

Multi-omic Characterization of the Taxa-function Relationship in Infant Gut Microbiomes

Quang P Nguyen

Dartmouth College <https://orcid.org/0000-0002-2072-3279>

Margaret R. Karagas

Dartmouth College

Juliette C. Madan

Dartmouth College Geisel School of Medicine

Erika Dade

Dartmouth College

Tom J. Palys

Dartmouth College

Hilary G. Morrison

Marine Biological Laboratory

Wimal W. Pathmasiri

UNC-Chapel Hill: University of North Carolina at Chapel Hill

Susan McRitche

UNC-Chapel Hill: University of North Carolina at Chapel Hill

Susan J. Sumner

UNC-Chapel Hill: University of North Carolina at Chapel Hill

H. Robert Frost

Dartmouth College

Anne G. Hoen (✉ Anne.G.Hoen@dartmouth.edu)

Dartmouth College

Research

Keywords: Infant gut microbiome, stool metabolome, prediction models, functional redundancy, metabolism

Posted Date: September 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-79308/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 Multi-omic characterization of the taxa-function relationship in infant gut microbiomes

2 Quang P. Nguyen^{1,2}, Margaret R. Karagas^{1,3}, Juliette C. Madan^{1,2,3,4}, Erika Dade¹, Tom J. Palys¹, Hilary G.
3 Morrison⁵, Wimal W. Pathmasiri^{6*}, Susan McRitche⁶, Susan J. Sumner⁶, H. Robert Frost^{2**} & Anne G. Hoen^{1,2,3**}

4 ¹Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA.

5 ²Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA

6 ³Children's Environmental Health & Disease Prevention Research Center at Dartmouth, Lebanon, NH, USA.

7 ⁴Division of Neonatology, Department of Pediatrics, Children's Hospital at Dartmouth, Hanover, NH, USA.

8 ⁵Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA.

9 ⁶Department of Nutrition, Nutrition Research Institute, University of North Carolina at Chapel Hill, Chapel Hill,
10 NC, USA

11 Corresponding author: anne.g.hoen@dartmouth.edu

12 *Corresponding author for metabolomic assays: wimal_pathmasiri@unc.edu

13 **Jointly supervised this research

14 **Email Addresses:**

15 QPN: Quang.P.Nguyen.GR@dartmouth.edu

16 MRK: Margaret.R.Karagas@dartmouth.edu

17 JCM: Juliette.Madan@dartmouth.edu

18 ED: Erika.F.Dade@dartmouth.edu

19 TJP: Thomas.J.Palys@dartmouth.edu

20 HGM: morrison@mbl.edu

21 WWP: wimal_pathmasiri@unc.edu

22 SLM: susan_mcritchie@unc.edu

23 SJS: susan_sumner@unc.edu

24 HRF: hildreth.r.frost@dartmouth.edu

25 AGH: Anne.G.Hoen@dartmouth.edu

27 **Abstract**

28 **Background:** The infant intestinal microbiome plays an important role in metabolism and immune
29 development with impacts on lifelong health. The linkage between the taxonomic composition of the
30 microbiome and its metabolic phenotype is undefined and complicated by redundancies in the taxon-function
31 relationships within microbial communities. To inform a more mechanistic understanding of the relationship
32 between the microbiome and health, we performed an integrative statistical and machine learning-based
33 analysis of microbe taxonomic structure and metabolic function in order to characterize the taxa-function
34 relationship in early life.

35 **Results:** Stool samples collected from infants enrolled in the New Hampshire Birth Cohort Study (NHBCS) at
36 approximately 6-weeks (n = 168) and 12-months (n = 282) of age were profiled using targeted and untargeted
37 nuclear magnetic resonance (NMR) spectroscopy as well as DNA sequencing of the V4-V5 hypervariable
38 region from the bacterial 16S rRNA gene. There was significant inter-omic concordance based on Procrustes
39 analysis of sample distances (6 weeks: p = 0.056; 12 months: p = 0.001), however this association was no
40 longer significant when accounting for phylogenetic relationships using generalized UniFrac distance metric (6
41 weeks: p = 0.376; 12 months: p = 0.069). Sparse canonical correlation analysis showed significant correlation,
42 as well as identifying sets of microbe/metabolites driving microbiome-metabolome relatedness. Performance of
43 machine learning models varied across different metabolites, with support vector machines (radial basis
44 function kernel) being the consistently top ranked model. However, predictive R² values demonstrated poor
45 predictive performance across all models assessed (avg. R²: -5.06% -- 6 weeks; -3.7% -- 12 months).

46 Conversely, the Spearman correlation metric was higher (avg. correlation: 0.344 – 6 weeks; 0.265 – 12
47 months). This demonstrated that taxonomic relative abundance was not predictive of metabolite
48 concentrations.

49 **Conclusions:** Our results suggest a degree of overall association between taxonomic profiles and metabolite
50 concentrations. However, lack of predictive capacity for stool metabolic signatures reflects, in part, the possible
51 role of functional redundancy in defining the taxa-function relationship in early life as well as the bidirectional
52 nature of the microbiome-metabolome association. Our results provide evidence in favor of a multi-omic
53 approach for microbiome studies, especially those focused on health outcomes.

54 **Keywords:** Infant gut microbiome, stool metabolome, prediction models, functional redundancy, metabolism

55 **Background**

56 The human gut microbiome is a complex and diverse system of microorganisms that co-inhabit the intestinal
57 lumen and play a crucial role in modulating human health and disease [1, 2]. The development of the
58 microbiota in early life is a sensitive process akin to primary ecological succession [3], and therefore both
59 reliant on and vulnerable to external perturbations. Studies have linked microbiome alterations to long-term
60 health consequences, including risk of obesity [4], type I diabetes [5], and inflammatory bowel disease [6]. As
61 such, there is a need to understand how the microbiome participates in the multifactorial pathways leading to
62 long-term disease outcomes. One key to this open question lies in the currently undefined relationship between
63 the taxonomic structure of the microbiome and its metabolic phenotype. Previous studies addressing this
64 question have mainly focused on DNA-based profiling of microbial functional potential, which, due to
65 complicated regulatory mechanisms at the cellular level beyond the genome, is not equivalent to the
66 microbiota's realized functional landscape [7].

67 There exists a bidirectional association between the metabolome and the microbiome in the gut [8, 9]. These
68 low molecular weight molecules can either be nutrients that shape the composition of the microbiome [10] or
69 important byproducts of host-microbe nutrient co-metabolism that help regulate host metabolic homeostasis

70 [11–13]. For example, members of the *Clostridium* clusters can produce and increase serum levels of
71 branched chain amino acids, which are markers for insulin resistance and diabetes [14, 15]. However, studies
72 suggest that the fecal metabolome specifically can be used as a readout of gut microbe metabolic functions.
73 Zierer et al. [16] showed in a large cohort of adult females (n = 786) from the TwinsUK study that around 60%
74 of the fecal metabolome is associated with microbial composition, where on average, 67% of variance in the
75 metabolome can be explained by the microbiome.

76 Recent studies have integrated the metabolome in microbiome analyses of health outcomes, most notably
77 Lloyd et al. [17] from the integrative Human Microbiome Project. However, these studies have mostly focused
78 on adult populations with specific metabolic disease etiologies such as inflammatory bowel disease. Only a
79 limited number of studies [18–23] have simultaneously profiled the gut microbiome and metabolome from
80 infant stool samples. These studies have preliminarily established that metabolomic profiles shift as taxonomic
81 abundances change between subject case/control status [18, 20, 21, 24]. Specifically, Ayeni et al. (n = 48) [19]
82 and Kisuse et al. (n = 35) [23] demonstrated that inter-sample distances calculated using metabolite
83 abundances are correlated with those calculated from taxonomic profiles using Mantel tests across African and
84 Asian cohorts. However, studies to date have either focused on preterm infants [18, 20, 21] or had small
85 sample sizes (less than 50) [19, 22, 23]. We identified a major gap in defining microbiome-metabolome
86 relatedness among infants from a population-based cohort capturing both early in infancy and near the first
87 birthday, with regards to determining the strength of association and to identify key contributors to the overall
88 concordance.

89 Here, we present our study investigating associations between microbe abundances assayed with 16S rRNA
90 sequencing and metabolomic profiles measured with ¹H NMR spectroscopy in a cohort of infants representing
91 a rural general population from the New Hampshire Birth Cohort Study (NHBCS). This is a unique
92 epidemiological cohort with multi-omic profiling of infant stool samples at multiple time points accompanied with
93 rich sociodemographic, dietary and health outcomes data [25]. Our study utilizes predictive modeling,

94 multivariate correlation methods and distance-based approaches to characterize the dynamic relationship
95 between the gut microbiome and the gut metabolome in early life.

96 **Results**

97 The overall workflow and subject selection process are described in Figure 1. Primary analyses were
98 performed on paired microbiome-metabolome data sets on samples collected at approximately 6 weeks (N =
99 158) and 12 months (N = 282) of age (65 subjects have paired samples collected at both time points). In order
100 to take advantage of the most samples from this study, each time point was analyzed separately with
101 sensitivity analyses performed on sample pairs. After processing and filtering, we evaluated a final taxonomic
102 dataset of 48 genera in 6 weeks samples and 72 genera in 12 months samples. Metabolomic profiles were
103 represented as 208 unique untargeted spectral bins and a concentration-fitting method [26] was used to
104 acquire specific relative concentrations of 36 targeted metabolites. These metabolites were chosen based on a
105 literature search (Table S1) for compounds known to be associated with commensal gut microbes. Results
106 presented here will primarily feature the targeted dataset, with accompanying figures and tables for the
107 untargeted data set in the supplemental section. Figure 1 shows the overall workflow including the sample
108 selection process.

109 **Study population**

110 Study subject characteristics are summarized in Table 1 separately for both subjects providing samples at 6-
111 week of age (n = 158) and 12-months of age (n = 282). Characteristic of our population, most infants are White
112 (97.5% among subjects contributing a 6-week sample; 95.4% among subjects contributing a 12-month
113 sample), delivered vaginally (6 weeks samples: 72.2%; 12 months samples: 70.9%) and did not have any
114 systemic antibiotic exposure during initial hospitalization following birth (6 weeks samples: 95.6%; 12 months
115 samples 97.2%). The average birth weight was also similar across subjects irrespective of the sample time
116 point, 3370 g (\pm 480) for infants contributing 6-week samples and 3430 g (\pm 528) for infants contributing 12-
117 month samples. Similarly, the average gestational age was 39.1 weeks (\pm 1.59) (6-week samples) and 39
118 weeks (\pm 1.7) (12-month samples). At the time of 6-week sample collection, 62% of infants had been

exclusively breastfed while by the time of 12-month sample collection, 59.2% of infants received breast milk supplemented with formula, however a large minority (35.1%) remained exclusively breastfed. There were more male than female infants in the cohort (53.8% male among infants contributing a 6-week sample; 56.4% male among infants contributing a 12-month sample). Maternal smoking during pregnancy was rare (6-week samples: 7%; 12-month samples: 5%).

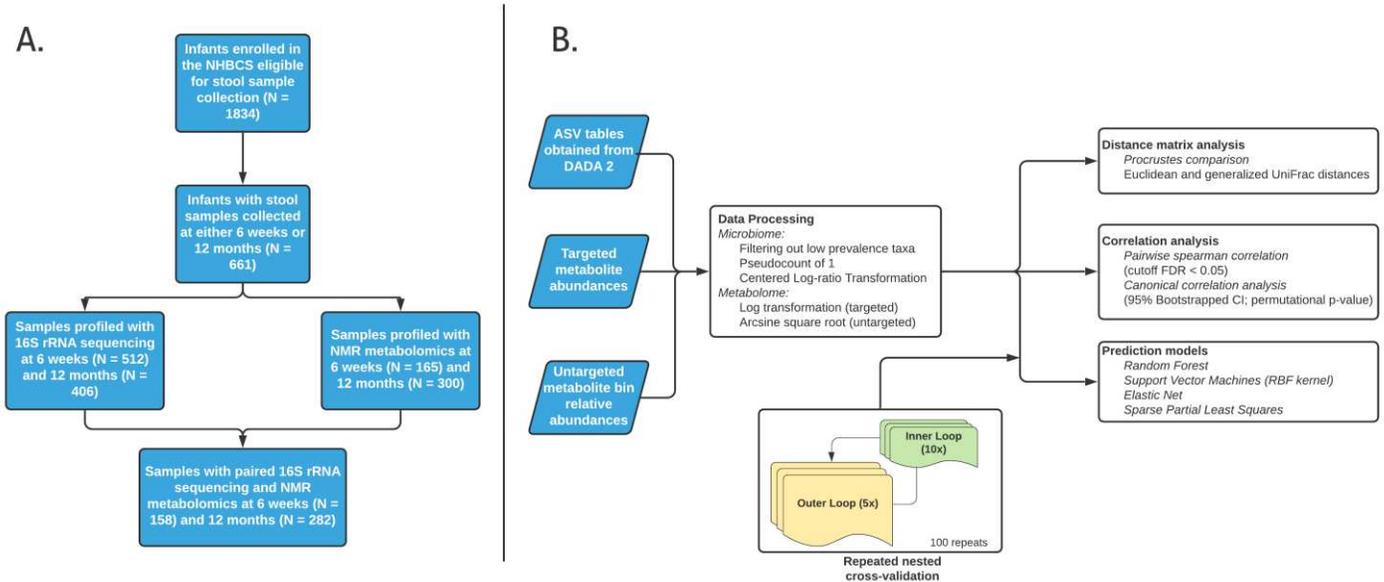


Figure 1. Overview of the analysis. Panel A describes the subject selection workflow and panel B describes the analytic pipeline.

Table 1. Selected characteristics of subjects contributing samples at 6 weeks (n = 158) and at 12 months of age (n = 282).

	6 weeks (n = 158)	12 months (n = 282)
Birthweight (grams)		
Mean (Standard Deviation)	3370 (480)	3430 (528)
Median [Minimum, Maximum]	3430 [1910, 4710]	3450 [1320, 4660]

Missing	2 (1.3%)	4 (1.4%)
Sex		
Male	85 (53.8%)	159 (56.4%)
Female	73 (46.2%)	123 (43.6%)
Feeding Mode Until Time of Sample Collection		
Unknown	6 (3.8%)	7 (2.5%)
Exclusively breastfed	98 (62%)	99 (35.1%)
Exclusively formula fed	13 (8.2%)	9 (3.2%)
Mixed	41 (25.9%)	167 (59.2%)
Delivery Mode		
Vaginal	114 (72.2%)	200 (70.9%)
Cesarean	44 (27.8%)	82 (29.1%)
Gestational Age (Weeks)		
Mean (SD)	39.1 (1.59)	39.0 (1.70)
Median [Minimum, Maximum]	39.1 [33.4, 43.0]	39.1 [29.1, 42.0]
Post-delivery infant systemic antibiotic exposure		
No	151 (95.6%)	274 (97.2%)
Yes	7 (4.4%)	8 (2.8%)
Maternal smoking during pregnancy		
No	143 (90.5%)	262 (92.9%)
Yes	11 (7.0%)	14 (5.0%)
Missing	4 (2.5%)	6 (2.1%)
Infant Race		

Other	4 (2.5%)	13 (4.6%)
White	154 (97.5%)	269 (95.4%)

129

130

Inter-omic sample distance comparison suggests overall concordance between data sets

131

Global concordance between the microbiome and the metabolome was observed across both time points and

132

metabolomic data sets (Figure 2A, Figure S1A) when analyzed using a symmetric Procrustes test with

133

samples ordinated by Euclidean distances (Methods). It is noted that the p-value at 6 weeks for the targeted

134

data set ($p = 0.057$) was only trending close to significant at the 0.05 level.

