Defining and Evaluating Microbial Contributions to Metabolite Variation in Microbiome-Metabolome Association Studies

Cecilia Noecker[1], Hsuan-Chao Chiu[1,2], Colin P. McNally[1], and Elhanan Borenstein[1,3,4,*]

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[2] Current affiliation: Office of Chief Technology Officer, MediaTek Inc., Hsinchu City, Taiwan 30078
[3] Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
[4] Santa Fe Institute, Santa Fe, NM 87501, USA
* Corresponding Author (elbo@uw.edu)

Abstract

Correlation-based analysis of paired microbiome-metabolome datasets is becoming a widespread research approach, aiming to comprehensively identify microbial drivers of metabolic variation. To date, however, the limitations of this approach have not been evaluated. To address this challenge, we introduce a mathematical framework to quantify the contribution of each taxon to metabolite variation based on uptake and secretion fluxes. We additionally use a multi-species metabolic model to simulate simplified gut communities, generating an idealized microbiome-metabolome dataset. We then compare observed taxon-metabolite correlations in this dataset to calculated ground-truth taxonomic contribution values. We find that correlation-based analysis poorly identifies key contributors even in these idealized settings, with extremely low predictive value and accuracy. Importantly, however, we demonstrate that the predictive value of correlation analysis is strongly influenced by both metabolite and taxon properties, as well as exogenous environmental variation. We finally discuss the practical implications of our findings for interpreting microbiome-metabolome studies.

Importance

Identifying the key microbial taxa responsible for metabolic differences between individual microbiomes is an important step towards understanding and manipulating microbiome metabolism. To achieve this goal, researchers commonly conduct microbiome-metabolome association studies, comprehensively measuring both the composition of species and the concentration of metabolites across a set of microbial community samples, and then testing for correlations between microbes and metabolites. Here, we evaluated the utility of this general approach by first developing a rigorous mathematical definition of the contribution of each microbial taxon to metabolite variation, and then examining these contributions in a simulated dataset of microbial community metabolism. We found that standard correlation-based analysis of our simulated microbiome-metabolome dataset identifies true contributions with very low accuracy, and that its performance depends strongly on specific properties of both metabolites and microbes, as well as on the surrounding environment. Combined, our findings can guide future interpretation and validation of microbiome-metabolome studies.

Introduction

Microbial communities have a tremendous impact on their surroundings, ranging from the degradation of environmental toxins (1) to the production of climate change-relevant metabolites (2). Host-associated communities, in particular, have a substantial impact on their hosts, and often produce a diverse set of metabolites that interact with numerous host pathways. In humans, such microbiome-derived metabolites have been identified as contributing factors to a wide array of diseases including heart disease (3), autism (4), non-alcoholic fatty liver disease (5), colon cancer (6), inflammatory bowel disease (7), and susceptibility to infection (8). Characterizing the ways microbial communities modulate their environments and the relationship between community structure and metabolic impact is therefore a major, timely, and complex challenge with promising implications for human health, as well as to environmental stewardship, agriculture, and industry.

When facing this challenge, perhaps the most important task is identifying specific community members that drive variation in metabolites of interest. Taxa responsible for observed metabolic differences across communities may be ideal targets for interventions aiming to modify metabolic phenotypes. Their identification, however, can be a daunting task. Complex microbial communities are often composed of hundreds or thousands of poorly characterized species, each with a unique and frequently unknown complement of metabolic capacities. Even when multiple species are known to possess the potential to synthesize or degrade a metabolite of interest, the metabolic activity of each species (and consequently, its contribution to metabolic variation) may be different

70 (9). Moreover, community ecology, interspecies interactions, and nutrient availability
71 (e.g., via diet) can all regulate and influence the metabolic activity of each species,
72 rendering the link between community members and metabolic products extremely
73 complex and challenging to infer (10–12).
74
75 To address this challenge and to identify community members that play an important
76 role in metabolic variation, a growing number of studies are now comprehensively
77 assaying multiple facets of community structure across samples, including, most
78 notably, taxonomic and metabolite compositions (13). For example, to investigate the
79 links between taxonomic shifts and metabolic phenotypes in the healthy vaginal
80 microbiome and in bacterial vaginosis, a recent study used a combination of 16S rRNA
81 qPCR, sequencing, and both global and targeted metabolomics (14). Another study,
82 aiming to identify taxonomic and metabolic features of resistance and susceptibility to *C.*
83 *dificile* infection in the mouse gut similarly applied 16S rRNA sequencing and global
84 metabolomics (15). In another example, researchers characterized metabolic and
85 microbial features of periodontitis in the oral microbiome before and after treatment,
86 combining 16S rRNA sequencing, shotgun metagenomic sequencing, and
87 metabolomics (16). These are just a few examples of a plethora of recent microbiome-
88 metabolome studies, investigating the metabolic effects of microbiome variation in the
89 contexts of chronic and infectious disease, agriculture, precision medicine, nutrition,
90 fermented food science, and more (17–24). Such multi-omic studies are also a major
91 focus of several large-scale initiatives to study both host-associated and environmental
92 microbiomes (25, 26).

Given the taxonomic and metabolomic profiles obtained via such microbiome-metabolome assays, the vast majority of studies rely on simple univariate correlation-based analyses to link variation in community ecology to variation in metabolic activity (11, 14, 15, 27–30). Such analyses specifically aim to identify species whose abundance across samples is correlated with the concentration of metabolites, often assuming that highly significant correlations reflect a direct mechanistic link between the taxon and metabolite in question. These studies further regularly assume that positive correlations imply synthesis and negative correlations imply degradation, or that targeting the microbe in question could be used to modulate the concentrations of the metabolites with which it is correlated. For example, a recent study characterizing the microbiome and metabolome in Spleen-yang-deficiency syndrome (29) concluded that a positive correlation between Bacteroides and mannose likely resulted from extracellular degradation of mannan into mannose by that taxon. Similarly, a study of antibiotic perturbations to the microbiome and metabolome stated that the presence of several weak positive and negative correlations between genera and arginine supported the conclusion that arginine levels may be affected by many community members with high functional redundancy (27).

Yet, to date, the extent to which a correlation-based analysis effectively detects direct metabolic relationships between taxa and metabolites is unclear. Obviously, a strong correlation between the abundance of a certain species and the concentration of a metabolite across samples *could* reflect direct synthesis or degradation of the

metabolite by that species, but could also arise due to environmental effects, precursor availability, selection, random chance, or co-occurrence between species. Similarly, cross-feeding, external host processes, and varying enzymatic regulation can mask a correlation even when this species does in fact contribute to observed metabolite variation. Indeed, previous studies have suggested that microbe-metabolite correlations must have a high rate of false positives (31), and a recent experimental study pairing microbiome-metabolome correlation analysis with *in vitro* monoculture validations found anecdotally that several observed correlations were in fact false positives (32).

