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9,** 1185–1188 (2012).](http://f1000.com/work/bibliography/396578)

[47. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).](http://f1000.com/work/bibliography/48791)

[48. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349,** 1101–1106 (2015).](http://f1000.com/work/bibliography/1074184)

[49. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Chapman and Hall/CRC, 1994). doi:10.1007/978-1-4899-4541-9](http://f1000.com/work/bibliography/324259)

[50. Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Statist.* **16,** 117–186 (1945).](http://f1000.com/work/bibliography/6239496)

[51. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. at <https://www.jstor.org/stable/2346101#metadata\_info\_tab\_contents>](http://f1000.com/work/bibliography/6239504)

[52. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–814 (2012).](http://f1000.com/work/bibliography/465803)

[53. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).](http://f1000.com/work/bibliography/433565)

[54. GitHub - slundberg/shap: A unified approach to explain the output of any machine learning model. at <https://github.com/slundberg/shap>](http://f1000.com/work/bibliography/6243802)

[55. Release history — scikit-learn 0.18.2 documentation. at <https://scikit-learn.org/0.18/whats\_new.html>](http://f1000.com/work/bibliography/6243884)

[56. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).](http://f1000.com/work/bibliography/121985)