135

Since the Procrustes test was performed on principal coordinate (PCoA) ordinations of sample distances, the

136

result is sensitive to the choice of dissimilarity metric. In addition to standard Euclidean distances, the

137

generalized UniFrac (gUniFrac) metric was also leveraged to account for phylogeny in calculating differences

138

between samples. With gUniFrac, the association was not significant at either time points for the targeted data

139

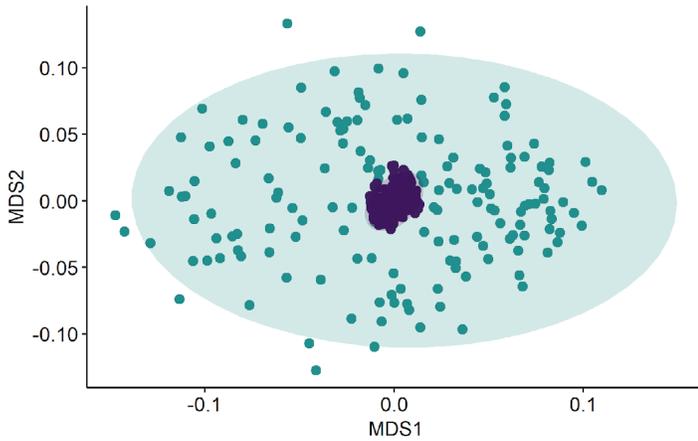
set only (Figure 2B), while the untargeted data set still maintained strong concordance (6 weeks samples – $p =$

140

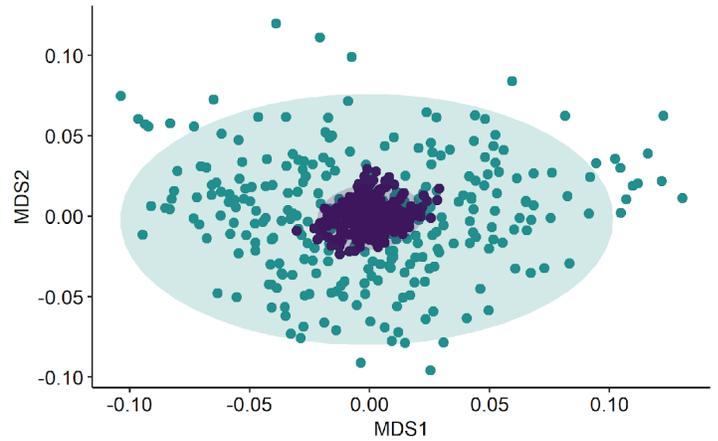
0.001; 12 months samples – $p = 0.006$; Figure S1B).

A. Euclidean-Euclidean

p-value = 0.057
Sum of Squares: 0.98

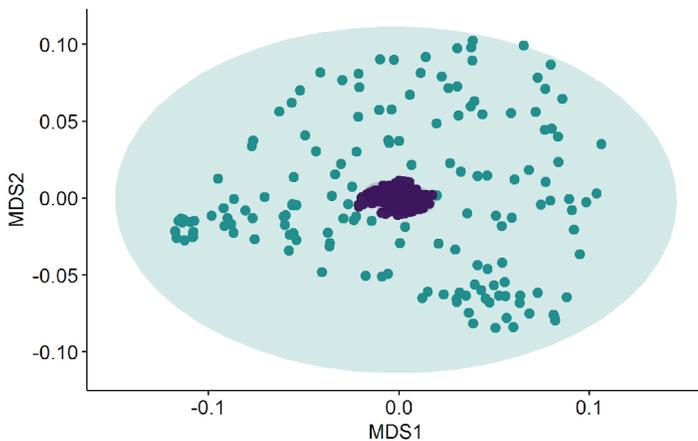


p-value = 0.001
Sum of Squares: 0.95

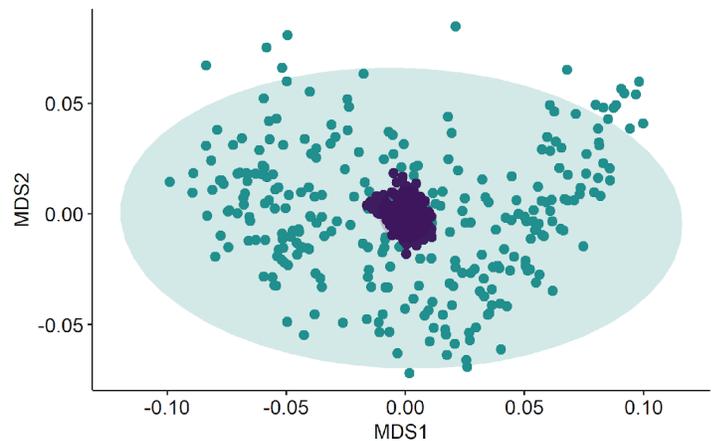


B. Gunifrac-Euclidean

p-value = 0.376
Sum of Squares: 0.99



p-value = 0.069
Sum of Squares: 0.99



6 weeks

12 months

Ordination ● Metabolites ● Taxonomy

141

142

143

144

145

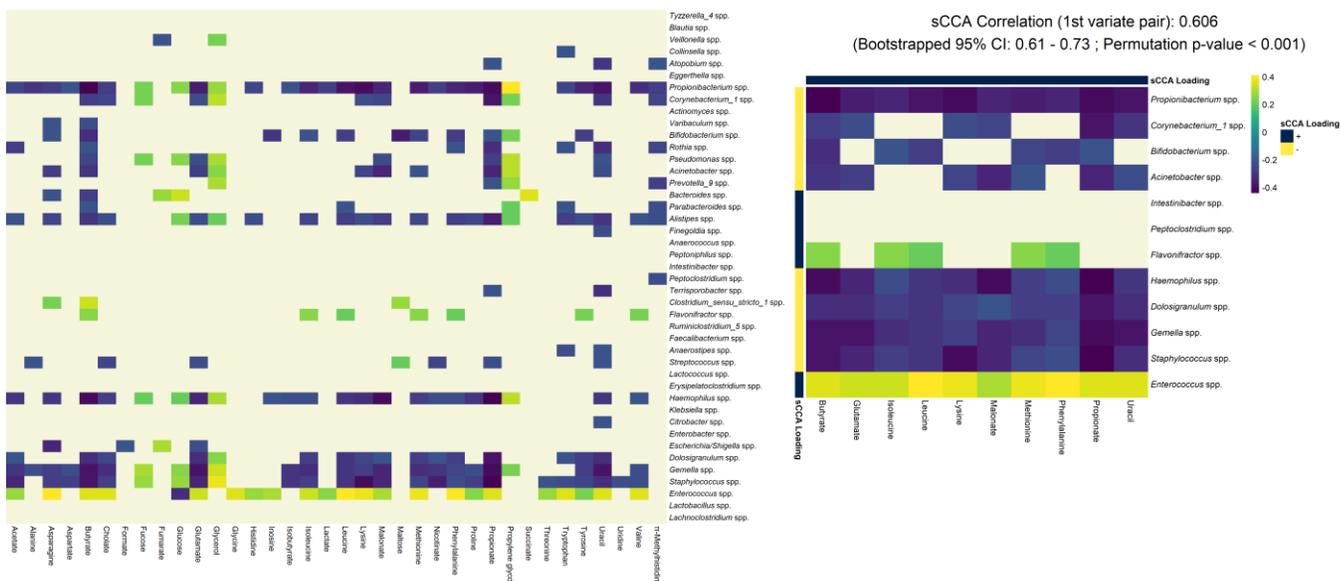
146

147

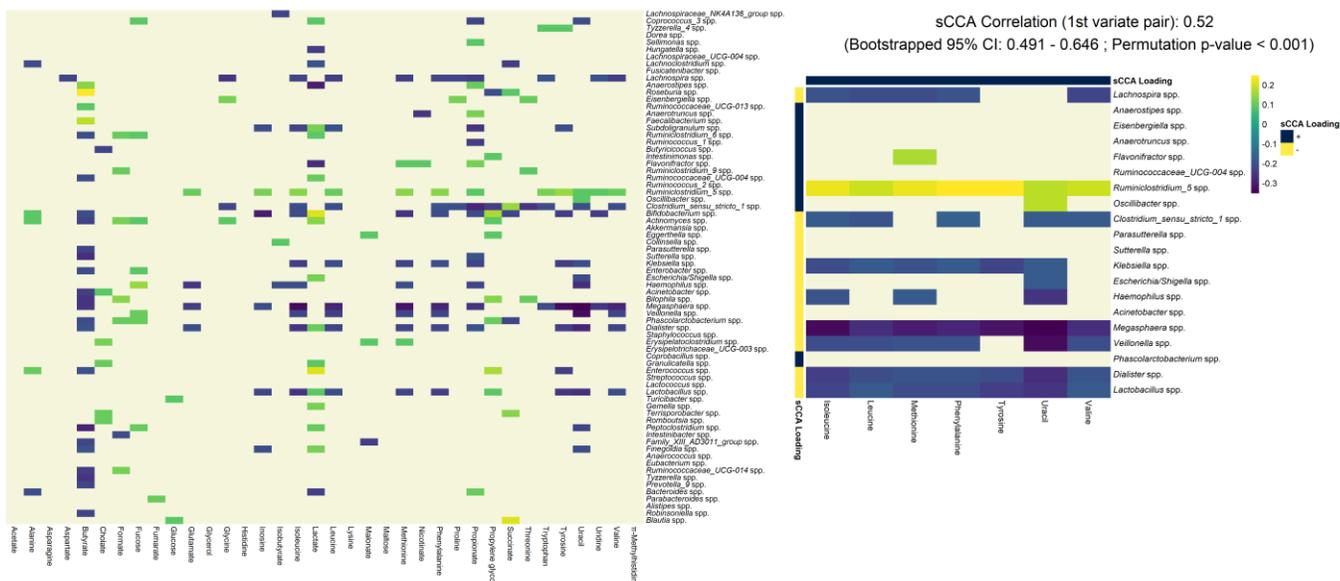
Figure 2: Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and taxonomic relative abundances for 6 weeks (left panels) (n = 158) and 12 months (right panels) (n = 262). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of log-transformed metabolite profiles. Bottom panel presents analyses based on gUniFrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of log-transformed metabolite profiles. There were significant associations between the

148 microbiome and the metabolome (both targeted and untargeted) when utilizing Euclidean distances, however
 149 this association goes away when the gUniFrac distance was employed for the targeted metabolites only.

A. 6 Weeks



B. 12 Months



150
 151 **Figure 3:** Pairwise Spearman correlation of concentration-fitted metabolites and genus-level taxonomic
 152 abundances for 6-weeks (panel A, N = 168) and 12-months (panel B, N = 282) infants. Left panel displays the
 153 overall correlation pattern, where non-significant correlations are not colored (false discovery rate (FDR)

154 controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected
155 by the sparse CCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped
156 95% confidence interval and permutation p-value are also reported. Significant microbiome-metabolome
157 correlation was observed at both time points, however no significant difference was found between the time
158 points.

159 **Sparse canonical correlation analyses reveal the core set of microbe-metabolite groups driving inter-** 160 **omic relatedness.**

161 Given broad concordance between the gut microbiome and metabolome from sample distance analyses, we
162 employed sparse canonical correlation analysis (SCCA) and pairwise Spearman rank correlation to ascertain
163 the strength of association as well as to identify core microbes and metabolites driving this relationship
164 (Methods).

165 The majority of taxa (65% of total genera at 6-weeks and 80% at 12-months) and metabolites (100% of total
166 metabolites at 6-weeks and 80% at 12-months) were part of significant (FDR threshold 0.05) Spearman
167 pairwise comparisons (Supplementary Note 1). This demonstrated a high level of congruence, where most
168 microbes are involved in metabolic processes captured in the stool metabolome. This was also reflected in the
169 significant multivariate correlation (permutation $p < 0.001$). However, at 6 weeks (correlation: 0.606 [0.61 -
170 0.73]), the degree of concordance was slightly higher than at 12 months (correlation: 0.52 [0.431 - 0.646]) but
171 this difference was not significant due to overlapping confidence intervals. The canonical correlation was
172 slightly higher in the untargeted data set (6 weeks: 0.636 [0.621 – 0.733]; 12 months: 0.49 [0.475 – 0.702]),
173 however the difference between time points was similar (Figure S2, Supplemental Note 2).

174 Using SCCA, we identified a core set of microbes and metabolites that are major contributors to the
175 multivariate correlation (Figure 3, right panels; Supplementary Notes 2). Selected microbes (in both the
176 targeted and untargeted data set) belonged to the Firmicutes, Actinobacteria and Proteobacteria phyla, as
177 those are the most commonly found phyla in the infant gut [25, 27]. However, previously established dominant
178 genera such as *Bifidobacterium*, *Bacteroides* and *Lactobacillus* were not consistently selected across both time

179 points. In the targeted data set *Bifidobacterium* was selected only at 6 weeks and *Lactobacillus* was only
180 selected at 12 months. Most notably, more microbes were selected at 12 months compared to 6 weeks in the
181 targeted data set, however in the untargeted data set this pattern was reversed (Figure S2, right panels). In
182 terms of selected metabolites, the majority of the selected metabolites in the targeted data set were amino
183 acids (Supplementary Table 1), with some short chain fatty acids (SCFAs) selected at the 6-week time point.

184 **Microbial community structure is weakly predictive of stool metabolite relative concentrations**

185 In order to determine how well the fecal metabolome acts as a functional representation of the gut microbiome,
186 we fitted metabolite-specific prediction models based on taxonomic profiles. Chosen models include random
187 forest (RF), elastic net (EN), support vector machines with radial basis kernel (SVM-RBF) and sparse partial
188 least squares (SPLS), all of which had previously been shown to work well with microbiome-associated
189 learning tasks [28]. Evaluation was based on predicted R-squared (R^2) and Spearman correlation coefficient
190 (SCC) as measured using 100 repeats of 5-fold nested cross validation (Methods).

191 Predictive performance was more dependent on the metabolite being predicted than by choice of model
192 (Figure 4, Supplementary Notes 3, Supplementary Files 1). Looking at predictive R^2 (Figure 4 panel A), the
193 average posterior mean performance across all models and metabolites was negative for both time points (-
194 5.6% at 6 weeks; -3.07% at 12 months), which indicated that for most prediction tasks the fitted model was
195 less predictive than a naïve, intercept only model. At 6 weeks 22.2% of metabolites had models that perform
196 significantly better than the null (lower bound of 95% credible interval larger than 0) while at 12 months 38.9%
197 of metabolites fit the classification. However, even among such metabolites, the maximum R^2 is only 11.8% at
198 6 weeks and 8.7% at 12 months. Conversely, SCC values were higher in comparison (cross-metabolite avg.:
199 0.339 at 6 weeks and 0.249 at 12 months) (Figure 4 panel B, Supplementary Notes 3). At 6 weeks, 83% of
200 metabolites were significantly more performant than the null, while at 12 months all metabolites were selected.
201 Using a more stringent cutoff as used by Mallick et al. [29], the majority of metabolites at 6 weeks (69.4% of
202 total metabolites) still remained as well predicted while conversely at 12 months only 38.9% (of total
203 metabolites) were predictable.

204 Results from the untargeted analysis showed higher performance values for both evaluation metrics
205 (Supplementary Note 3). Specifically, 56.7% of metabolites bins at 6 weeks and 42.7% of bins at 12 months
206 had R^2 values significantly higher than 0. However, under SCC, while 57% of metabolite bins at 6 weeks had
207 SCC values significantly larger than 0.3 cutoff, only 28.8% of metabolite bins at 12 months fit this criterion.
208 Despite better performance, the overall average values were still low, suggesting that across the entire
209 metabolome few metabolites were highly predictable.

210 Despite weak predictive performance values, we were still interested in determining a model that stands out as
211 the most appropriate for our prediction task. Aggregating performance across metabolites stratified by model
212 for both evaluation metrics (Figure 5, top panel), it can be observed that the average performances were
213 similar (Supplementary Notes 3), for which no semi-targeted analyses performed better on average than the
214 naive model under R^2 . This is further illustrated when model performance was aggregated by rank using Borda
215 scores (Figure 5, bottom panel). A higher score indicated that a model was selected as the top choice more
216 times than others, where an even score distribution across models corroborated the suggestion that no model
217 was best across all prediction tasks. That said, SVM-RBF seemed to be the highest scoring model, particularly
218 for the 6-week time point. The untargeted analysis also found similar results (Figure S3).