Importantly, two crucial challenges hinder a comprehensive and systematic evaluation of correlation-based analysis. The first is the lack of a rigorous general definition of a microbe's contribution to metabolite variability. While establishing the main taxonomic contributors to metabolite variation may be straightforward for specialized, well-characterized metabolites that are synthesized by just a single taxon, it can be much less clear for metabolites that can be synthesized (and/or degraded or modified) by many different taxa in the community. The second challenge is the absence of ground truth data on the nature of microbe-metabolite relationships. While limited data on the taxa driving metabolite shifts can be obtained from comparative mono- and co-culture studies (32–34), large-scale and comprehensive datasets that link species and metabolite abundances in the context of a complex community, for which the precise impact of each species on observed metabolite variation is known, are currently not available.

In this study, we address these two challenges, combining a framework for quantifying microbial contributions with a model-based simulated dataset. Specifically, we first introduce a generalizable and rigorous mathematical framework for decomposing observed metabolite variation and quantifying the contribution of each community member to this variation based on uptake and secretion fluxes. Second, we use a dynamic multi-species genome-scale metabolic model to simulate the metabolism of a set of simple microbial communities and to generate an idealized dataset of paired taxonomic and metabolomic abundances, with complete information on metabolite fluxes, microbial growth, interspecies interactions, and environmental influences. Applying our mathematical framework to this simulated dataset, we could then compare calculated contribution values to observed taxon-metabolite correlations and evaluate the ability of correlation-based analyses to identify key microbial contributors. We were additionally able to investigate factors that shape the relationship between community composition and metabolism in depth and to identify specific properties and mechanisms that impact the performance of microbiome-metabolome correlation studies.

Notably, given the objectives of this study, we intentionally focus on characterizing microbiome-metabolome relationship in a model-based, tractable, and well-defined setting. Indeed, our metabolic model may not perfectly capture the complex and diverse mechanisms that are at play in host-associated communities; however, considering the scope of this study, accurately modeling the metabolism of a specific community may not be crucial. Rather, for our analysis, we want our simulated data to recapitulate broad

trends observed in naturally occurring microbial ecosystems, as indeed has been observed in similar models (35–39). Moreover, utilizing this model-based approach allows us to dissect the relationship between community composition and metabolic phenotypes without the complexities inherent to *in vivo* communities (including spatial heterogeneity, measurement error, inter-microbial signaling, or strain-level variation), and with variation in the concentrations of environmental metabolites resulting exclusively from microbial metabolic activity. Analyzing the ability of a correlation-based analysis to detect true microbial drivers of metabolite variation in these simplified, best-case settings provides a baseline for the expected performances of such analyses in real microbiome-metabolome studies.

Results

Quantifying the impact of individual microbial species on variation in metabolite concentrations

In this study, we consider a microbial community as an idealized system, consisting of a population of multiple microbial species in a shared, well-mixed, biochemical environment. Each species uptakes necessary metabolites from the shared environment, performs a variety of metabolic processes to promote its growth, and secretes certain metabolites back into the shared environment. We additionally assume that certain nutrients flow into the environment and that microbial cells and metabolites are diluted over time. These processes can represent, for example, the inflow of dietary nutrients and the transit through the gut in the context of the gut microbiome. For

simplicity, we primarily consider a constant inflow and dilution rate, as in a chemostat setting. Accordingly, a microbiome-metabolome study can be conceived as analyzing a set of several such communities (at a certain point in time), each with a different composition of microbial species and correspondingly variable environmental metabolite concentrations. We focus initially on a controlled setting with identical nutrient inflow across all microbiomes, but later examine the impacts of differences in nutrient inflow between communities.

Given this setting, we first sought to establish a rigorous and quantitative framework for defining the impact of each microbial species (or any taxonomic grouping) in the community on the variation observed in the concentration of a given metabolite across community samples. We focused on species that *directly* modulate the environmental concentration of a given metabolite via synthesis or degradation, ignoring indirect effects via, for example, the synthesis of a precursor substrate that could impact the metabolic activity of other species. We noted that the total concentration of a metabolite in the environment can be represented as the sum of cumulative synthesis or degradation fluxes of this metabolite by each of the n species in the community, as well as cumulative environmental fluxes (e.g., total nutrient inflow and dilution). Formally, the metabolite concentration, M, can therefore be expressed as a sum of n dependent random variables m_i, where each m_i denotes the overall synthesis or degradation of the metabolite by each species, along with an additional random variable m_{env}, denoting the overall impact of environmental processes.

$$M = \sum_{i=1}^{n} m_i + m_{env}$$

As discussed above, when analyzing microbiome-metabolome datasets, the goal is often to identify taxa responsible for *changes* in the concentration of a metabolite of interest across a set of samples. Accordingly, here we wish to quantify the *contribution* of each species to the *variance* in the concentration of that metabolite across samples. Specifically, in the formulation above, *var(M)* depends on the variance in the constituent microbial and environmental factors, as well as the covariance between these components. This variance can then be linearly separated into *n+1* terms, representing the contribution of each species (and of any environmental nutrient fluxes) to the total variation in the metabolite:

$$var(M) = \sum_{i=1}^{n} c_i + c_{env}\ ;\ c_i = var(m_i) + \sum_{j \neq i} cov\left(m_i, m_j\right) + cov(m_i, m_{env})$$

If the nutrient inflow is constant across samples, its effect can be ignored and its contribution to the variance is 0. Additionally, in a chemostat setting, the dilution of each metabolite can be accounted for in the calculation of each contribution, as it depends strictly on the dilution rate and on previous metabolite concentrations (Methods). Finally, in order to compare species contributions across metabolites and to represents the relative share of the total variance of a given metabolite that is attributable to species *I*, we defined the *relative* contribution to variance $\hat{c}_i$ of each species *i* to metabolite *M* by

normalizing contribution values by the metabolite's total variance:

$$\hat{c}_i = \frac{c_i}{var(M)}$$

This framework for calculating microbial contribution values provides a systematic measure of the causal impact of each taxon on observed variation in the environmental concentration of each metabolite, distilling the effect of complex ecological and metabolic interactions to a concise and interpretable set of quantities. Moreover, the obtained contribution profile is a linear decomposition of observed metabolic variation, wherein the sum of contributions of all species equals the observed variation in the metabolite. Notably, when a species' activity has large negative covariances with the activities of other community members, contribution values can be negative. Such negative contribution values indicate that a species' secretion or uptake of that metabolite varies in a way that mitigates the activity of others. Correspondingly, contribution values can be greater than 1, reflecting scenarios in which a species in fact generates more variation of this metabolite than is ultimately observed, but that its impact is mitigated by other species.