219 **Sensitivity analyses**

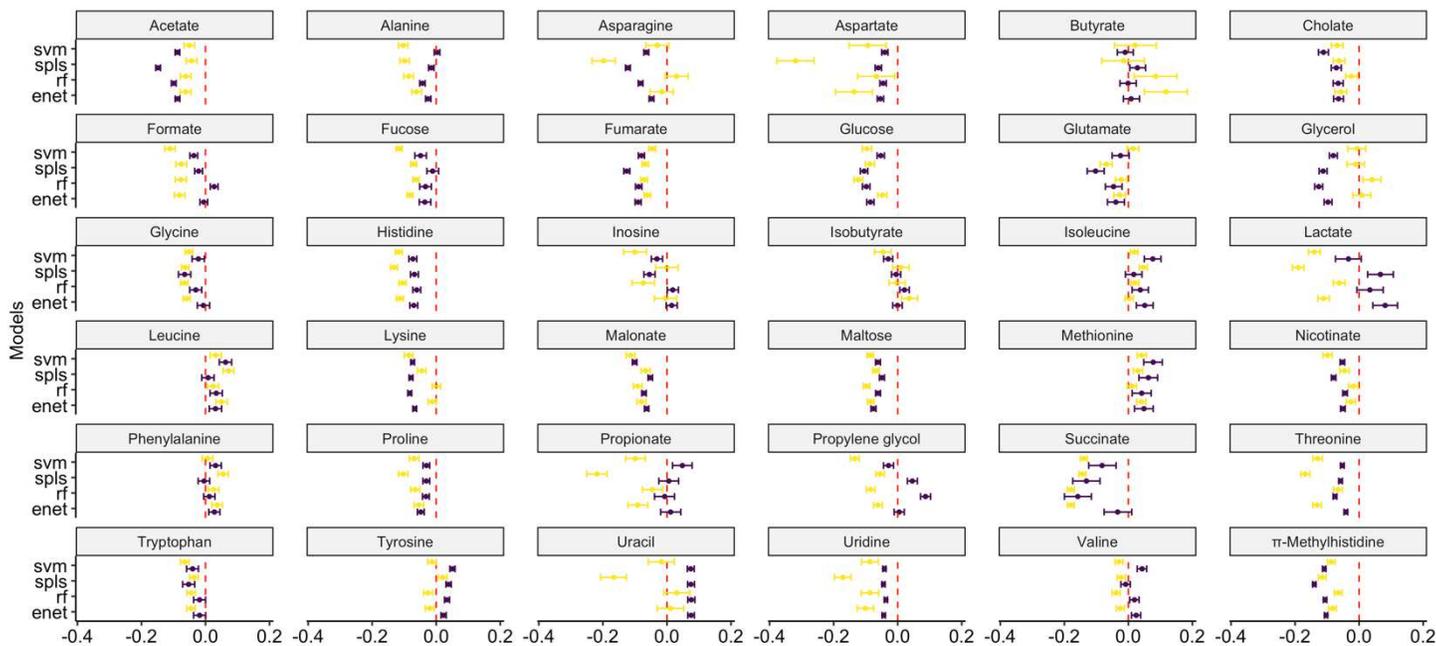
220 We performed both Procrustes and correlation analyses on a data set restricted to the 65 subjects with paired
221 samples collected at both time points (6 weeks and 12 months). Each time point was analyzed separately as in
222 our main analysis. In the targeted data set, significant Procrustes concordance was observed at 12 months (p-
223 value = 0.003) but not at 6 weeks (p-value = 0.106). This association was no longer significant when
224 considering taxonomic ordination using the gUniFrac distance metric (6 weeks). Surprisingly, in the untargeted
225 data set, no association was observed across both time points and choice of distance metric (Figure S5, S6).
226 In the canonical correlation analyses, significance was only observed in the targeted data set at 6 weeks only
227 (6 weeks: permutation p-value = 0.044; 12 months: permutation p-value = 0.388). Even though most
228 correlations were not significantly different from the permuted null, the canonical correlation coefficient is

229 higher at 6 weeks compared to 12 months in both the targeted (6 weeks: 0.676 [0.661 – 0.765]; 12 months:
230 0.52 [0.484 - 0.663]), and untargeted (6 weeks: 0.703 [0.685 – 0.788]; 12 months: 0.444 [0.52 – 0.705]) data
231 sets (Figure S7, S8).

232 Furthermore, to ascertain the uncertainty of model choice, we evaluated all selected modelling approaches
233 with simulated data sets based on bootstrapped resampling of taxonomic relative abundances (Figure S4). For
234 the first simulation scenario, models were assessed against generated metabolite concentrations under
235 different degrees of model saturation (number of taxa associated with the outcome) and association strength
236 (signal to noise ratio). As expected, model performance asymptotically reached perfect prediction with
237 increasing signal strength and model saturation, which demonstrated that prediction models were able to
238 capture predictive associations should they arise even in sparse microbiome data sets. Most notably,
239 simulation performance differed more by signal-to-noise ratio than model saturation, which indicated that the
240 strength of association plays a larger role in the observed weak predictive performance than the number of
241 taxa involved. Surprisingly, we obtained very similar results to our real data values under our lowest simulation
242 setting (model saturation = 5%; signal-to-noise ratio 0.5). As such, it can be suggested that the lack of
243 predictability is due to weak coupling rather than model choice.

A. R-squared

Time — 12M — 6W



B. Spearman Correlation

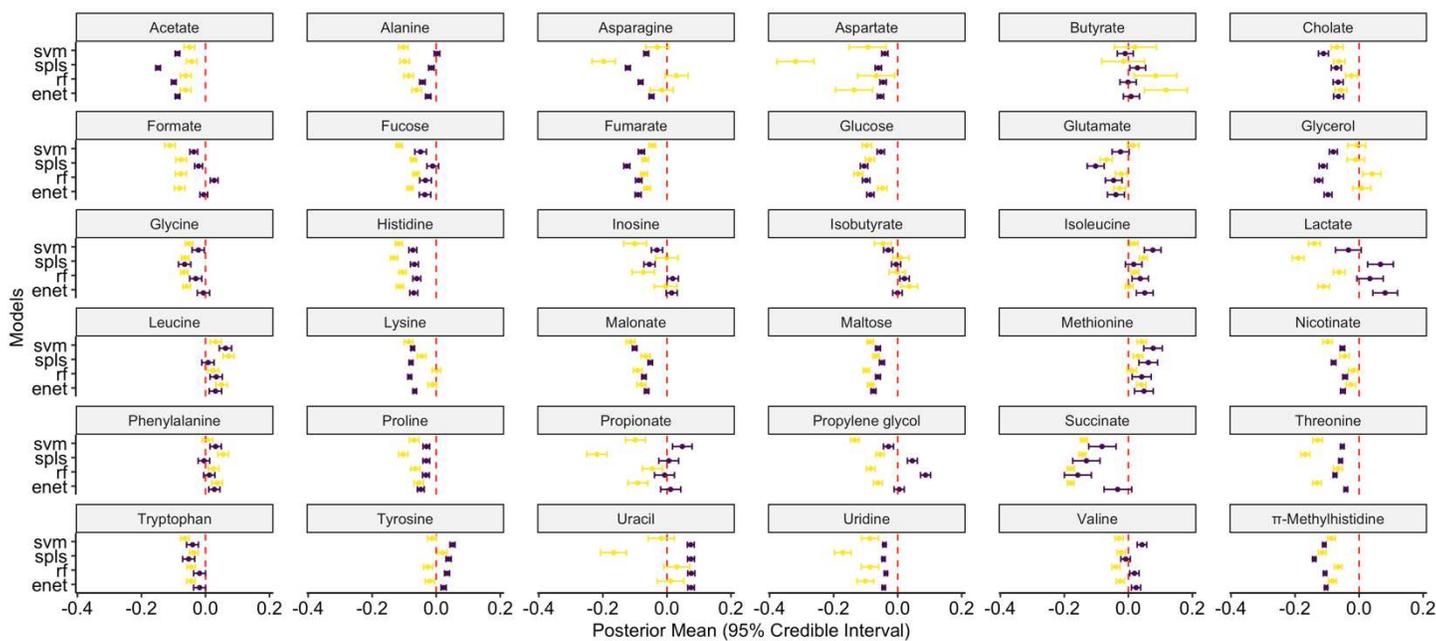
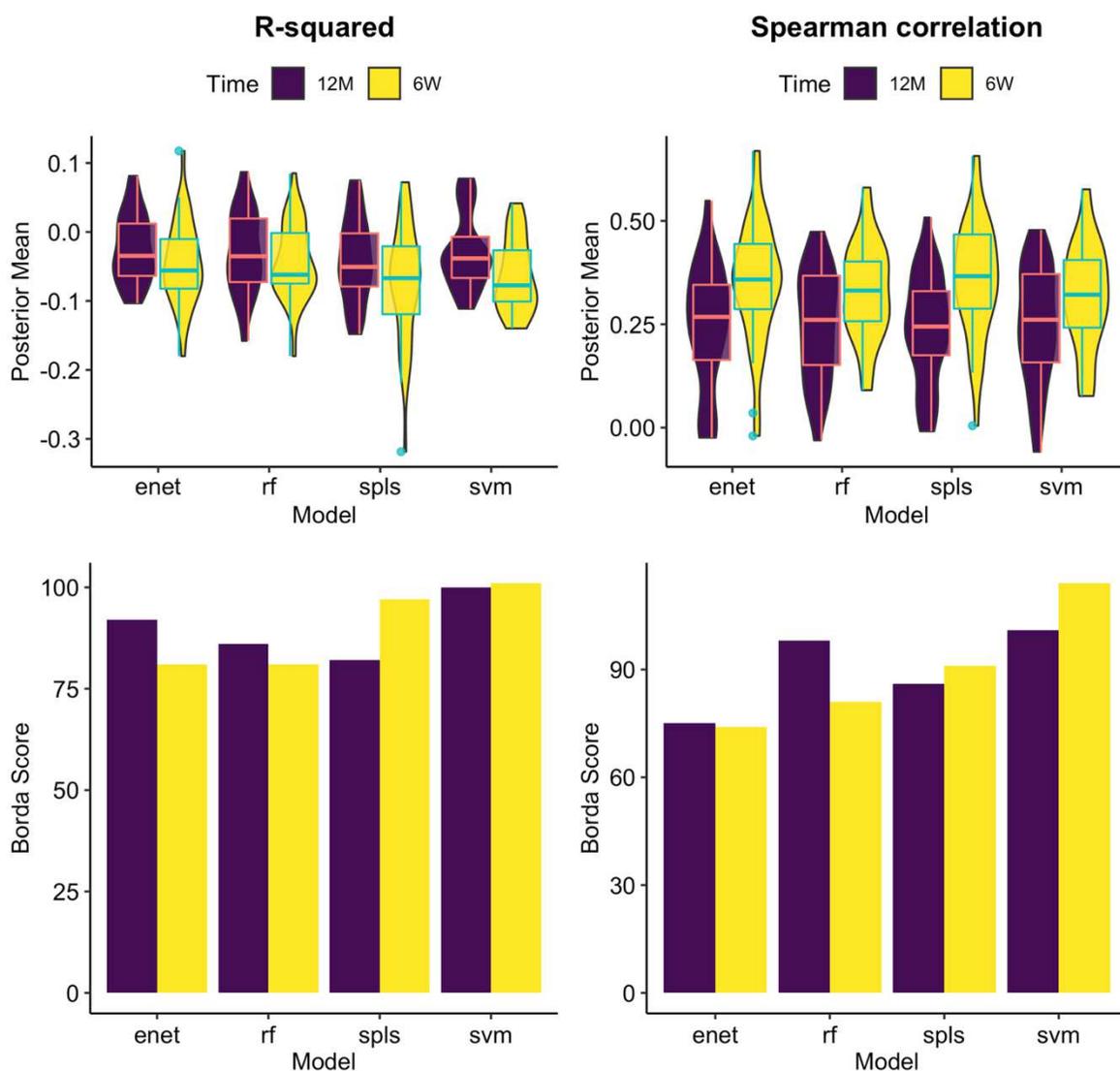


Figure 4. Forest plots of each prediction performance metric (R-squared – Panel A, Spearman correlation – Panel B) for each time point (6 weeks (n = 158), 12 months (n = 282)) across all 36 metabolites and 4 machine learning models. 95% credible interval and predictive posterior means were generated using Bayesian

248 modelling of the evaluation statistic (Methods) after 100 repeats of 5-fold nested cross validation. Red vertical
249 lines indicate a value of 0 for the evaluation metric (equivalent to null model). Metabolites were classified as
250 predictable if the null value did not lie within the estimated 95% credible interval. For most metabolites,
251 predictive performance was not significantly better than null models.

252



253

254 **Figure 5.** Comparative analysis predictive model performance across all metabolites in the targeted dataset for
255 both 6-weeks (n = 158) and 12-months (n = 282) time points. Top panel shows superimposed boxplots and

256 violin plots of the distribution of predictive posterior mean for each evaluation metric across all 36 metabolites.
257 Bottom panels show aggregated model rankings for all metabolites using R-squared (left) and Spearman
258 correlation (right) using Borda scores (Methods). Higher scores indicate that a model was consistently selected
259 as a better performing. Relatively similar Borda scores and cross-metabolite average predictive performances
260 indicate that no model was clearly the most performant. However, support vector machines (with radial basis
261 function kernel) was highest scoring model.

262 **Discussion**

263 In this study, we analyzed the relationship between fecal microbial taxonomic abundances and metabolite
264 concentrations with multi-omic profiling via paired targeted sequencing of the 16S rRNA gene and H¹ NMR
265 metabolomics at multiple time points. Ecological, statistical and machine learning approaches were applied to
266 provide a multi-faceted view of this association. To our knowledge, this study is one of the few comprehensive
267 investigations addressing the microbiome/metabolome interface in a large general population cohort of infants.

268 **The microbiome is significantly correlated but weakly predictive of the metabolome**

269 Overall global concordance was found from three independent methods (Procrustes analysis, SCCA and
270 univariate Spearman correlation), consistent with previous studies on both infant [19, 24] and adult populations
271 [17, 30]. This overall effect was found at both time points, suggesting there coupling exists throughout infancy
272 even when there is a high inter- and intra-individual variability in taxonomic compositions [27].

273 Although the microbiome and the metabolome were significantly correlated in our analyses, most metabolites
274 were not predictable across chosen machine learning models when evaluated by R² statistic. Even among the
275 small number of metabolites that are significantly predictable compared to the null model, the maximum
276 performance values were still low for both the untargeted and targeted analyses. Since all fitted models had
277 been shown to be suited for microbiome-associated prediction tasks [28, 31] as well as covering both non-
278 linear and linear associations, the lack of predictability is unlikely due to model choice. Our data driven
279 simulations further supported this assessment, where low predictability by R² was observed when there is low

280 signal-to-noise ratio. However, when evaluated using SCC, the models seem to perform better. Most of the
281 metabolome (57% of metabolite bins) were significantly well predicted at SCC cut off criteria of 0.3. This result
282 is similar to Mallick et al. [29], which also performs microbiome to metabolome prediction using data from a
283 cohort of adults. The study reported 53.8% of metabolites classified as well-predicted under the elastic net
284 model with the same cutoff. Despite similar percentages of the metabolome being well predicted, the maximum
285 cross-model average correlation value is 0.647 at 6 weeks and 0.431 at 12 months, which is lower than those
286 seen in Mallick et al. Low R^2 compared to SCC demonstrated that even though models were able to anticipate
287 to a certain degree the trends in metabolite abundances, they were not able to accurately predict the exact
288 concentrations.

289 Even though previous research suggests that the fecal metabolome is a functional representation of the gut
290 microbiome [16], metabolites are often resultant of both host and microbial processes [11, 13]. This means for
291 each metabolite, there are host-associated components that drive measured concentrations [32]. We
292 hypothesize that metabolites that are more predictable are those which are associated with microbe-specific
293 metabolism. As such, predictability can be a marker to identify whether a metabolite is host or microbiome
294 driven.