It is also worth noting that our analytical decomposition of contributions to variance is mathematically equivalent to calculating the Shapley values for the variance in metabolite concentrations (see Methods and Figure S1). Shapley value analysis is a game theory technique that defines an individual's contribution to a collective outcome, and has been shown to be the only general definition that is efficient, linear, symmetric,

252 and assigns zero values to null contributors (40). A similar, Shapley value-based
253 approach was recently applied to address the related problem of identifying the primary
254 taxonomic contributors to differential functional abundances in metagenomic data (41).
255

256 ***A multi-species metabolic model for generating complex microbiome-***
257 ***metabolome data***

258 We next set out to generate a large-scale dataset of microbiome-metabolome profiles
259 with complete information about metabolite uptake and secretion fluxes. To this end, we
260 used a multi-species metabolic model to simulate the growth, dynamics, metabolism,
261 and environment of a simple microbial community. This model is based on a previously
262 introduced genome-scale framework for modeling the metabolism of multi-species
263 communities and for tracking the metabolic activity of each community member over
264 time (42, 43). Briefly, this framework assumes that each species optimizes its growth
265 selfishly given available nutrients in the shared environment and predicts the metabolic
266 activity for each species in short time increments using Flux Balance Analysis (44). After
267 each increment, the model uses the predicted metabolic activities of the various species
268 to update the biomass of each species and the concentration of metabolites in the
269 shared environment (hence, potentially impacting the growth and metabolism of other
270 species in subsequent time steps). Full details of this model and simulation parameters
271 can be found in the Methods.
272

273 We specifically modeled a simplified gut community that was previously explored
274 experimentally (45). This community includes 10 representative gut species, spanning

275 the major clades found in the human gut and collectively encoding the key metabolic
276 processes taking place in this environment, including breakdown of complex dietary
277 polysaccharides, amino acid fermentation, and removal of fermentation end products
278 via sulfate reduction and acetogenesis. Genome-scale metabolic models of these 10
279 species were obtained from the AGORA collection (38) – a recently introduced set of
280 high-quality gut-specific metabolic models. To mimic the experimental gnotobiotic
281 mouse setting (45), we simulate growth in a chemostat, with a nutrient inflow mimicking
282 the content of a standard corn-based mouse chow, and a dilution rate consistent with
283 mouse transit time and gut volume. While maintaining this nutritional environment, we
284 systematically explored the landscape of possible community compositions, varying the
285 initial relative abundance of each species from 10% to 60% (with a consistent total
286 abundance equal to the community carrying capacity), resulting in a total of 61 different
287 community compositions. For the analysis below, we simulated growth for 144 hours (as
288 576 15-minute time steps). For most community compositions considered, this
289 simulation time consisted of an initial stabilization period followed by a period of near-
290 steady-state equilibrium (Figure 1A). Notably, across the various simulations, some
291 species maintained high abundances throughout the course of the simulation, while
292 others reverted to lower levels.
293
294 Throughout the course of each simulation, we recorded the abundances of each
295 species, the secretion and uptake rate of each metabolite by each species (as well as
296 internal reaction fluxes), and the concentration of each metabolite in the environment
297 (Figure 1A-B), thereby obtaining a comprehensive dataset describing species

298 composition, metabolic activities, and metabolite concentrations across 61 different
299 communities. To mirror the typical structure of a microbiome-metabolome cross-
300 sectional dataset, we specifically considered the abundances of species and the
301 concentrations of metabolites in the environment at the end of each simulation (i.e.,
302 after the final time point; see Figure 1). 60 of the 68 metabolites present in the nutrient
303 inflow (46) exhibited at least some variation across communities, as did 18 additional
304 microbially-produced metabolites. Metabolite variation was generally low (median
305 coefficient of variation 0.021), reflecting a relatively stable nutrient environment, yet 25
306 metabolites (32%) did have a coefficient of variation greater than 0.1. For downstream
307 analysis, we excluded metabolites without substantial measurable variance across
308 samples, filtering those with variance at or below the 25th percentile. This resulted in a
309 dataset of 52 variable metabolites, of which 14 are purely microbially-produced
310 metabolites, 9 are microbially-produced but also present in the nutrient inflow, and 29
311 are introduced only through the nutrient inflow. Of these 52 variable metabolites, 47 are
312 utilized by any member of the community (including 18 that are cross-fed in at least one
313 simulation). The final species compositions and the final concentrations of several key
314 metabolites across all simulations are shown in Figure 2A-F.
315
316 Exploring this dataset, we found that species composition and metabolite
317 concentrations exhibited complex patterns and biologically reasonable distributions
318 (Figure S2) (47). Several metabolic processes known to occur in the mammalian gut
319 were replicated by our simulations, including, for example, conversion of acetate to
320 butyrate by *E. rectale* (46), and production of key microbial metabolites such as 4-

321 aminobutyric acid (GABA), indole, and succinate. Cross-feeding relationships were
322 observed frequently (18 metabolites), including cross-feeding of 6 amino acids, whose
323 exchange is widespread in host-associated microbiota (48).
324
325 Clearly, the model and simulations described above represent a gross simplification of
326 the microbiome's structure, dynamics, and function. Importantly, however, this
327 simplification is also an important strength. Specifically, the data obtained from these
328 simulations provide a unique opportunity to examine the relationship between
329 community dynamics and metabolic activity in a realistic, yet tractable model of
330 community metabolism where complete information about the activity and fluxes of each
331 microbial species is available (Figure S3). Indeed, our multi-species model captures
332 many of the intricacies of bacterial genome-scale metabolism and the
333 interconnectedness (both within and between species) of multiple metabolic processes,
334 yet without additional complexities inherent to *in vivo* communities. Furthermore, in our
335 simulations, variation in the concentrations of environmental metabolites results
336 exclusively from microbial metabolic activity, with no variation in nutrient inflow or other
337 non-microbial sources, providing a controlled setting for evaluating the relationship
338 between community members and metabolite concentrations.
339

340 ***Metabolite variation is driven by diverse microbial mechanisms***

341 Given the simulated dataset described above (for which uptake and secretion fluxes are
342 known), we applied our contribution framework to calculate the contribution of each
343 species to the variation observed in each of the 52 variable metabolites (Figure S4).