295 However, poor overall performance even amongst predictable metabolites suggests the possible role of other
296 ecological processes. We hypothesized functional redundancy, an aspect ubiquitous in microbial communities
297 [33], plays an important role in this weak coupling. Under functional redundancy, high variability in taxonomic
298 profiles would correspond to little or no change in metabolite abundances, making it difficult to form accurate
299 predictions. Functional redundancy is usually a marker for ecosystem resilience [34]. This is evidenced when
300 Procrustes analyses were adjusted for phylogenetic relatedness using the gUniFrac distance metric where
301 inter-omic association in the targeted data set is no longer significant compared to using Euclidean distances.
302 gUniFrac adjusts for phylogeny by weighting the differences in proportions of each taxa across two samples by
303 the branch length from constructed evolutionary trees [35]. Therefore, the distance between two samples might
304 be larger under gUniFrac if the contributing taxa are evolutionarily divergent. This suggests that samples with

305 similar metabolic profiles might be numerically comparable (cluster together under Euclidean distances) but
306 with evolutionarily divergent taxonomic compositions. This supports the hypothesis that phylogenetically distant
307 gut microbes can replace each other while performing similar roles, a hallmark of functional redundancy [33,
308 36]. Since this was only observed in the targeted data set, we contend that the effect of redundancy is
309 localized to certain metabolites and is less pronounced globally. Functional redundancy is also consistent with
310 previous research in human associated microbiomes [37].

311 **Taxa and metabolites selected to be core to the microbiome-metabolome correlation reveal the**
312 **importance of amino acid metabolism.**

313 Taxa and metabolites with non-zero loading coefficients in SCCA analyses were identified as factors driving
314 this overall correlation. The SCCA procedure utilized a L_1 -penalized matrix decomposition of the cross-product
315 matrix akin to a LASSO regression problem [38], which means that variables were selected based on their
316 importance to the overall covariance between taxa and metabolite abundances.

317 In the targeted data set, Firmicutes, Actinobacteria and Proteobacteria were the most represented phyla
318 across both targeted and untargeted data sets, consistent with the major phyla present in infant guts compared
319 to adults [25, 27, 39]. At 6 weeks, the Firmicutes selected were evenly split between the *Clostridia* and *Bacilli*
320 classes, with *Peptostreptococcaceae* being the most dominant family among *Clostridia*. Alternatively, at 12
321 months, most Firmicutes were only of the *Clostridia* class, mostly represented by the *Ruminococcaceae* and
322 *Lachnospiraceae* families. Both *Clostridia* and *Bacilli* classes have been shown to be important in the
323 production of SCFAs [40, 41]. SCFAs (which includes compounds such as butyrate, formate, propionate and
324 lactate) are important metabolites obtained primarily from colonic microbial fermentation of carbohydrates that
325 escape digestion in the small intestines [42]. Butyrate, for example, is an energy source for colonocytes [43] as
326 well as participating in the maintenance of the gut epithelial barrier through mucin production [44]. Similarly,
327 propionate is part of the gluconeogenesis pathway in liver hepatocyte cells, which is core to lipid and energy
328 metabolism in liver [45]. Most importantly, SCFAs participate in immune programming in early life, where the
329 reduction in SCFA producing bacteria is associated with inflammatory bowel disease [46, 47]. SCFA

330 production in early life is linked to the *Bifidobacterium* and *Bacteroides* catabolism of human milk
331 oligosaccharides (HMO) [48–50], which explains the selection of the *Bifidobacterium* genera at 6 weeks where
332 infants are exclusively on a milk-based diet. The selection of SCFA-associated bacteria further demonstrates
333 the functional role that SCFA play in the interface between the microbiome and the host [42]. This is further
334 supported by the fact that two SCFAs in our targeted list (butyrate and propionate) were selected as important
335 variables in SCCA analyses.

336 Despite the importance of SCFAs, the most selected metabolites are of the amino acids class (7 out of 10
337 metabolites selected at 6 weeks were amino acids). Although microbes can catabolize amino acids for energy
338 [51], this reaction is not energetically efficient [10]. However, microbes will process amino acids in
339 environments with low carbohydrate availability, such as in the distal colon [9, 52], where one of the end point
340 products is SCFAs. This is further supported by the fact that members of the *Peptostreptococcaceae* family
341 (most represented family of selected members of the *Clostridia* class at 6 weeks) has been suggested to be
342 core to the metabolism of amino acids to SCFAs [51]. However, looking across both time points in our analysis,
343 negative correlation with amino acids did not correspond to a positive association with SCFAs (Figure 3),
344 suggesting that amino acid depletion in the fecal metabolome is most likely either to conserve resources from
345 reducing amino acid biosynthesis [10], generate energy in a low carbohydrate environment, or reduce toxic
346 byproducts of protein catabolism [53] rather than to produce SCFAs. However, there could be a possibility that
347 SCFA was produced but rapidly absorbed, which could not be captured by cross-sectional fecal metabolomics.

348 **The microbiome is more tightly coupled with the metabolome in early infancy**

349 Results suggest some level of significant difference in microbiome-metabolome coupling across development.
350 Canonical correlation, while not significantly different, were lower at 12 months than at 6 weeks, suggesting a
351 time-varying effect. When looking at predictability, we observed a higher number of well predicted metabolites
352 at 6 weeks compared to 12 months. Among those selected as well predicted metabolites, the average
353 performance values (both R^2 and SCC) were higher. This is also replicated in the global untargeted data set.

354 There are various factors that can contribute to the difference in microbiome-metabolome coupling between
355 infants at 6 weeks and 12 months. First, there exists substantive differences in dietary patterns for those
356 included in our analysis. The majority of infants at 6 weeks (62%) were exclusively breastfed, while that
357 number is markedly less (35%) at 12 months, at which time infants are also consuming complimentary solid
358 family foods. This transition in diet to solid foods have been shown to induce a change in the gut microbiome
359 composition and diversity due to increased amounts of fiber and protein [54, 55], which might favor certain
360 microbes over others. Such changes in diet, particularly the cessation of breastmilk intake, also contributed
361 towards the development of infant gut microbiomes towards a more “adult like” state [27, 54]. We hypothesized
362 that earlier in life when infants are only consuming a limited type of food (predominantly breast milk or formula),
363 the microbiome participates more actively in host-microbiome co-metabolic activity as infants are more reliant
364 on microbes to breakdown complex nutrients [56]. Conversely, at one year of age where the microbiome has
365 matured, this relationship is not as strongly coupled as a larger share of the metabolome comes from host-
366 produced metabolites.

367 However, as analyses were conducted within each timepoint independently with little subject overlap, further
368 investigations are required to make more conclusive statements about the potential time-varying effect of
369 microbiome-metabolome coupling. Particularly, aside from differences in diet, factors such as differences in
370 antibiotic exposure [57] and maternal covariates [58] might result in differences between time points. In future
371 studies we hope to examine this factor using samples across multiple time points for the same infants.

372 **Limitations**

373 This study has various limitations. First, we utilized partial 16S rRNA gene sequencing instead of shotgun
374 whole genome sequencing, which removes our ability to incorporate the functional metagenome into the
375 microbiome/metabolome interface, as well as limiting our taxonomic resolution to the genus level for most of
376 the analysis [59]. We hypothesized this lack of resolution contribute to overall lack of predictability, as well as
377 limiting the interpretability of variables selected by the SCCA process as species and strain level differences
378 can result in completely separate metabolic contributions [60].

379 Second, our cohort includes only infants from the NHBCS, a population-based cohort reflecting mostly rural
380 and White demographics of northern New England in the United States. While this increases confidence in the
381 internal validity of our study, this homogeneity in race and geography limits the generalizability of our results to
382 other populations.

383 Third, our study is a cross-sectional survey of microbiome-metabolome relationships at two different time
384 points. This means that we cannot capture associations relating to metabolites that are highly produced and
385 consumed. This means that the metabolites selected might not be representative of the intricate relationship
386 between the microbiome and the metabolome. This interpretation is further limited by the lack of annotation for
387 our untargeted metabolite bins, which cannot be compensated by the small number of metabolites selected for
388 the targeted analyses.

389 Finally, each time point was analyzed independently with only 65 subjects with samples in both time points. As
390 such, this limits the ability to explore the differences in coupling across the first year of life.

391 **Conclusion**

392 In conclusion, we conducted one of the first large-scale multi-omics analysis of the microbiome-metabolome
393 relationship using samples from a large birth cohort study at 2 time points (6 weeks and 12 months). Although
394 we found global concordance between the microbiome and the metabolome, the inter-omic concordance is
395 weak, where bacterial abundances cannot predict metabolite concentrations. In addition to the involvement of
396 host-produced metabolites as part of the metabolome, we suggest the role of functional redundancy as a
397 possible mechanism behind the lack of coupling identified. Additionally, we were able to identify metabolites
398 and microbes driving the overall correlation. Results pointed to support the importance of SCFA metabolism
399 particularly at 6 weeks, as well as the role of amino acid metabolism, either as a source of SCFA and energy in
400 the absence of carbohydrates, or as a general mechanism for microbes to save energy as they incorporate
401 amino acids around their environment.

We conclude that although the metabolome is a functional output of the microbiome, there exists massive challenges in being able to trace specific microbial contributions to host-microbe metabolism due to the complexity of factors such as functional redundancy. As such, we recommend studies to profile both the microbiome and the metabolome, as aspects of microbial metabolic contributions cannot be found solely through one omic data set. This is particularly important in settings where it is important to have a mechanistic understanding of the role of microbes such as developing of microbiome therapies [61].

Methods

Study population

Subjects for this study were from the New Hampshire Birth Cohort Study (NHBCS) who provided infant stool samples at 6-weeks and 12-months after birth. As described in previous studies [25, 58], NHBCS is a prospective study of mother-infant dyads in New Hampshire, USA. Participants eligible are pregnant women between the ages of 18 and 45 years old, currently receiving routine prenatal care at one of the study clinics, consuming water out of a private well with no intention to move prior to delivery. The Center for the Protection of Human Subjects at Dartmouth provided institutional review board approval. All methods were carried out in accordance with guidelines. Written informed consent was obtained for participation from all subjects for themselves and their children. Comprehensive sociodemographic, exposure and outcome data such as infant feeding method, delivery mode, maternal smoking status, etc. were collected for all participants through surveys, medical records and telephone interviews conducted during pregnancy, about 6 weeks postpartum, and updated every 4 months up until first year of age and every 6 months thereafter.

Collection of infant stool samples

Infant stool samples were collected at 6-weeks and 12-months. Stool samples were provided in diapers and stored by subjects in their home freezer (-20°C) until they were able to return it to the study site. Stool was thawed at 4 °C so that it could be aliquoted into cryotubes. Stools collected for 16S rRNA gene sequencing were aliquoted (range 350-850 mg) into 3ml RNeasy lysis buffer and homogenized before storing at -80 °C. Stools

426 collected for metabolomic analysis were aliquoted (1-2 grams) into 15ml centrifuge tubes before storing at -80
427 °C.

428 **Taxonomic profiling using 16S rRNA targeted gene sequencing**

429 RNAlater stool samples were thawed and DNA was extracted using the Zymo Fecal DNA extraction kit (Cat#
430 D6010, Zymo Research, Irvine, CA), according to the manufacturer's instructions. For each sample extraction,
431 400ul RNAlater stool slurry (50–100 mg of stool) was used to isolate DNA. Extractions were performed in
432 batches of multiple samples and included a composite RNAlater stool positive control and a RNAlater negative
433 control. Lysis was performed using 750ul Lysis Buffer in ZR BashingBead™ Lysis Tubes (0.5 mm beads),
434 mixed and then shaken on a Disruptor Genie for 6 min. Eluted DNA was quantified on a Qubit™ fluorometer
435 using the Qubit™ dsDNA BR Assay. Average coefficient of variation of DNA yields (ng/ul) for composite
436 RNAlater stool positive controls was 28%. No DNA was ever detectable in negative control elutions.
437 Concentrations of DNA samples used for 16S rRNA gene sequencing range from 1 ng/ul to 25 ng/ul.

438 The V4-V5 hypervariable region of bacterial 16S rRNA gene was sequenced at Marine Biological Laboratory in
439 Woods Hole, MA, using standard approaches [62, 63]. As described previously [25, 58], 16S rDNA V4-V5
440 amplicons were generated from purified genomic DNA samples using fusion primers. The use of forward
441 primers containing one of eight five-nucleotide barcodes between the Illumina-specific bridge and sequencing
442 primer regions and the 16S-specific region and a single reverse primer containing 1 of 12 Illumina indices
443 enables 96 samples per lane multiplexing. Amplifications were done in triplicate with one negative control for
444 internal quality control at MBL. We used qPCR (Kapa Biosystems) to quantify the amplicon pool, and one
445 Illumina MiSeq 500 cycle paired end run to sequence each pool of 96 libraries. We demultiplex and divided
446 datasets using Illumina MiSeq reporter and a custom Python script. Demultiplexed reads derived from Illumina
447 sequencing were denoised and quality filtered using DADA2 (v. 1.12.1) [64] in R (v. 3.6.1) [65]. Illumina
448 adapter sequences were removed prior using cutadapt (v. 1.18). We then use the RDP classifier implemented

449 natively in the DADA2 R package with SILVA database (v. 128) to profile the taxonomy of identified amplicon
450 sequence variants (ASVs).

451 **Functional profiling using untargeted and targeted ¹H NMR metabolomics**

452 ¹H NMR metabolomics was performed in collaboration with the NIH Eastern Regional Comprehensive
453 Metabolomics Resource Core (RCMRC) at UNC Chapel Hill. De-identified stool aliquots were shipped on dry
454 ice and immediately stored at -80 °C for metabolomics analysis. Samples were thawed and ~150mg of stool
455 samples were transferred to MagNA Lyser tubes after recording the weight. Samples were then homogenized
456 with 50% acetonitrile in water by using the Omni Bead Disruptor (Omni International, GA, USA). Homogenized
457 samples were centrifuged at 16000 rcf and the supernatant was separated into another tube. An aliquot (1000
458 uL, 100 mg equivalent of fecal mass) was transferred into an Eppendorf tube and lyophilized overnight. The
459 dried extract was reconstituted in 700 uL of NMR master mix (containing 0.2M phosphate buffer, 0.5 mM DSS-
460 d6 (internal standard), and 0.2% sodium azide (preventing bacterial growth)), vortexed on a multi tube vortexer
461 at speed 5 for 2 min and centrifuged at 16000 rcf for 5 min. A 600 µl aliquot of the supernatant was transferred
462 into pre-labeled 5mm NMR tubes. Additionally, study pooled quality control (QC) samples (created from
463 randomly selected study samples) and batch pooled QC samples were generated from supernatants of study
464 samples and aliquots of supernatants were dried and reconstituted similar to study samples described above
465 and used for QC purposes.

466 ¹H NMR spectra of feces extracts were acquired on a Bruker 700 MHz NMR spectrometer using a 5 mm
467 cryogenically cooled ATMA inverse probe and ambient temperature of 25 C. A 1D NOESY presaturation pulse
468 sequence (noesygprr1d [66, 67], [recycle delay, RD]-90°-t1-90°-tm-90°-acquire free induction decay (FID))
469 was used for data acquisition. For each sample, 64 transients were collected into 64k data points using a
470 spectral width of 12.02 ppm, 2 s relaxation delay, 10 ms mixing time, and an acquisition time of 3.899 s per
471 FID. The water resonance was suppressed using resonance irradiation during the relaxation delay and mixing
472 time. NMR spectra were processed using TopSpin 3.5 software (Bruker-Biospin, Germany). Spectra were zero
473 filled, and Fourier transformed after exponential multiplication with line broadening factor of 0.5. Quality control

474 measures included review of each NMR spectrum for line shape and width, phase and baseline of spectra, and
475 tight clustering of QC samples in Principal Component Analysis [68]. NMR bin data (0.49-9.0 ppm) were
476 generated (untargeted data) excluding water (4.73-4.85 ppm) using intelligent bucket integration of 0.04 ppm
477 bucket width with 50% looseness using ACD Spectrus Processor (ACD Labs Inc, Toronto, Canada).The
478 integrals of each bin were normalized to the total spectral intensity of each spectrum and transferred to
479 analysis software. This resulted in a collection of spectral bins with bin-specific relative abundances, which will
480 be called the untargeted data. In addition, relative concentration of library-matched metabolites (selected from
481 the literature implicated to be important in host-microbe relationships - Table S1) was determined by using
482 Chenomx NMR Suite 8.4 Professional software [26].This data set will be called the targeted data set.

483 **Software and tools**

484 All analyses were performed using the R programming language (v. 3.6.3) [65] and associated packages. All
485 data wrangling steps were performed using *phyloseq* [69], *plyr* and *tidyverse* packages [70], as well as the
486 *compositions* package [71] for log-ratio transformations. All figures were generated using the *ggplot2* [72],
487 *cowplot* [73], *viridis* [74] and *pheatmap* [75] packages. Additionally, the *tidymodels* [76] suite of packages was
488 utilized to assist in all modelling tasks. Specific packages used for modelling will be enumerated below. All
489 scripts as well as intermediary analysis objects are available on github with all dependencies and their versions
490 (https://github.com/qpmnguyen/infant_metabolome_microbiome).

491 **Data transformation and normalization**

492 For microbiome data, we retained all ASVs present in at least 10% of samples [29] and added one
493 pseudocount to all cells [77]. We then subsequently aggregated all ASVs to the genus taxonomic level [28] and
494 converted data to relative proportions using total read counts by sample to account for differential sequencing
495 depth. We further filtered out taxa with mean relative proportion < 0.005% [78]. To address the compositional
496 problem induced by a sum to one constraint, we apply the centered log ratio transformation (CLR), which is
497 often used to remove such constraints in microbiome data sets [79].

Given w as a compositional vector with elements w_1, \dots, w_p where $\sum_1^p w_i = 1$, we can define the centered log ratio transformation $clr(w)$ as z :

$$z = \left[\log \left(\frac{w^1}{g(w)} \right), \dots, \log \left(\frac{w^p}{g(w)} \right) \right]$$

where $g(w) = \left[\prod_{i=1}^p w^i \right]^{\frac{1}{p}}$ is the geometric mean of the composition. Consequently, z exists in a $p - 1$ dimensional Euclidean space and is constrained to sum to 0. The CLR transformation is favored compared to other statistically equivalent log-ratio transformations due to its scale invariant property and ease of interpretation [80].

For metabolomic data sets, we employed different transformations to approximate homoscedasticity depending on the data type (targeted vs untargeted). For targeted metabolites, we performed a $\log(x + 1)$ transformation while for untargeted metabolites we utilized the arcsine square root transformation which has been previously used for transforming composition metabolomics data sets [29].

Distance matrix analyses

Principal coordinates analysis (PCoA) was performed using the *pcoa* function from the *ape* package in R [81] with sample distance matrices. The PCoA procedure seeks to represent high dimensional multivariate data sets in lower dimensions through eigen decomposition of the doubly centered distance matrix. PCoA allows the usage of non-Euclidean distances between samples such as ecological indices, which makes it a preferable method for sample ordination compared to principal component analysis (PCA).

We constructed Euclidean distance matrices for both metabolic and taxonomic profiles post data transformation described in the previous section. Additionally, gUniFrac distances (alpha = 0.5) [62] were considered for taxonomic data using the implementation provided in the package *MiSPU* [63]. gUniFrac requires a phylogenetic tree, of which an approximate maximum likelihood phylogenetic tree was constructed with representative ASV sequences using FastTree (v 2.1) [64]. Multiple sequence alignment was performed using the *AlignSeqs* function from the *DECIPHER* package in R [65] and trees were midpoint rooted using

521 *phytools* [66]. Since multiple sequence alignment is not conserved under filtering and aggregation of ASVs,
522 gUniFrac distance calculations were performed with pre-filtered ASV-level abundances normalized to relative
523 abundances.

524 The first two axes of constructed ordinations were then compared using a symmetric Procrustes procedure
525 implemented in the *protest* function in the *vegan* package [67]. Procrustes superimposes two ordinations by
526 translating and rotating the coordinates, which preserves the general structure of the data. This method
527 performs a superimposition fit between two data sets minimizing the sum-of-squared differences (m^2), which
528 describes the degree of concordance between the two configurations normalized to unit variance. Significance
529 is obtained by testing against the permuted null using a permutation test. This method was shown to have
530 more power while also limiting type I error compared to the traditional Mantel test in ecological analysis [82].
531 Significance was determined using a permutation test on the sum of squared differences with 999
532 permutations [68].

533 **Sparse canonical correlation and Spearman correlation analyses**

534 Sparse canonical correlation analysis (sCCA) was performed to identify strongly associated metabolite-
535 microbe groups. sCCA seeks to find linear combinations of variables from each dataset that maximizes the
536 correlation with each other while simultaneously thresholding variable specific weights to induce sparsity and
537 performing variable selection. The correlation coefficient in the first canonical variate quantifies the overall
538 degree of multivariate associations. As such, sCCA is a popular method in integrating multi-omics datasets
539 with the ability to select more biologically relevant sets of features compared to traditional ecological methods
540 such as co-inertia analysis [83]. In this study, we use the sCCA implementation in the package *PMA* in R [38]
541 which uses a novel penalized matrix decomposition procedure to achieve sparsity [84]. We tune
542 hyperparameters using a permutation approach in the *CCA.permute* function ($nperms = 50$) prior to fitting the
543 final model. We obtain the correlation coefficients as a measure of overall correlation between the two data
544 sets and calculated a bootstrapped 95% confidence interval ($nboot = 5000$) as well as performing a
545 permutation test ($nperm = 1000$) at the 0.05 significance level. In order to keep the structure of the data set

546 across different permutations, we use the function *randomizeMatrix* from the package *picante* in R [85] using
547 the *richness* null model, which randomizes community abundances within samples to maintain sample species
548 richness.

549 Pairwise Spearman correlations were also performed using the *cor* function in R. Hypothesis testing was done
550 using *cor.test*, with multiple hypothesis testing correction using the Benjamini-Hochberg procedure using
551 *p.adjust*. An FDR value of 0.05 is used as cutoff for significance pairwise correlations. Visualization was done
552 using *heatmap* package in R.

553 **Predictive modelling and evaluation**

554 We choose candidate models based on previous research utilizing supervised learning with microbiome
555 associated prediction tasks [28, 29, 31]. Specifically, we chose random forest (RF) [86], support vector
556 machine with radial basis function kernel (SVM-RBF) [87], elastic net (EN) [88] and sparse partial least
557 squares (SPLS) [89], which have all been shown to perform with high-dimensional predictors. These models
558 also support linear and non-linear associations between the microbiome and the outcome of interest. Model
559 fitting, parameter tuning, and evaluation were done using *caret* package in R [90]. Parallel processing was
560 performed using the *doParallel* [91] and *parallel* packages.

561 We evaluate prediction performance by performing 100 repeats of 10-fold nested cross validation, whereby
562 within each training fold is a separate 5-fold cross-validation procedure done to perform hyperparameter
563 selection when appropriate with parameter grids modelled after Pasolli et al. [31]. For RF, we set the number of
564 trees to be 500, and the number of features used in each decision tree to be the square root of the number of
565 the original features. For SVM-RBF, we tuned across a grid for the regularization parameter C (values
566 $2^{-5}, 2^{-3}, \dots, 2^{15}$) and the kernel width parameter γ (values $2^{-15}, 2^{-13}, \dots, 2^3$). For EN, we tuned over a grid of
567 the regularization parameter λ and the L_1 to L_2 penalty ratio α , where for each α value (spaced by 0.1) between
568 0 (equivalent to a LASSO model) and 1 (equivalent to a ridge regression model), we evaluate 100 lambda
569 values chosen by the *glmnet* procedure. For SPLS, we kept the concavity parameter κ constant at 0.5 while
570 tuning the number of components K (values 1,2, ...,10) and the thresholding parameter η (values 0.1,0.2, ...,0.9).

571 We utilize standard regression evaluation metrics include predictive R-squared (R^2) and Spearman correlation
572 coefficient (SCC) which are defined as:

$$573 \quad R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \text{ (} y \text{ and } \hat{y} \text{ are true and predicted values)}$$

$$574 \quad SCC = \frac{cov(y_{(i)}, \hat{y}_{(i)})}{\sigma_{y_{(i)}} \sigma_{\hat{y}_{(i)}}} \text{ (} y_{(i)} \text{ and } \hat{y}_{(i)} \text{ are ranked true values and predicted values)}$$

575 These statistics were chosen due to their ability to capture two different aspects of the regression task.

576 Predictive R^2 captures the predicted residual sum of squares (PRESS) normalized by the total sum of squares
577 of the outcome, thereby measuring predictive performance while also putting it into context of a naive, intercept
578 only model. On the other hand, SCC quantifies the monotonic association between true and predicted values,
579 providing perspective as to whether or not the predicted values can capture the overall trend of the outcome.

580 Prior to evaluation, all metabolites were back transformed to their original scale. In order to perform
581 comparisons between models across time points and metabolites as well as ascertaining the uncertainty of
582 each evaluation metric, a Bayesian approach as presented in [92]. Specifically, a generalized Bayesian
583 hierarchical linear model (with identity link and gaussian standard error) in the following form was fitted for
584 each metabolite:

$$\text{EvaluationStatistic} \sim \text{Model} + (1|\text{repeat}) + (1|\text{repeat}:\text{fold})$$

585 This model assumes that the distribution of the evaluation statistic as a linear function of model assignment,
586 with random intercepts varying among repeats and for folds within each repeat. Models were fitted using
587 implementation in the R package *tidyposterior* [93] using default weakly informative priors as described in the
588 *rstanarm* package [94]. Using this model, a predictive posterior mean and 95% credible interval can be
589 generated. The posterior mean is then used to rank the best performing model for each metabolite according
590 to the evaluation metric of interest. Ranks are then aggregated using the Borda method [95] to generate Borda
591 scores. In detail, for each metabolite, 4 points are added to the top ranked model, 3 points to the second
592 ranked model and so on. The model with the highest total points for each metric is the most performant model
593 aggregated across all prediction tasks.

594 **Simulation design**

595 Simulations were performed to examine the behavior of models under known association/null settings in order
596 to validate findings. For the first simulation scenario, a linear association between genus-level taxonomic
597 abundance and log transformed metabolite concentrations were simulated. The predictor matrix were
598 bootstrapped resamples of the community matrix post data processing. β coefficient values were sampled from
599 the standard normal distribution $N(0,1)$ values for each genus would have a probability p (0.05, 0.1, 0.5, 0.95) of
600 being 0 which determines the sparsity of the coefficients (or the level of model saturation). We generate
601 metabolite outcome values Y following the model

$$Y = \beta_0 + X\beta + \epsilon$$

602 where X is the $n \times p$ simulated taxonomic predictor matrix, β is the $p \times 1$ previously defined coefficient vector,
603 $\epsilon \sim N(\mu = 0, \sigma = \sigma_\epsilon)$ is the standard normal noise vector. Similar to Xiao et al. 2018 [96] and Shi et al. 2016
604 [97], we set all $\beta_0 = \frac{6}{\sqrt{10}}$ and $\sigma_\epsilon = \frac{\sigma(\beta_0 + X\beta)}{SNR}$ where signal-to-noise ratio (SNR) are set at 0.5, 0.7, 3, 5 to simulate
605 both situations where noise is higher than signal and vice versa. For each simulation setting, 100 data sets
606 were generated.

607 For the second simulation scenario, null models were assessed through a permutation procedure using the
608 *picante* package in R as described earlier. A total of 500 permutations was performed for each model.

609 To evaluate the predictive capacity of models for each simulation scenario, each data set was split into a train
610 and test set (80% train; 20% test). Within each training set, a 10-fold cross validation procedure was employed
611 to tune any hyperparameters. Similar evaluation metrics were assessed as described in the model fitting
612 section.

613 **List of abbreviations**

614 NHBCS: New Hampshire Birth Cohort Study

615 NMR: Nuclear Magnetic Resonance

616 PCoA: Principal Coordinates Analysis
617 gUniFrac: Generalized Unique Fraction
618 ASV: Amplicon Sequence Variant
619 FDR: False Discovery Rate
620 sCCA/CCA: Sparse Canonical Correlation Analysis
621 SCFA: Short Chain Fatty Acids
622 RF: Random Forest
623 EN: Elastic Net
624 SVM-RBF: Support Vector Machines with Radial Basis kernel Function
625 SPLS: Sparse Partial Least Squares
626 CLR: Centered Log Ratio transformation

627 **Additional Files**

628 **Additional File 1:** Contains supplementary notes (Notes 1 – 3), supplementary figures (Figures S1-S8) and
629 supplementary table 1. (DOCX 2429KB)

630 **Additional File 2:** Pairs of microbes and targeted metabolites with significant spearman correlations at q-value
631 < 0.05 with their correlation values for each time point. (XLS 22KB)

632 **Additional File 3:** List of targeted metabolites with predictive $R^2 > 0$ or spearman correlation > 0.3 for each
633 model across all time points. (XLSX 17KB)

634 **Declarations**

635 **Ethics approval and consent to participate**

636 The Center for the Protection of Human Subjects at Dartmouth provided institutional review board approval. All
637 methods were carried out in accordance with the guidelines. Written informed consent was obtained for
638 participation from all subjects for themselves and their children.

639 **Availability of data and material**

640 The 16S rRNA gene sequencing datasets used in this study are stored in the National Center for
641 Biotechnology Information (NCBI) Sequence Read Archive: <http://www.ncbi.nlm.nih.gov/sra> [under accession](#)
642 [number PRJNA296814](#). The raw and processed metabolomics data and associated de-identified metadata are
643 available at the NIH Common Fund Metabolomics Data Repository (www.metabolomicsworkbench.org, Study
644 ID#: XXXX – *in progress*). All intermediary analysis objects and scripts are available on [Github](#).

645 **Competing interests**

646 The authors declare that they have no competing financial interests

647 **Funding**

648 This research is supported by funding from the National Institutes of Health (grants NLM R01LM012723,
649 NIGMS P20GM104416, NCI P30CA023108, NCI R21CA253408, NLM K01LM012426, NIH UG3 OD023275,
650 NIEHS P01ES022832 and EPA RD-83544201). None of the funding bodies had a role in the design of the
651 study or the collection, analysis, and interpretation of data and in writing of the manuscript.

652 **Authors' contributions**

653 QPN, AGH and HRF designed the research. QPN performed the statistical analysis and wrote the paper. QPN,
654 AGH, HRF, MRK, and JCM interpreted the results. ED performed bioinformatic processing and maintained the
655 data. TJP performed laboratory extractions of samples. HGM provided technical and methodological support
656 for sequencing data. SM, SJS and WWP provided technical and methodological support for metabolomic. AGH
657 and HRF jointly supervised this research. AGH has primary responsibility for all the content. All authors have
658 reviewed the manuscript, offered critical feedback, and approved the final version.

659 **Acknowledgements**

660 We thank the participants of the New Hampshire Birth Cohort Study and to the study staff, without whom this
661 research would not be possible. We would also like to thank all members of the Microbiome research group at
662 Dartmouth College for all advice and additional assistance in preparing this manuscript.

663 **Authors' Information**

664 **Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA**

665 Quang P. Nguyen, Erika Dade, Tom J. Palys, Juliette C. Madan, Margaret R. Karagas & Anne G. Hoen
666 **Division of Neonatology, Department of Pediatrics, Children’s Hospital at Dartmouth, Hanover, NH,**
667 **USA**

668 Juliette C. Madan
669 **Children’s Environmental Health & Disease Prevention Research Center at Dartmouth, Lebanon, NH,**
670 **USA**

671 Juliette C. Madan, Margaret R. Karagas & Anne G. Hoen
672 **Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH, USA**

673 Anne G. Hoen & H. Robert Frost
674 **Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA**

675 Hilary G. Morrison
676 **Department of Nutrition, Nutrition Research Institute, University of North Carolina at Chapel Hill,**
677 **Chapel Hill, NC, USA**

678 Wimal W. Pathmasiri, Susan McRitche & Susan J. Sumner

679

680 **References**

681 1. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol.*
682 2015;31:69–75.

683 2. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. Development of the Human Infant Intestinal
684 Microbiota. *PLOS Biol.* 2007;5:e177.

685 3. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia
686 in the developing infant gut microbiome. *Proc Natl Acad Sci.* 2011;108 Supplement 1:4578–85.