344 The resulting contribution values can be used as ground-truth information about the link
345 between microbial activity and environmental metabolites.
346
347 To highlight the nature and utility of such contribution values, and to demonstrate how
348 metabolic fluxes translate into contribution profiles, we first describe our results for
349 several example metabolites (Figure 2). Putrescine, an amino acid fermentation
350 product, is an example of the simplest case, in which one microbial species – *E. coli* –
351 synthesizes a metabolite that is not utilized or modified by other community members.
352 Variation in the environmental concentration of putrescine was hence fully determined
353 by the level of secretion from *E. coli*, which is therefore assigned a relative contribution
354 of 1 (Figure 2B). Tetradecanoic acid, in contrast, was introduced (at a constant rate) via
355 the nutrient inflow and utilized by the three *Bacteroides* species in the community to
356 varying degree (primarily by *B. ovatus* and to a slightly lesser extent by *B.*
357 *thetaiotaomicron*). The calculated contribution values successfully attributed variation in
358 the environmental concentration of this metabolite to these three species, and correctly
359 captured the difference in the magnitude between their effects (Figure 2C). Variation in
360 uracil, another metabolite introduced via the nutrient inflow, was mainly driven by large
361 shifts in its uptake by *B. ovatus,* but this effect is partially masked by *E. rectale,* which
362 reduced its uptake when *B. ovatus*' flux was high and vice versa. Other species also
363 utilized uracil, but at relatively similar levels across samples, and accordingly with
364 relatively little impact on its variation. These complex patterns were all captured by the
365 contribution profile obtained by our framework, with *B. ovatus* assigned a high positive
366 contribution, *E. rectale* assigned an intermediate *negative* contribution, and other

species assigned relatively negligible contribution values (Figure 2D). More complex species-metabolite relationships were also accurately and effectively summarized. Contribution values for acetate, for example, reflected the cross-feeding interactions that underlie variation in its concentration (Figure 2E). It was introduced to the shared environment by several species (primarily *C. symbiosum*), but most of its variation ultimately depended on the level of uptake by *E. rectale.* Finally, the contribution profile of succinate demonstrates how extremely strong interspecies interactions can produce contribution values much greater than the observed variance (Figure 2F). In the simulated data, this metabolite was synthesized by *B. hydrogenotrophica,* but was almost always fully utilized by other community members. The calculated contributions suggest that if the synthesis of succinate by *B. hydrogenotrophica* would not have been offset by uptake from other species, the variance in succinate concentration across samples would have been 71.7 times higher than is actually observed. (Note that the difference between positive and negative is always 1.)

Examining the complete set of variable metabolites and calculated contribution values revealed similar patterns of interactions (Figure S4). Specifically, as for the metabolites discussed above, negative contributions and/or contribution values greater than 1 were widespread. Nearly all metabolites (50 out of 52) had at least one species with a negative contribution value, and 36 had at least one species with a contribution value greater than 1. Of the 32 other metabolites with negative contributions, 29 were present in the nutrient inflow and their negative contributions result from competition between species for their uptake. This prevalence of negative and extreme values suggests that

390 strong negative interspecies interactions have substantial impacts on metabolite
391 concentrations, and that often, observed variation in a given metabolite's concentration
392 is the complex outcome of multiple species generating and offsetting much higher
393 variation.
394
395 It is also important to note that while the average metabolic uptake/secretion flux of
396 each species and the magnitude of its contribution to a given metabolite were generally
397 significantly correlated (Spearman, $p < 0.01$ for 49 of the 52 metabolites), the species
398 with the highest flux was often *not* the largest contributor to variation (26 of the 52
399 metabolites). Similarly, the variance in a species' flux was significantly correlated with its
400 contribution for 48 of the metabolites, but for 9 metabolites the species with the most
401 variable flux was still not the largest contributor (due to differences in whether variable
402 flux generated by one species is compensated by variation in the flux of another). These
403 findings suggest that even if the magnitude and variation of species uptake and
404 secretion fluxes across a set of microbiome samples are known (rather than just the
405 abundances of species, which is the only measure usually assayed), metabolic
406 interdependence between species would still make true contributor species challenging
407 to identify.
408
409 Combined, the observations above highlight the complex relationship between species
410 activity and measured metabolite concentrations, demonstrating the important role of
411 both direct and indirect species interactions. This complex relationship, observed even
412 in the idealized settings of our simulation model, is potentially markedly more complex

413 than what is assumed by many microbiome-metabolite association-based analyses.

414

415 ***Correlation analysis fails to detect true microbial contributors to metabolite***

416 ***variation***

417 Given our observations above, we next set out to comprehensively assess how well
418 pairwise correlation analysis (commonly used for analyzing microbiome-metabolome
419 data) can detect true taxonomic contributors to metabolite variance. Put differently, we
420 evaluated the extent to which a correlation between species abundance and metabolite
421 concentration across samples captures the true causative contribution of a species'
422 metabolic activity to observed metabolite variation.

423

424 Following numerous microbiome-metabolome studies (14, 23, 28, 49), we considered
425 identifying species-metabolite relationships as a classification task, aiming to identify for
426 each metabolite the set of species that are primarily responsible for the variation
427 observed in its concentration across samples. To this end, we defined *key contributor*
428 species for each metabolite as those with a contribution value greater than 10% of the
429 total positive contribution values. This resulted in a set of 83 species-metabolite key
430 contributor pairs, representing true links between species activity and metabolite
431 variation. On average, each metabolite had only 1.6 contributors (Figure S5), although
432 7.5 species on average had utilized or synthesized each metabolite at any point. 31.3%
433 of these contributions occurred via synthesis reactions, 66.3% via utilization, and 2.4%
434 (2 instances) via both processes. We then calculated the Spearman rank correlations
435 between species abundances and metabolite concentrations across samples, and used

436 a *p*-value threshold of 0.01 to define significant correlation between species and
437 metabolites. This produced a set of 191 significant species-metabolite correlations,
438 representing putative species-metabolite links.
439
440 Comparing this set of significant species-metabolite correlations to the set of species-
441 metabolite key contributors clearly illustrated the difficulty of using univariate
442 associations to infer mechanistic contributions (Figure 3). Indeed, of the 191 significant
443 species-metabolite correlations, the vast majority (141) were false positives
444 (corresponding to a positive predictive value of only 26.2%), and did not represent true
445 contributor relationships (Figure 3A). Moreover, more than a third of these false positive
446 species-metabolite pairs (51 out of 141) had *no* mechanistic connection; i.e., the
447 species did not ever use or produce the metabolite in question. Furthermore, for 12
448 variable metabolites (out of 52), none of the key contributors were successfully detected
449 by a correlation analysis. The overall accuracy was somewhat higher (66.5%), reflecting
450 the high number of non-contributors that are also not correlated. Using a stricter cutoff
451 ($p < 0.0001$, equivalent to a Bonferroni-corrected value of 0.05) only improved the
452 positive predictive value to 33% and the accuracy to 77.1%. Indeed, a ROC curve
453 analysis (Figure 3B) produced an area under the curve of 0.72, and overall correlations
454 and scaled contribution values were only weakly associated (Figure 3C), suggesting
455 that these findings can only be partially mitigated by changing classification thresholds.
456 Metabolites of different classes had generally similar correspondence between
457 correlations and contributions (Figure 3D).
458