- 687 4. Stanislowski MA, Dabelea D, Wagner BD, Iszatt N, Dahl C, Sontag MK, et al. Gut Microbiota in the First 2
688 Years of Life and the Association with Body Mass Index at Age 12 in a Norwegian Birth Cohort. *mBio*. 2018;9.
689 doi:10.1128/mBio.01751-18.
- 690 5. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The Dynamics of the
691 Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host Microbe*.
692 2015;17:260–73.
- 693 6. Arrieta M-C, Stiemsma LT, Amenyogbe N, Brown EM, Finlay B. The Intestinal Microbiome in Early Life:
694 Health and Disease. *Front Immunol*. 2014;5. doi:10.3389/fimmu.2014.00427.
- 695 7. Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. *Trends Microbiol*. 2018;26:563–74.
- 696 8. Turnbaugh PJ, Gordon JI. An invitation to the marriage of metagenomics and metabolomics. *Cell*.
697 2008;134:708–13.
- 698 9. Ewald DR, Sumner SC. Human Microbiota, Blood Group Antigens, and Disease. *Wiley Interdiscip Rev Syst*
699 *Biol Med*. 2018;10:e1413.
- 700 10. Oliphant K, Allen-Vercoe E. Macronutrient metabolism by the human gut microbiome: major fermentation
701 by-products and their impact on host health. *Microbiome*. 2019;7:91.
- 702 11. Heinken A, Thiele I. Systems biology of host–microbe metabolomics. *Wiley Interdiscip Rev Syst Biol Med*.
703 2015;7:195–219.
- 704 12. Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, et al. Symbiotic gut microbes modulate human
705 metabolic phenotypes. *Proc Natl Acad Sci*. 2008;105:2117–22.
- 706 13. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-Gut Microbiota Metabolic
707 Interactions. *Science*. 2012;336:1262–7.

- 708 14. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BAH, et al. Human gut
709 microbes impact host serum metabolome and insulin sensitivity. *Nature*. 2016;535:376–81.
- 710 15. Neis EPJG, Dejong CHC, Rensen SS. The Role of Microbial Amino Acid Metabolism in Host Metabolism.
711 *Nutrients*. 2015;7:2930–46.
- 712 16. Zierer J, Jackson MA, Kastenmüller G, Mangino M, Long T, Telenti A, et al. The fecal metabolome as a
713 functional readout of the gut microbiome. *Nat Genet*. 2018;50:790–5.
- 714 17. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of
715 the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569:655–62.
- 716 18. Wandro S, Osborne S, Enriquez C, Bixby C, Arrieta A, Whiteson K. The microbiome and metabolome of
717 pre-term infant stool is personalized, and not driven by health outcomes including necrotizing enterocolitis and
718 late-onset sepsis. 2018.
- 719 19. Ayeni FA, Biagi E, Rampelli S, Fiori J, Soverini M, Audu HJ, et al. Infant and Adult Gut Microbiome and
720 Metabolome in Rural Bassa and Urban Settlers from Nigeria. *Cell Rep*. 2018;23:3056–67.
- 721 20. Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Fofanova T, Nelson A, et al. Longitudinal development of
722 the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls.
723 *Microbiome*. 2017;5. doi:10.1186/s40168-017-0295-1.
- 724 21. Younge NE, Newgard CB, Cotten CM, Goldberg RN, Muehlbauer MJ, Bain JR, et al. Disrupted Maturation
725 of the Microbiota and Metabolome among Extremely Preterm Infants with Postnatal Growth Failure. *Sci Rep*.
726 2019;9:1–12.
- 727 22. Brink L, Chintapalli S, Mercer K, Piccolo B, Adams S, Bowlin A, et al. Early Postnatal Diet Differentially
728 Affects the Fecal Microbiome and Metabolome (FS04-02-19). *Curr Dev Nutr*. 2019;3 Supplement_1.
729 doi:10.1093/cdn/nzz048.FS04-02-19.

- 730 23. Kisuse J, La-ongkham O, Nakphaichit M, Therdtatha P, Momoda R, Tanaka M, et al. Urban Diets Linked to
731 Gut Microbiome and Metabolome Alterations in Children: A Comparative Cross-Sectional Study in Thailand.
732 *Front Microbiol.* 2018;9. doi:10.3389/fmicb.2018.01345.
- 733 24. Hill CJ, Lynch DB, Murphy K, Ulaszewska M, Jeffery IB, O’Shea CA, et al. Evolution of gut microbiota
734 composition from birth to 24 weeks in the INFANTMET Cohort. *Microbiome.* 2017;5:4.
- 735 25. Madan JC, Hoen AG, Lundgren SN, Farzan SF, Cottingham KL, Morrison HG, et al. Effects of Cesarean
736 delivery and formula supplementation on the intestinal microbiome of six-week old infants. *JAMA Pediatr.*
737 2016;170:212–9.
- 738 26. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. Targeted Profiling: Quantitative Analysis of ¹ H
739 NMR Metabolomics Data. *Anal Chem.* 2006;78:4430–42.
- 740 27. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of
741 the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe.* 2015;17:690–703.
- 742 28. Zhou Y-H, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait
743 Prediction. *Front Genet.* 2019;10. doi:10.3389/fgene.2019.00579.
- 744 29. Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, et al. Predictive metabolomic
745 profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun.* 2019;10:1–11.
- 746 30. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, et al. Integrative analysis of the
747 microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships.
748 *Microbiome.* 2013;1:17.
- 749 31. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large
750 Metagenomic Datasets: Tools and Biological Insights. *PLOS Comput Biol.* 2016;12:e1004977.

- 751 32. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, Bokatzian S, et al. AMON: annotation of
752 metabolite origins via networks to integrate microbiome and metabolome data. *BMC Bioinformatics*.
753 2019;20:614.
- 754 33. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, et al. Function and functional
755 redundancy in microbial systems. *Nat Ecol Evol*. 2018;:1.
- 756 34. Allison SD, Martiny JBH. Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad*
757 *Sci*. 2008;105 Supplement 1:11512–9.
- 758 35. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition
759 with environmental covariates using generalized UniFrac distances. *Bioinformatics*. 2012;28:2106–13.
- 760 36. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Yunta RG, Okuda S, et al. Species–function
761 relationships shape ecological properties of the human gut microbiome. *Nat Microbiol*. 2016;1:16088.
- 762 37. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the
763 human gut microbiota. *Nature*. 2012;489:220–30.
- 764 38. Witten D, Tibshirani R, Gross S, Narasimhan B. PMA: penalized multivariate analysis. 2019.
765 <https://CRAN.R-project.org/package=PMA>.
- 766 39. Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of
767 the gut microbiome in early childhood from the TEDDY study. *Nature*. 2018;562:583–8.
- 768 40. Lopetuso LR, Scalfaferrri F, Petito V, Gasbarrini A. Commensal Clostridia: leading players in the
769 maintenance of gut homeostasis. *Gut Pathog*. 2013;5:23.
- 770 41. Ilinskaya ON, Ulyanova VV, Yarullina DR, Gataullin IG. Secretome of Intestinal Bacilli: A Natural Guard
771 against Pathologies. *Front Microbiol*. 2017;8. doi:10.3389/fmicb.2017.01666.

- 772 42. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on
773 human metabolism. *Gut Microbes*. 2016;7:189–200.
- 774 43. LeBlanc JG, Chain F, Martín R, Bermúdez-Humarán LG, Courau S, Langella P. Beneficial effects on host
775 energy metabolism of short-chain fatty acids and vitamins produced by commensal and probiotic bacteria.
776 *Microb Cell Factories*. 2017;16. doi:10.1186/s12934-017-0691-z.
- 777 44. Peng L, Li Z-R, Green RS, Holzman IR, Lin J. Butyrate Enhances the Intestinal Barrier by Facilitating Tight
778 Junction Assembly via Activation of AMP-Activated Protein Kinase in Caco-2 Cell Monolayers. *J Nutr*.
779 2009;139:1619–25.
- 780 45. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K, et al. Gut-derived short-chain
781 fatty acids are vividly assimilated into host carbohydrates and lipids. *Am J Physiol-Gastrointest Liver Physiol*.
782 2013;305:G900–10.
- 783 46. Corrêa-Oliveira R, Fachi JL, Vieira A, Sato FT, Vinolo MAR. Regulation of immune cell function by short-
784 chain fatty acids. *Clin Transl Immunol*. 2016;5:e73.
- 785 47. Schulthess J, Pandey S, Capitani M, Rue-Albrecht KC, Arnold I, Franchini F, et al. The Short Chain Fatty
786 Acid Butyrate Imprints an Antimicrobial Program in Macrophages. *Immunity*. 2019;50:432-445.e7.
- 787 48. James K, Bottacini F, Contreras JIS, Vigoureux M, Egan M, Motherway MO, et al. Metabolism of the
788 predominant human milk oligosaccharide fucosyllactose by an infant gut commensal. *Sci Rep*. 2019;9:1–20.
- 789 49. Lawson MAE, O'Neill IJ, Kujawska M, Gowrinadh Javvadi S, Wijeyesekera A, Flegg Z, et al. Breast milk-
790 derived human milk oligosaccharides promote *Bifidobacterium* interactions within a single ecosystem. *ISME J*.
791 2020;14:635–48.
- 792 50. Marcobal A, Barboza M, Sonnenburg ED, Pudlo N, Martens EC, Desai P, et al. *Bacteroides* in the Infant
793 Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways. *Cell Host Microbe*. 2011;10:507–14.

794 51. Dai Z-L. Amino acid metabolism in intestinal bacteria: links between gut ecology and host health. *Front*
795 *Biosci.* 2011;16:1768.

796 52. Macfarlane GT, Macfarlane S. Bacteria, Colonic Fermentation, and Gastrointestinal Health. *J AOAC Int.*
797 2012;95:50–60.

798 53. Fan P, Li L, Rezaei A, Eslamfam S, Che D, Ma X. Metabolites of Dietary Protein and Peptides by Intestinal
799 Microbes and their Impacts on Gut. *Curr Protein Pept Sci.* 2015;16:646–54.

800 54. Moore RE, Townsend SD. Temporal development of the infant gut microbiome. *Open Biol.* 2019;9:190128.

801 55. Laursen MF, Bahl MI, Michaelsen KF, Licht TR. First Foods and Gut Microbes. *Front Microbiol.* 2017;8.
802 doi:10.3389/fmicb.2017.00356.

803 56. Milani C, Duranti S, Bottacini F, Casey E, Turrone F, Mahony J, et al. The First Microbial Colonizers of the
804 Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol Mol Biol*
805 *Rev.* 2017;81:e00036-17.

806 57. Coker MO, Hoen AG, Dade E, Lundgren S, Li Z, Wong AD, et al. Specific class of intrapartum antibiotics
807 relates to maturation of the infant gut microbiota: a prospective cohort study. *BJOG Int J Obstet Gynaecol.*
808 2020;127:217–27.

809 58. Lundgren SN, Madan JC, Emond JA, Morrison HG, Christensen BC, Karagas MR, et al. Maternal diet
810 during pregnancy is related with the infant stool microbiome in a delivery mode-dependent manner.
811 *Microbiome.* 2018;6:109.

812 59. Zhang C, Zhao L. Strain-level dissection of the contribution of the gut microbiome to human metabolic
813 disease. *Genome Med.* 2016;8. doi:10.1186/s13073-016-0304-1.

- 814 60. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and
815 dynamics in the expanded Human Microbiome Project. *Nature*. 2017. doi:10.1038/nature23889.
- 816 61. Lemon KP, Armitage GC, Relman DA, Fischbach MA. Microbiota-Targeted Therapies: An Ecological
817 Perspective. *Sci Transl Med*. 2012;4:137rv5.
- 818 62. Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM, et al. Sewage Reflects the
819 Microbiomes of Human Populations. *mBio*. 2015;6. doi:10.1128/mBio.02574-14.
- 820 63. Huse SM, Young VB, Morrison HG, Antonopoulos DA, Kwon J, Dalal S, et al. Comparison of brush and
821 biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human
822 subjects. *Microbiome*. 2014;2:5.
- 823 64. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution
824 sample inference from Illumina amplicon data. *Nat Methods*. 2016;13:581–3.
- 825 65. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2019.
826 <https://www.R-project.org/>.
- 827 66. Beckonert O, Keun HC, Ebbels TMD, Bundy J, Holmes E, Lindon JC, et al. Metabolic profiling,
828 metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.
829 *Nat Protoc*. 2007;2:2692–703.
- 830 67. Dona AC, Jiménez B, Schäfer H, Humpfer E, Spraul M, Lewis MR, et al. Precision high-throughput proton
831 NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem*.
832 2014;86:9887–94.
- 833 68. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, et al. Guidelines and
834 considerations for the use of system suitability and quality control samples in mass spectrometry assays
835 applied in untargeted clinical metabolomic studies. *Metabolomics Off J Metabolomic Soc*. 2018;14:72.

- 836 69. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of
837 microbiome census data. PLoS ONE. 2013;8:e61217.
- 838 70. Wickham H. The split-apply-combine strategy for data analysis. J Stat Softw. 2011;40:1–29.
- 839 71. van den Boogaart KG, Tolosana-Delgado R, Bren M. compositions: Compositional data analysis. 2019.
840 <https://CRAN.R-project.org/package=compositions>.
- 841 72. Wickham H. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York; 2016.
842 <https://ggplot2.tidyverse.org>.
- 843 73. Wilke CO. cowplot: Streamlined plot theme and plot annotations for “ggplot2.” 2019. [https://CRAN.R-](https://CRAN.R-project.org/package=cowplot)
844 [project.org/package=cowplot](https://CRAN.R-project.org/package=cowplot).
- 845 74. Garnier S. viridis: Default color maps from “matplotlib.” 2018. <https://CRAN.R-project.org/package=viridis>.
- 846 75. Kolde R. pheatmap: Pretty heatmaps. 2019. <https://CRAN.R-project.org/package=pheatmap>.
- 847 76. Kuhn M, Wickham H. tidymodels: Easily install and load the “tidymodels” packages. 2020. [https://CRAN.R-](https://CRAN.R-project.org/package=tidymodels)
848 [project.org/package=tidymodels](https://CRAN.R-project.org/package=tidymodels).
- 849 77. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust
850 Inference of Microbial Ecological Networks. PLOS Comput Biol. 2015;11:e1004226.
- 851 78. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly
852 improves diversity estimates from Illumina amplicon sequencing. Nat Methods. 2013;10:57.
- 853 79. Aitchison J. The Statistical Analysis of Compositional Data. J R Stat Soc Ser B Methodol. 1982;44:139–77.
- 854 80. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And
855 This Is Not Optional. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.02224.

- 856 81. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.
857 Bioinformatics. 2018;35:526–8.
- 858 82. Peres-Neto PR, Jackson DA. How well do multivariate data sets match? The advantages of a Procrustean
859 superimposition approach over the Mantel test. *Oecologia*. 2001;129:169–78.
- 860 83. Cao D-S, Liu S, Zeng W-B, Liang Y-Z. Sparse canonical correlation analysis applied to -omics studies for
861 integrative analysis and biomarker discovery. *J Chemom*. 29:371–8.
- 862 84. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal
863 components and canonical correlation analysis. *Biostat Oxf Engl*. 2009;10:515–34.
- 864 85. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for
865 integrating phylogenies and ecology. *Bioinformatics*. 2010;26:1463–4.
- 866 86. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.
- 867 87. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of
868 the Fifth Annual Workshop on Computational Learning Theory. New York, NY, USA: ACM; 1992. p. 144–152.
869 doi:10.1145/130385.130401.
- 870 88. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc Ser B Stat*
871 *Methodol*. 2005;67:301–20.
- 872 89. Chun H, KeleÅ S. Sparse partial least squares regression for simultaneous dimension reduction and
873 variable selection. *J R Stat Soc Ser B Stat Methodol*. 2010;72:3–25.
- 874 90. Wing MKuhnC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. Caret: classification
875 and regression training. 2019. <https://CRAN.R-project.org/package=caret>.

- 876 91. Corporation M, Weston S. doParallel: Foreach parallel adaptor for the “parallel” package. 2019.
877 <https://CRAN.R-project.org/package=doParallel>.
- 878 92. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a Change: a Tutorial for Comparing Multiple
879 Classifiers Through Bayesian Analysis. :36.
- 880 93. Kuhn M. tidyposterior: Bayesian analysis to compare models using resampling statistics. 2018.
881 <https://CRAN.R-project.org/package=tidyposterior>.
- 882 94. Brilleman S, Crowther M, Moreno-Betancur M, Buros Novik J, Wolfe R. Joint longitudinal and time-to-event
883 models via Stan. https://github.com/stan-dev/stancon_talks/.
- 884 95. Lin S. Rank aggregation methods. *Wiley Interdiscip Rev Comput Stat*. 2010;2:555–70.
- 885 96. Xiao J, Chen L, Yu Y, Zhang X, Chen J. A Phylogeny-Regularized Sparse Regression Model for Predictive
886 Modeling of Microbial Community Data. *Front Microbiol*. 2018;9. doi:10.3389/fmicb.2018.03112.
- 887 97. Shi P, Zhang A, Li H. Regression analysis for microbiome compositional data. *Ann Appl Stat*.
888 2016;10:1019–40.

889 **Figure and Table Title and Legend**

890 **Figure 1.** Overview of the analysis. Panel A describes the subject selection workflow and panel B describes
891 the analytic pipeline.

892 **Table 1.** Selected characteristics of subjects contributing samples at 6 weeks (n = 158) and at 12 months of
893 age (n = 282).

894 **Figure 2:** Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and
895 taxonomic relative abundances for 6 weeks (left panels) (n = 158) and 12 months (right panels) (n = 262). Top
896 panels present analyses based on ordinations from Euclidean distances of genus level abundances after
897 centered log ratio transformation and Euclidean distances of log-transformed metabolite profiles. Bottom panel

presents analyses based on gUniFrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of log-transformed metabolite profiles. There were significant associations between the microbiome and the metabolome (both targeted and untargeted) when utilizing Euclidean distances, however this association goes away when the gUniFrac distance was employed for the targeted metabolites only.

Figure 3: Pairwise Spearman correlation of concentration-fitted metabolites and genus-level taxonomic abundances for 6-weeks (panel A, N = 168) and 12-months (panel B, N = 282) infants. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (false discovery rate (FDR) controlled q -value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sparse CCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval and permutation p -value are also reported. Significant microbiome-metabolome correlation was observed at both time points, however no significant difference was found between the time points.

Figure 4. Forest plots of each prediction performance metric (R-squared – Panel A, Spearman correlation – Panel B) for each time point (6 weeks (n = 158), 12 months (n = 282)) across all 36 metabolites and 4 machine learning models. 95% credible interval and predictive posterior means were generated using Bayesian modelling of the evaluation statistic (Methods) after 100 repeats of 5-fold nested cross validation. Red vertical lines indicate a value of 0 for the evaluation metric (equivalent to null model). Metabolites were classified as predictable if the null value did not lie within the estimated 95% credible interval. For most metabolites, predictive performance was not significantly better than null models

Figure 5. Comparative analysis predictive model performance across all metabolites in the targeted dataset for both 6-weeks (n = 158) and 12-months (n = 282) time points. Top panel shows superimposed boxplots and violin plots of the distribution of predictive posterior mean for each evaluation metric across all 36 metabolites. Bottom panels show aggregated model rankings for all metabolites using R-squared (left) and Spearman correlation (right) using Borda scores (Methods). Higher scores indicate that a model was consistently selected as a better performing. Relatively similar Borda scores and cross-metabolite average predictive performances

923 indicate that no model was clearly the most performant. However, support vector machines (with radial basis
924 function kernel) was highest scoring model.

925 **Figure S1.** Inter-omics Procrustes biplots comparing PCoA ordinations of untargeted metabolite profiles and
926 taxonomic relative abundances for 6 weeks (left panels) (n = 158) and 12 months (right panels) (n = 262). Top
927 panels present analyses based on ordinations from Euclidean distances of genus level abundances after
928 centered log ratio transformation and Euclidean distances of arcsine square root transformed metabolite
929 relative abundances. Bottom panel presents analyses based on generalized Unifrac distance of amplicon
930 sequence variant (ASV) relative abundances and Euclidean distances of arcsine square root transformed
931 metabolite relative abundances.

932 **Figure S2.** Pairwise Spearman correlation of metabolite bins and genus-level taxonomic abundances for 6-
933 weeks (panel A, N = 168) and 12-months (panel B, N = 282) infants. Left panel displays the overall correlation
934 pattern, where non-significant correlations are not colored (false discovery rate (FDR) controlled q-value <
935 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sparse CCA
936 procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence
937 interval and permutation p-value are also reported.

938 **Figure S3.** Comparative analysis predictive model performance across all metabolites in the untargeted
939 dataset for both 6-weeks (n = 158) and 12-months (n = 282) timepoints. Top panel shows superimposed
940 boxplots and violin plots of the distribution of predictive posterior mean for each evaluation metric across all
941 208 spectral bins. Bottom panels show aggregated model rankings for all metabolites using R-squared (left)
942 and spearman correlation (right) using Borda scores (Methods).

943 **Figure S4.** Results for positive (Panel A) and negative simulations (Panel B). Positive simulations were
944 conducted based on bootstrapped resamples of the original data (12-month time point) and a normally
945 distributed outcome vector which represented a log-transformed metabolite profile. Different levels of model
946 saturation (horizontal, model sparsity (spar) at 0.05, 0.1, 0.5, 0.95) and effect sizes (vertical, signal-to-noise
947 ratio (snr) at 0.5, 0.7, 3, 5) were assessed, with 100 data sets generated for each setting combination.

948 Negative simulations were conducted based on permutations of the original data (12-month time point), with a
949 total of 1000 permutations. Highly negative outliers were removed for the purposes of visualization

950 **Figure S5.** Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and
951 taxonomic relative abundances in the sensitivity analyses for 6 weeks (left panels) (n = 65) and 12 months
952 (right panels) (n = 65). Top panels present analyses based on ordinations from Euclidean distances of genus
953 level abundances after centered log ratio transformation and Euclidean distances of arcsine square root
954 transformed metabolite relative abundances. Bottom panel presents analyses based on generalized Unifrac
955 distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of arcsine square
956 root transformed metabolite relative abundances

957 **Figure S6.** Inter-omics Procrustes biplots comparing PCoA ordinations of untargeted metabolite bin relative
958 concentrations and taxonomic relative abundances in the sensitivity analyses for 6 weeks (left panels) (n = 65)
959 and 12 months (right panels) (n = 65). Top panels present analyses based on ordinations from Euclidean
960 distances of genus level abundances after centered log ratio transformation and Euclidean distances of arcsine
961 square root transformed metabolite relative abundances. Bottom panel presents analyses based on
962 generalized Unifrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean
963 distances of arcsine square root transformed metabolite relative abundances

964 **Figure S7.** Pairwise spearman correlation of concentration-fitted targeted metabolite concentrations and
965 genus-level taxonomic abundances for 6-weeks (panel A, N = 65) and 12-months (panel B, N = 65) infants in
966 sensitivity analyses. Left panel displays the overall correlation pattern, where non-significant correlations are
967 not colored (FDR controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and
968 metabolites selected by the sCCA procedure. Additionally, correlation coefficient of the first sCCA variate pair,
969 bootstrapped 95% confidence interval (nboot = 5000) and permutation p-value (nperm = 1000) are also
970 reported.

971 **Figure S8.** Pairwise spearman correlation of untargeted metabolite bin relative concentrations and genus-level
972 taxonomic abundances for 6-weeks (panel A, N = 65) and 12-months (panel B, N = 65) infants in sensitivity

973 analyses. Left panel displays the overall correlation pattern, where non-significant correlations are not colored
974 (FDR controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites
975 selected by the sCCA procedure. Additionally, correlation coefficient of the first sCCA variate pair,
976 bootstrapped 95% confidence interval (nboot = 5000) and permutation p-value (nperm = 1000) are also
977 reported

978 **Table S1.** Metabolites selected for targeted analysis and their potential biological functions.

Figures

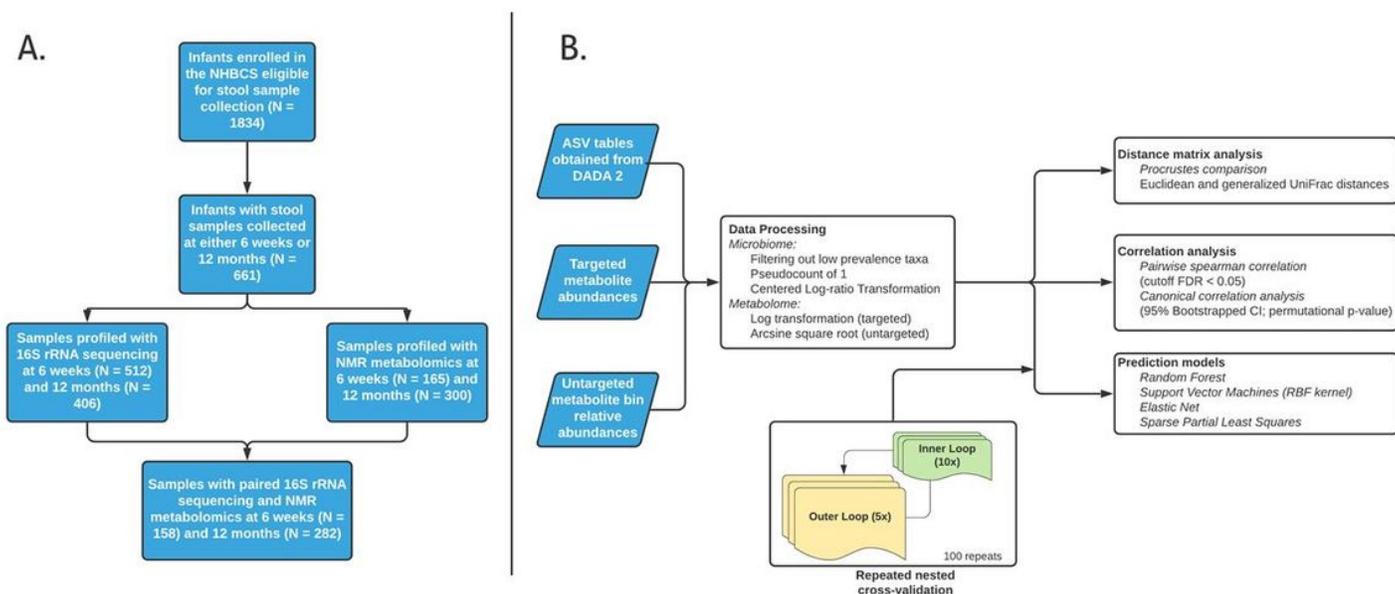


Figure 1

Overview of the analysis. Panel A describes the subject selection workflow and panel B describes the analytic pipeline.

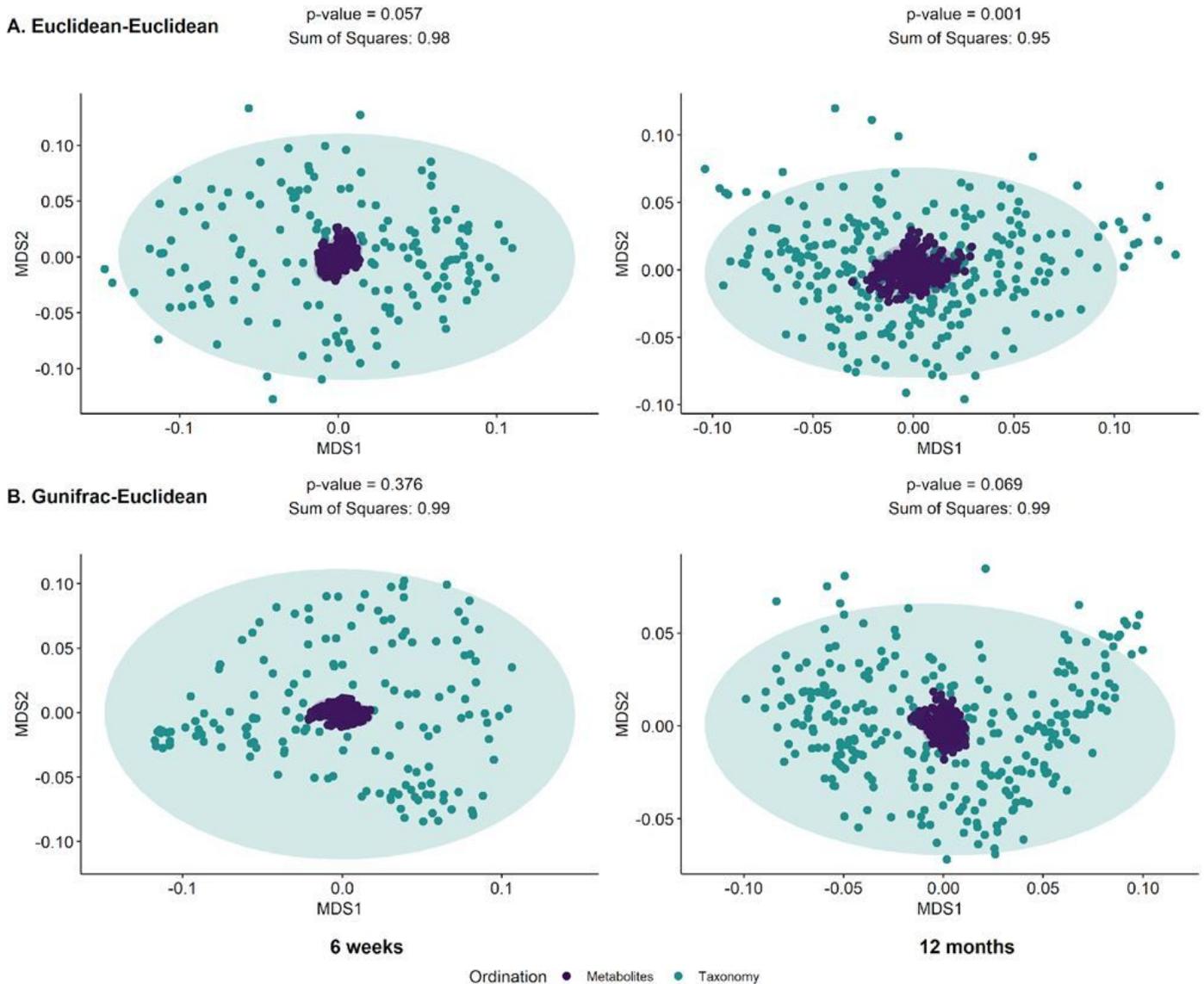
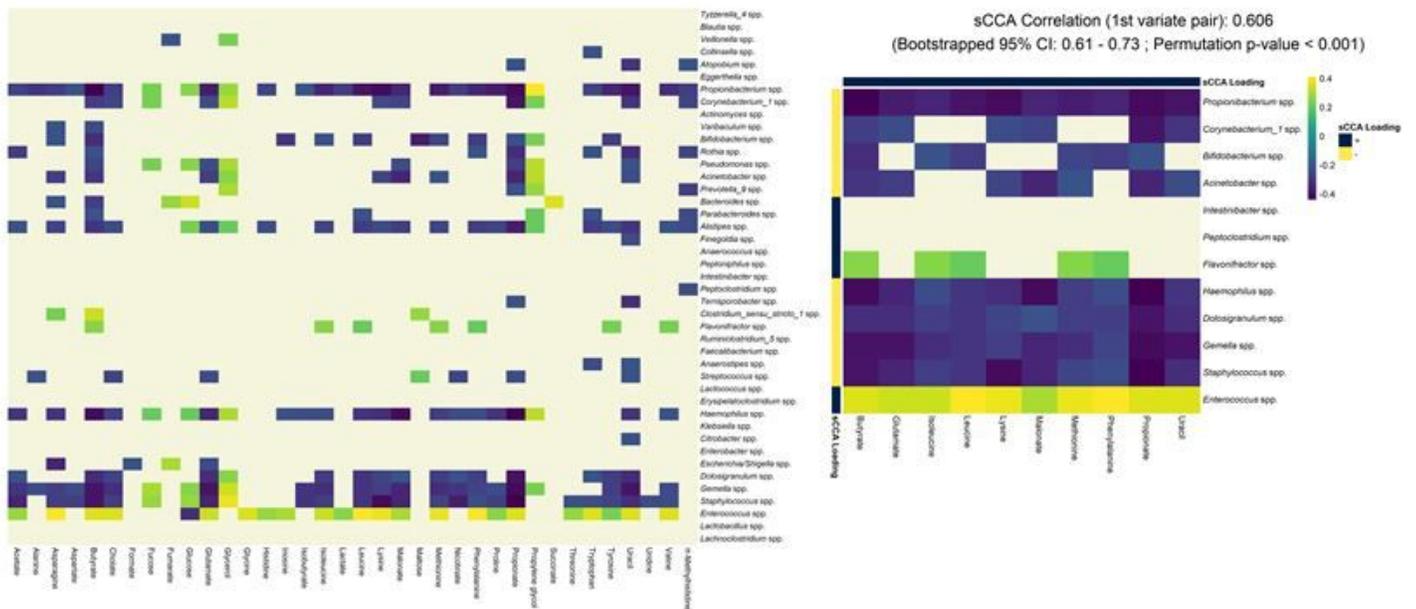


Figure 2

Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and taxonomic relative abundances for 6 weeks (left panels) ($n = 158$) and 12 months (right panels) ($n = 262$). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of log-transformed metabolite profiles. Bottom panel presents analyses based on gUniFrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of log-transformed metabolite profiles. There were significant associations between the microbiome and the metabolome (both targeted and untargeted) when utilizing Euclidean distances, however this association goes away when the gUniFrac distance was employed for the targeted metabolites only.