459 Notably, key contributors for purely microbially-produced metabolites were not identified
460 more accurately than those for metabolites in the nutrient inflow (66% versus 67%),
461 which is perhaps not surprising since we used a constant inflow across samples (but
462 see also our analysis below with variable inflow). Across species, contributions were
463 identified most accurately for *D. piger*, which had a relatively low number of
464 contributions (Figures 3E and S5C), but the positive predictive value was nonetheless
465 <50% for all species.
466
467 Using an alternative classification task, aiming to detect all microbes that affect variation
468 in a given metabolite across samples regardless of whether their effects are ultimately
469 reflected in the observed concentrations, provided qualitatively similar findings
470 (Supplementary Text, Figure S5).
471
472 ***Species and metabolite properties explain discrepancies between correlations***
473 ***and contributions***
474 Our analysis above demonstrated that correlations between species abundances and
475 metabolite concentrations can often be only poorly associated with true contribution of
476 species to metabolite variation. We therefore next investigated the origins of such
477 discrepancies, seeking to identify factors that lead to a significant species-metabolite
478 correlation when the species in fact does *not* contribute to that metabolite variation (i.e.,
479 false positives), and factors that mask such correlation when the species *does* in fact
480 contribute to this metabolite variation (i.e., false negatives).
481

482 To determine whether the identity of the species or metabolite in question can explain
483 inaccurate identifications of key contributors, we used a regression-based analysis.
484 Specifically, we considered all species-metabolite non-contributor pairs, and fitted a
485 logistic regression model to predict whether a species-metabolite pair exhibited
486 significant correlation (false positive), based on either species identities, metabolite
487 identities, or both (Methods). We then compared these three models using a likelihood
488 ratio test to assess which of these features (i.e., species or metabolite identities) are
489 informative. We similarly considered all species-metabolite key contributor pairs
490 separately, again fitting a logistic regression model based on species identities,
491 metabolite identities, or both to predict whether a pair failed to exhibit significant
492 correlation (false negative).
493
494 For non-contributors, we found that false positives can be explained largely by species
495 identity (likelihood ratio test (LRT) for inclusion of species terms $p < 10^{-13}$). Incorporating
496 both species and metabolite identities did not significantly improve the model (LRT for
497 metabolite terms p=0.72). This finding suggests that false positives – correlations
498 observed between species and metabolites to which they in fact did not contribute – are
499 the outcome of interactions at the species level, regardless of the metabolite in
500 question. This impact of strong interactions between features has been described
501 extensively in other data types (50, 51). Indeed, examining the 141 false positives
502 identified above, we found that many can be explained by the relationships between the
503 three dominant species in this community: *E. rectale*, *B. thetaiotaomicron*, and *B.*
504 *ovatus*. These species competed strongly for carbon sources (and utilized their

505 maximum allocation of sucrose, glucose, and fructose at nearly every step of the
506 simulation), and their abundances were therefore negatively correlated. As a result,
507 metabolites that varied due to the activity of one of these species were also frequently
508 correlated with the other two. In total, 32 false positive correlations paired one of these
509 species with a metabolite for which another species in this trio was a key contributor.
510 More generally, we found that the probability of a false positive correlation for a
511 particular species and metabolite depended on the species' correlation with the true key
512 contributors for that metabolite (p=0.006, Spearman rho between share of false
513 positives and interspecies correlation; Figure 4A). Moreover, the maximum correlation
514 each species had with any other species is a strong predictor of its overall specificity,
515 which varies widely from 33.3% for *E. rectale* to 92% for *D. piger* (Spearman rho=-0.84,
516 p=0.002).
517
518 In the case of key contributors, we found that false negative correlations can be
519 explained largely by metabolite identity (LRT for metabolite terms p=0.002; although the
520 species involved was also somewhat informative with LRT p=0.08). Put differently, a
521 lack of correlation between the abundance of a key contributor species and the
522 concentration of the metabolite to which it contributed was determined mainly by the
523 nature of the metabolite in question. This lack of correlation between a given metabolite
524 and its contributors could have resulted from competition or exchange of a metabolite
525 between multiple species, such that none of the involved species end up strongly
526 associated with the final outcome on their own. Indeed, across all metabolites, the
527 average correlation between a metabolite and its key contributors is negatively

528 associated with its number of key contributors (Spearman rho=-0.45, *p*=0.0008). The
529 number of key contributors for any metabolite was also thus negatively associated with
530 the sensitivity of contributor detection for that metabolite (Spearman rho=-0.48,
531 *p*=0.0004; Figure 4B). We further hypothesized that false negative outcomes might be
532 more common for metabolites with more or larger negative species contributions, since
533 these, by definition, mask or compensate for the activity of key contributor species.
534 While all metabolites with a false negative outcome did have at least one species with a
535 negative contribution value, as mentioned above, this was true for nearly all analyzed
536 metabolites (50/52), and the number of negative contributing species was not
537 associated with the occurrence of a false negative correlation (*p*=0.86, Wilcoxon rank
538 sum test). Moreover, we also did not observe any effect of the average concentration of
539 a metabolite on the sensitivity and accuracy of its detection via correlation analysis, nor
540 of whether it is secreted, utilized, or cross-fed (Figure 4C). In summary, our analysis
541 suggests that the largest factor explaining whether a metabolite's key contributor can be
542 detected by a correlation analysis is simply whether there are other community
543 members (key contributors) that also impact the observed concentration of that
544 metabolite.
545

546 ***Environmental fluctuations in metabolite concentrations impact detection of key***
547 ***contributors***

548 Our analyses above all focused on a single simulated dataset in which the nutrient
549 inflow was constant across all samples, meaning that metabolite variation was fully
550 governed by microbial activity. However, in reality, metabolite variation can and does

551 arise also from non-microbial sources, potentially affecting both the landscape of key
552 microbial contributors and our ability to detect them via correlation-based analyses. To
553 explore the impact of environmental fluctuations, we therefore ran several sets of
554 additional simulations designed to emulate experimental settings with varying degrees
555 of nutrient fluctuation. In the case of the gut microbiome, such fluctuations can
556 represent, for example, changes in diet composition, host consumption or absorption, or
557 other non-microbial processes. In these simulations, we maintained the same set of 61
558 initial species compositions but added small amounts of stochastic noise to the nutrient
559 inflow, sampling inflow concentrations for each compound in each simulation from a
560 normal distribution with a mean equal to the compound's original inflow rate and a
561 standard deviation ranging from 0.5% to 10% of the mean in 8 increments (Methods).
562 For each of the resulting 8 datasets, we again calculated contribution values (with the
563 added element of the nutrient inflow as a potential contributor to variance), identified key
564 contributors, and compared them with the results of a correlation analysis.
565
566 Examining the obtained contribution values, we found, as expected, that variation in
567 inflow quantities can outweigh the variation in microbial fluxes, and that as the variation
568 in inflow increases, its contribution to metabolite variation increased at the expense of
569 the contributions of community members (Figure 5A). As a result, the number of key
570 contributions attributed to each species decreased for metabolites in the nutrient inflow
571 (Figure 5B). Interestingly, however, some species lost their contributions more gradually
572 than others, and in some cases even became key contributors for additional metabolites
573 (Figure 5B).