A. 6 Weeks



B. 12 Months

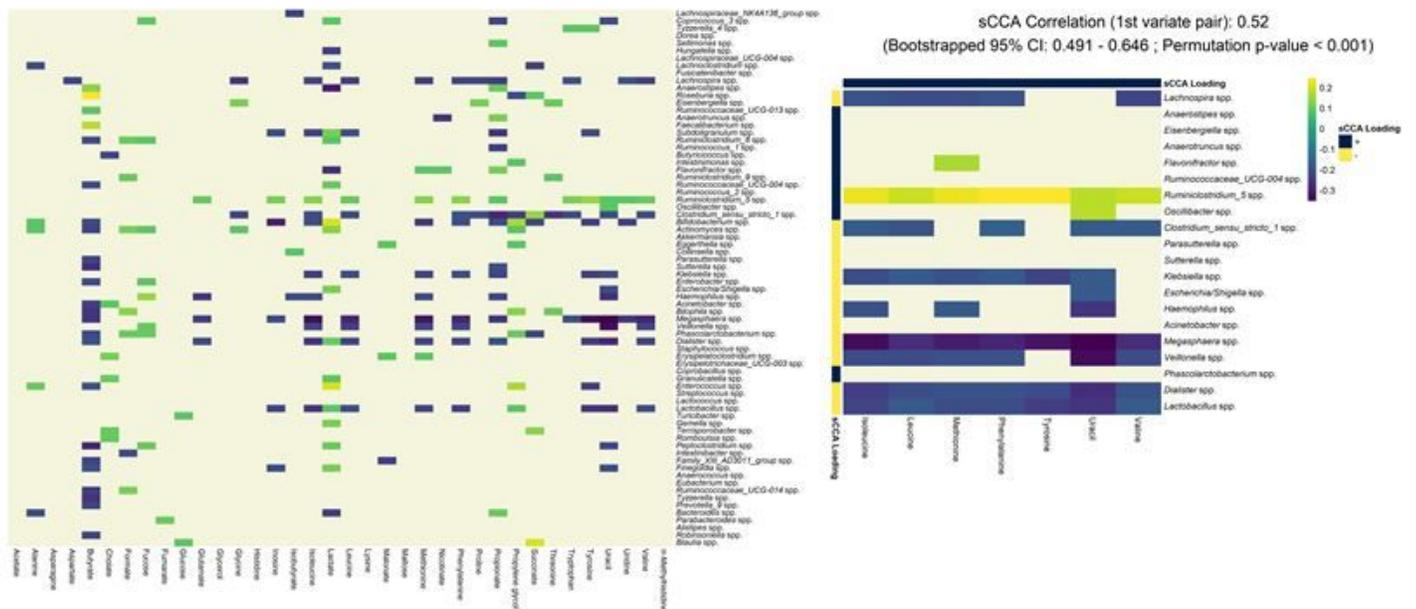
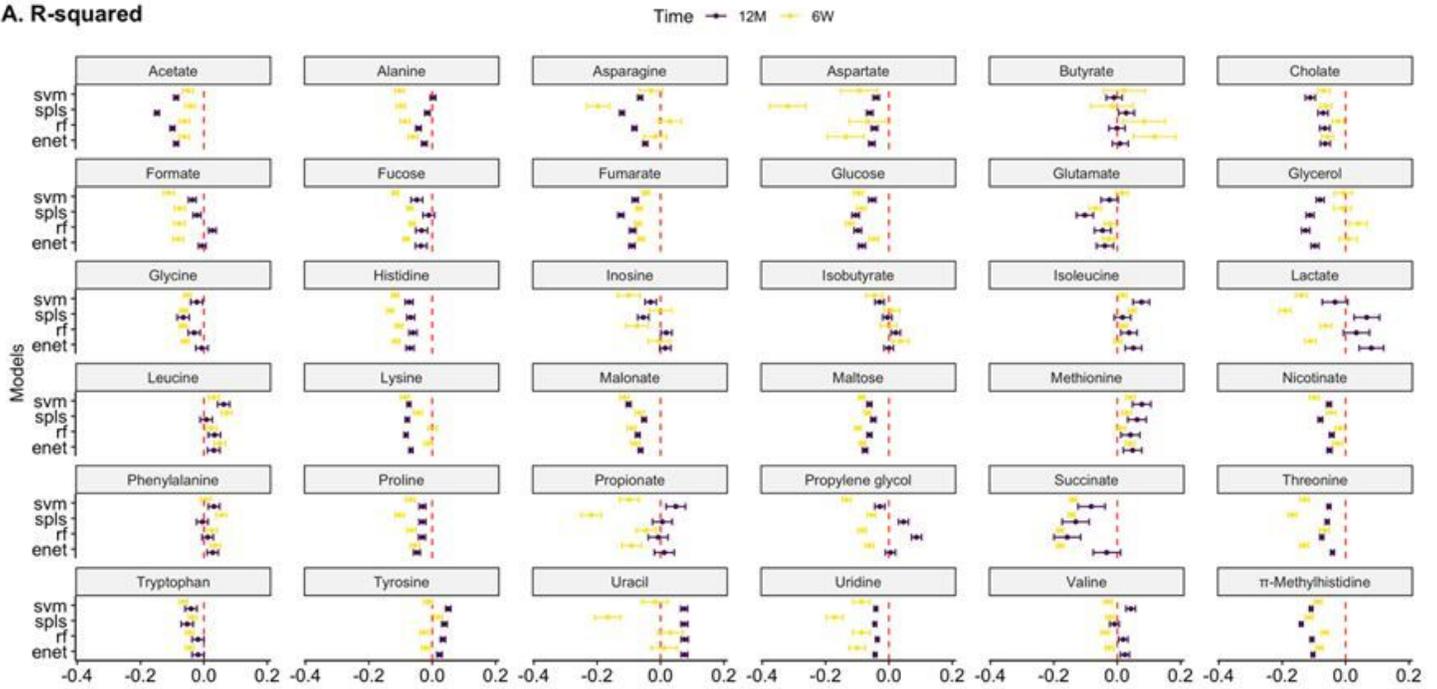


Figure 3

Pairwise Spearman correlation of concentration-fitted metabolites and genus-level taxonomic abundances for 6-weeks (panel A, N = 168) and 12-months (panel B, N = 282) infants. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (false discovery rate (FDR) controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sparse CCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval and permutation p-value are also reported. Significant microbiome-metabolome correlation was observed at both time points, however no significant difference was found between the time points.

A. R-squared



B. Spearman Correlation

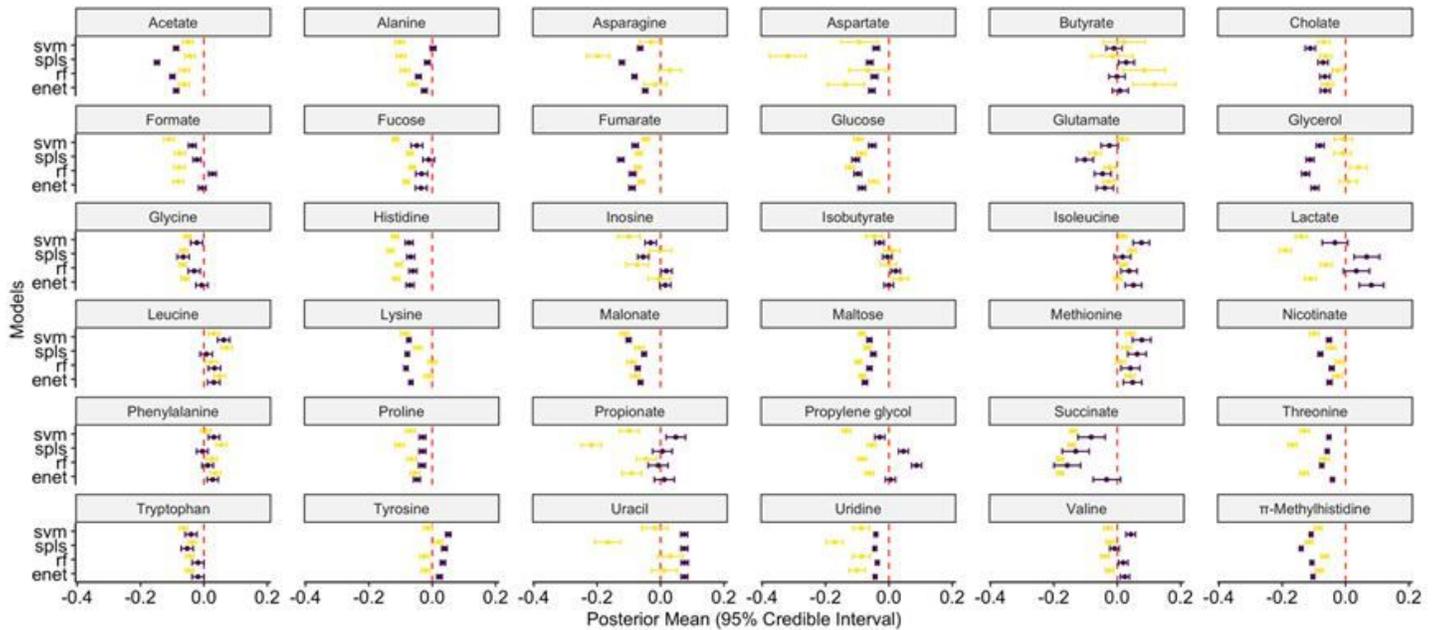


Figure 4

Forest plots of each prediction performance metric (R-squared – Panel A, Spearman correlation – Panel B) for each time point (6 weeks (n = 158), 12 months (n = 282)) across all 36 metabolites and 4 machine learning models. 95% credible interval and predictive posterior means were generated using Bayesian modelling of the evaluation statistic (Methods) after 100 repeats of 5-fold nested cross validation. Red vertical lines indicate a value of 0 for the evaluation metric (equivalent to null model). Metabolites were classified as predictable if the null value did not lie within the estimated 95% credible interval. For most metabolites, predictive performance was not significantly better than null models

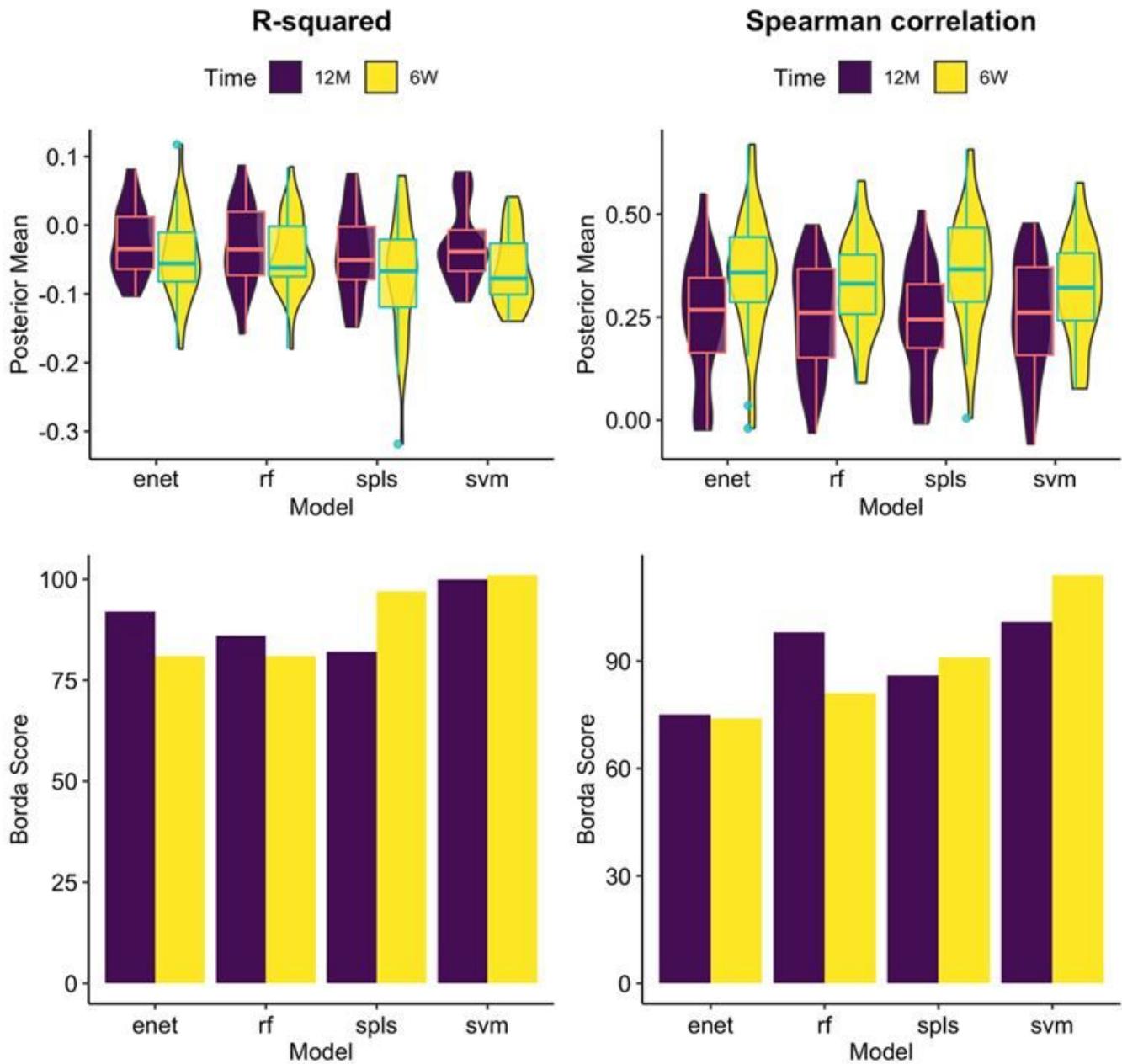


Figure 5

Comparative analysis predictive model performance across all metabolites in the targeted dataset for both 6-weeks ($n = 158$) and 12-months ($n = 282$) time points. Top panel shows superimposed boxplots and violin plots of the distribution of predictive posterior mean for each evaluation metric across all 36 metabolites. Bottom panels show aggregated model rankings for all metabolites using R-squared (left) and Spearman correlation (right) using Borda scores (Methods). Higher scores indicate that a model was consistently selected as a better performing. Relatively similar Borda scores and cross-metabolite average predictive performances indicate that no model was clearly the most performant. However, support vector machines (with radial basis function kernel) was highest scoring model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [additionalfile3.xlsx](#)
- [additionalfile2.xlsx](#)
- [additionalfile1.docx](#)