We next examined how correlation-based detection of key microbial contributors was affected by these inflow fluctuations. We assigned each of the 52 metabolites in each of the 9 datasets (the original dataset with no inflow fluctuations and the 8 datasets with varying degree of fluctuations) to bins according to the level of contribution attributed to the inflow for this metabolite at that degree of fluctuation (see Methods). We then evaluated the performance of correlation analysis for each bin separately. The share of true key contributors naturally decreased rapidly with increasing environmental contribution, as did the number of significantly correlated species-metabolite pairs (Figure 5C). Importantly, however, the sensitivity of correlations decreased substantially with the level of contribution attributed to the inflow, but the specificity in fact increased from 67.7% to 96.2% (Figure 5D). This suggests that while environmental fluctuations disrupted the signal linking microbial species with the metabolites they impact, they also disrupted indirect associations between species and metabolites (false positives). Overall, however, the AUC did not change significantly with increasing environmental contribution (Figure S6A), and the positive predictive value is similarly relatively stable (and never rose higher than 41%). Interestingly, the detection of some metabolites not present in the inflow was also affected by inflow fluctuations in a similar manner (Supplementary Text, Figure S6B).

Discussion: Insights and implications for microbiome-metabolome analyses

Above, we have investigated the ability of correlation-based analyses to detect key microbial contributors responsible for variation in metabolite concentrations across samples. Our findings suggest that microbe-metabolite correlation analysis may be a useful approach for exploratory analyses, but highlight some of the limitations and caveats of such microbiome-metabolome studies and identified several factors that impact the relationship between community composition and metabolite concentrations. Below, we elaborate on a set of practical conclusions and their implications for the analysis and interpretation of microbiome-metabolome studies.

Association-based analyses of microbiome-metabolome assays have low predictive value for detecting direct species-metabolite relationships and require conservative interpretation. Microbiome-metabolome association studies have been previously proposed as a powerful tool for the identification of causal mechanisms of microbiome metabolism (52), and indeed, such studies often present detected associations as evidence for mechanistic relationships (11, 27, 29). However, our analysis suggested that the positive predictive value of significant species-metabolite correlations for identifying true microbial contributors can be extremely poor (less than 50% across all settings, and as low as 10% in the context of large environmental fluctuations). Recent experimental studies pairing microbiome-metabolite correlation analysis with *in vitro* monoculture validations have similarly observed many false

618 positive correlations (32). Additionally, given the somewhat low sensitivity observed in
619 our analysis, a lack of association is not necessarily sufficient to reject a hypothesis that
620 a particular microbial taxon impacts a particular metabolite. Identified correlations
621 between microbial taxa and metabolites should therefore be interpreted very
622 conservatively and used mostly to prioritize microbe-metabolite relationships for follow-
623 up validation studies (e.g., via culture-based studies or germ-free model organism
624 colonization). One potential approach for improving the predictive value of such
625 correlation-based analyses is to examine whether they replicate across multiple
626 conditions. Indeed, we found that a correlation does provide stronger evidence for a
627 contributor relationship if it persists across several different contexts. Across our 9
628 simulated datasets with varied environmental fluctuations, the 41 species-metabolite
629 pairs that were significantly correlated in every dataset were 2.2 times more likely to
630 denote true key contributor relationships than other significant correlations (Fisher exact
631 test, p=0.03), although their positive predictive value was still relatively low (39.5%).
632
633 **The predictive power of correlation-based analysis is species-, metabolite-, and**
634 **context- dependent.** In our dataset, metabolites varied widely in both contribution
635 profiles and in their detectability via correlation analysis. In particular, the key
636 contributors for metabolites acted upon by fewer species were identified more readily.
637 Correlation analysis may thus identify microbes involved in specialized secondary
638 metabolic processes (e.g. products of complex biosynthetic pathways) more readily
639 than those involved in more widespread processes. Therefore, correlation-based
640 approaches may be more informative for analyzing compounds that are specific to a

641 small number of taxa, but accurate dissection of the taxa controlling variation in widely-
642 trafficked metabolites may require more detailed analysis and experimentation.
643 Similarly, we found that species-metabolite correlations for species that are strongly
644 associated with other taxa (e.g., those with tight interactions with other community
645 members) are often spurious, suggesting that such correlations should be regarded less
646 confidently.
647
648 **External metabolic fluctuations can strongly impact the detection of microbial**
649 **contributions.** Our analysis of the impact of environmental fluctuations suggested that
650 the presence of environmental variability from a diverse set of samples could in fact
651 increase correlation specificity. We also found that the sensitivity of correlation analysis
652 rapidly decreased with increasing environmental fluctuations (from 60% to 9%). These
653 observations suggest that while a strictly controlled environment (e.g., a fixed diet) is
654 intuitively expected to increase the strength of microbiome-metabolome studies, its
655 value depends on the study priorities. Specifically, if the goal is to identify clear-cut
656 microbial drivers of healthy- and disease-associated metabolite shifts, the presence of
657 environmental variability could be beneficial as it may reduce the rate of false positive
658 associations. In contrast, for studies searching for a particular microbial taxon's
659 involvement in a particular process (e.g. aiming to determine whether an ingested
660 probiotic impacts aspects of gut metabolism), a more controlled environment may be
661 favorable. It should however be noted that are findings were based on environmental
662 fluctuations that were uniform and independent, which may not hold for real-life
663 environmental fluctuations such as diet variation. It is also worth noting that in our

664 simulations, microbial fluxes for some environmental metabolites could be drowned out
665 by as little as 0.5% variation in nutrient inflow quantities, while others still had
666 substantial microbial contributions even with 10% variation in inflow. When interpreting
667 an observed association, the scale of possible microbial variation relative to external
668 variation should therefore be taken into account.
669
670 **Mechanistic reference information can improve the predictive power of**
671 **microbiome-metabolome studies.** In our simulated dataset, 36% of the false positive
672 correlations occurred between a metabolite and a species that was in fact not capable
673 of uptaking or secreting that metabolite. Ruling out such falsely detected links would
674 substantially improve the positive predictive value of a correlation-based analysis. One
675 approach for doing so is by utilizing genomic information, which can be obtained or
676 predicted for many microbial taxa (53). By coupling such genomic information with
677 metabolic databases such as KEGG or MetaCyc (54, 55), researchers can filter out
678 correlation-based links that are likely not feasible causative relationships. Further
679 improvement can be obtained by integrating such reference information directly into the
680 analysis. Indeed, we previously introduced a computational framework, termed
681 MIMOSA (56), that utilizes a simple community-wide metabolic model to assess
682 whether measured metabolite variation is consistent with shifts in community metabolic
683 potential, and to identify potential contributing taxa. MIMOSA has been applied to varied
684 host-associated microbiomes from varied body sites and from human and mouse hosts
685 (12, 57, 58). Applying MIMOSA to the simulated species-metabolite dataset analyzed
686 above (Methods), we found that it indeed identified key contributors significantly more

accurately than a correlation-based analysis, with an AUC of 0.89 (Figure 6). Notably, in this analysis, we assumed MIMOSA has access to the correct set of metabolic reactions possessed by each species. Using standard less-complete information obtained directly from the KEGG database (as done regularly when using this tool) reduced the number of metabolites that could be analyzed from 52 to 39, with improved specificity (96%) and positive predictive value (61%) and an ultimately comparable AUC (0.74). Combined, these findings suggest that reference model-based approaches can provide stronger evidence for mechanistic relationships than strictly correlation-based methods, but their use depends on complete and high-quality metabolic reference databases.

Future opportunities and challenges

Microbiome-metabolome studies have an important role in microbial ecology research. They specifically have great potential to dissect the metabolic interactions of complex microbial communities, and to unify “top down” and “bottom up” microbiome research approaches by providing mechanistic information at a systems level. Moreover, from a translational perspective, microbiome-metabolome studies can inform efforts to design targeted therapies to alter specific microbial or metabolic features of a community (13). Such interventions require first identifying putative targets, which in many cases may entail identifying the key contributor species that drive observed shifts in a particular beneficial or detrimental metabolic phenotype.

708 Importantly, while we show here that a correlation-based analysis may be limited in its
709 ability to identify these key microbe-metabolite links, this does not necessarily imply an
710 inherent limitation of microbiome-metabolome data. For example, analyzing our data,
711 we found that species abundance is in fact a very good proxy for metabolic activity
712 (median correlation of 0.996 between abundance and flux for all species-metabolite
713 pairs). When we further examined whether false negative associations in our dataset
714 stem from a disconnect between the abundance of a species and its metabolite uptake
715 or secretion rates, we identified only 2 undetected key contributor pairs that could be
716 explained by such a discrepancy. This analysis suggests that taxonomic abundance
717 data is sufficient to explain and model community metabolic variation to great extent,
718 despite common concerns about potential discrepancies between community
719 composition and function. It also suggests that metatranscriptomic expression data may
720 not provide much additional value for this purpose, as other studies have indicated (53,
721 59, 60).
722
723 Given the increasing prevalence of microbiome-metabolome studies, their promise, and
724 the caveats of association-based research discussed above, further development of
725 computational and statistical methods for analyzing such datasets is clearly needed.
726 Possible directions include the use of multi-species dynamic metabolic models that can
727 replicate experimental observations (61), multivariate approaches for deconvolving
728 interactions between species and the environment (62, 63), and probabilistic methods
729 that can integrate prior information while allowing for other unknown mechanisms (31,
730 64).

There is also a continued need for gold standards to evaluate new methods. This study is only a first step in that direction and has analyzed one specific type of research question: identifying microbial taxa directly responsible for variation in metabolite concentrations between samples in a cross-sectional study design. Although this focus describes many recent microbiome-metabolome studies, other studies may address a wide range of complementary research questions, and correspondingly, the desired "ground truth" can take different forms. Additionally, our findings rely on a single *in silico* system that does not capture many aspects of community metabolism. Further studies should also consider additional variables such as community diversity, sample size, measurement error, and other types of environmental variation. Ongoing technology developments in mass spectrometry and stable isotope probing will ideally enable future evaluation analyses using experimental, quantitative, species-specific community flux data to define key microbial contributors (65, 66). Such evaluations can also take advantage of datasets comparing community microbiome-metabolome data with *in vitro* monoculture or mono-colonization data (32–34).

Ultimately, much remains to be learned about the many processes through which complex microbial communities shape their environment. The first major call for the application of metabolomics to microbiome research, published 10 years ago (67), noted that new methods will be necessary to integrate genomic and metabolic data and inform the prediction of community metabolic properties from metagenomes. Now that microbiome-metabolome datasets are widely available, ongoing development of

754 analysis methods for these studies has great potential to generate new knowledge.
755 Moreover, future work in this area stands to benefit from the utility of dynamic,
756 multiscale metabolic modeling. Detailed mechanistic simulations are used widely in
757 astronomy, climate science, and other fields to make methodological choices and
758 assess possible experimental outcomes when ground truth measurements are
759 unavailable or difficult to obtain (68, 69). An analogous strategy in microbiome research
760 may be similarly fruitful.

761

Methods

Derivation of species contributors to variation

We derived an expression representing the contribution of each species to the variance in the concentration of each metabolite. While we describe this calculation in terms of species, a similar calculation could be done at the level of phyla, strains, or any grouping of the community for which metabolite secretion and uptake fluxes are available.

The concentration of a given metabolite *M* at the end of a single simulation run is a function of the uptake and secretion fluxes (responding to the species' degradation and synthesis activities) of the *n* species, the environmental inflow over all time steps m_{in}, and the dilution m_{out} out of the chemostat over all time steps:

$$M = \sum_{i=1}^{n} m_i + m_{in} - m_{out}$$

The value of m_{out} at a given time step *t* is the product of the dilution rate *D* and the metabolite concentration at the previous time point (see above). This fact can be used to express m_{out} in terms of all the previously recorded environmental inflow and microbial activities. The metabolite concentration at any time point *t*, *M(t)*, is then equal to:

$$M(t) = \sum_{k=1}^{t-1}\left[(1-D)^{t-k-1}\sum_{i=1}^{n} m_{ik}\right] + m_{in}\sum_{k=1}^{t-1}(1-D)^{k},$$

where m_{ik} represents the activity of species *i* at a single time point *k.* We can then ignore dilution outflow by replacing each activity value m_i in the final concentration calculation above with a value corrected for the mitigating effect of chemostat dilution over the course of the simulation up to time *t*, defined here as m_i*. m_i* represents the total amount of a compound secreted or uptaken by species *i*, minus the share of that quantity that is eventually diluted out over the course of the simulation.

$$m_i^* = \sum_{k=1}^{t-1}(1-D)^{t-k-1} m_{ik},$$

and thus,

$$M = m_{in} + \sum_{i=1}^{n} m_i^*$$

In this work, we refer to "environmental fluctuations" as the effect of the independently parameterized nutrient inflow, m_{in}, and where not otherwise specified we use m_i to imply m_i*, a species activity quantity that accounts for the corresponding subsequent dilution out of the system.

Using the expression above, *var(M)* can then be clearly expressed as a sum of correlated environmental and microbial random variables:

802 $$var(M) = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(m_i, m_j) + \sum_{i=1}^{n} cov(m_i, m_{env})$$

803 $$= \sum_{j=1}^{n} var(m_j) + var(m_{env}) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} cov(m_i, m_j) + 2\sum_{i=1}^{n} cov(m_i, m_{env})$$

804

805 This expression can then be partitioned additively into *n+1* terms representing the
806 contribution of each microbial species and of fluctuations in the environmental nutrient
807 inflow.

808

809 $$c_i = \sum_{j=1}^{n} cov(m_i, m_j) + cov(m_i, m_{env}) = var(m_i) + \sum_{j \neq i} cov(m_i, m_j) + cov(m_i, m_{env})$$

810

811 ***Multi-species Dynamic Flux Balance Analysis modeling***

812 In this study, we simulated the growth and metabolism of a community of 10
813 representative gut species that was previously explored experimentally (45). We
814 specifically utilized a previously introduced multi-scale framework for modeling the
815 dynamics and metabolism of multiple microbial species in a well-mixed shared nutrient
816 environment (42, 70). This framework assumes that each species in the community
817 aims to maximize its own growth on a short time scale given available nutrients, and
818 uses Flux Balance Analysis to predict the growth and metabolic activity of each species
819 at this short time scale (44). The shared environment is then iteratively updated based
820 on the species' predicted growth, uptake, and secretion rates, such that metabolic

821 interactions are mediated via the environment as a natural byproduct of species
822 activities, rather than being explicitly modeled (43).
823
824 We used genome-scale metabolic model reconstructions of the 10 community members
825 from the AGORA collection (38), which have been consistently curated to remove or
826 modify thermodynamically unfavorable reactions, remove futile cycles, and confirm
827 growth in anaerobic environments on expected carbon sources, with additional curation
828 for several biosynthesis pathways. The COBRA toolbox was used to convert each
829 AGORA model to MATLAB format (71). The growth and metabolism of the 10-species
830 community were simulated in a chemostat setting in 15-minute time intervals. We set
831 the chemostat volume to be approximately equal to a mouse gut (0.00134 liter (72)). We
832 similarly set metabolite inflows to emulate the macronutrient and micronutrient
833 quantities in a corn-based mouse chow (45) (provided in Supplementary Data 1).
834
835 The simulations were performed following a previously introduced procedure (42),
836 repeated for each time step t_n: First, the maximum uptake rate for all metabolites by all
837 species, denoted as v_{jk} for metabolite j and species k, were calculated based on
838 Michaelis-Menten single-substrate kinetics, with assumed universal values for maximum
839 rate V_{max} and transporter affinity K_m for all metabolites (provided in Supplementary Data
840 1). v_{jk} was further constrained based on an allocation of the metabolite's environmental
841 concentration to each species in proportion with its biomass. Then, the steady state
842 reaction fluxes for each species k at time point t_n were determined by maximizing the
843 growth rate μ_k, within the obtained constraints on environmental metabolite uptake. To

obtain a single and consistent flux solution for each species, the total flux activity for each species (i.e., the sum of absolute fluxes given the predicted optimal growth rate) was minimized, under the assumption that organisms prefer to operate their metabolism with minimal enzymatic cost (73). The optimal flux solutions were solved using linear programming with GLPK (www.gnu.org/software/glpk). With the resulting flux and growth rate information, the total biomass of each species *k*, $bio_k(t_n)$, was updated for the next time point t_{n+1}, using a standard exponential growth function incorporating dilution:

$$bio_k(t_{n+1}) = bio_k(t_n)e^{\mu_k \Delta t} - bio_k(t_n)D\Delta t,$$

where *D* is the dilution rate. We set *D* to 0.0472 per hour, in order to obtain community growth rates consistent with the observed average growth rate of the three most abundant species growing under 47 different carbon conditions (74). The total amount of uptake or secretion for each species *k* and metabolite *j* over a single time step was then calculated as previously derived (42):

$$m_{FBA}^{jk}(t_n) = \frac{v_{jk}}{\mu_k} * \ bio_k(t_n)(e^{\mu_k \Delta t} - 1),$$

where v_{jk} is the rate of uptake or secretion specified by the FBA solution for that species and metabolite at that time point, μ_k is the species growth rate, $bio_k(t_n)$ is the species abundance, and Δt is the size of the time step. Finally, combining the flux solutions of all species, nutrient inflow, and dilution, along with the steady state

assumption of no intracellular metabolite accumulation, the concentration of a given metabolite in the shared nutrient environment at the next time point, $M_j(t_{n+1})$ can be updated as:

$$M_j(t_{n+1}) = M_j(t_n) + m_{FBA}^j(t_n) + m_{in}^j \Delta t - M_j(t_n) D \Delta t,$$

where $m_{FBA}^j(t_n)$ is the metabolic impact from all species considering their abundance and their uptake and secretion rates of metabolite *j*, and m_{in}^j is the inflow rate of metabolite *j*. This process of calculating uptake rates, Flux Balance Analysis solutions, and updated metabolite concentrations was then repeated iteratively for the duration of the simulation.

Each simulation was run for a period of 144 hours or 576 time steps. This time period was long enough for most simulation runs to approach a steady state composition: specifically, in >65% of the simulations analyzed in our study, the change in abundance in any species over the final 3 hours was less than 0.01% of the carrying capacity (see below), and all had no changes greater than 0.3% of the capacity over that period. The concentrations of species and metabolites, the species growth rates, and the solved rates of all reactions for each species (including uptake and secretion) were recorded in each step of each simulation and used for subsequent analyses (Supplementary Data 1 and 2).

890 ***Simulation initialization parameters***

891 We fixed the initial total abundances of microbes to the carrying capacity for this system
892 and media, which was estimated to be 0.433 units of biomass. This capacity was
893 calculated as the average final total abundance from a set of simulations with varying
894 compositions and low initial abundances. We then varied the relative abundances,
895 increasing the abundance of one species at a time at the expense of all other species
896 equally. Specifically, for each species, we ran simulations in which the ratio of that
897 species' initial abundance relative to all other species was 2, 3, 4.5, 6, 9, and 13 times
898 (equating to a range in relative abundance of 10% to 60% for each species). This
899 resulted in a total of 61 simulation runs (one with all species starting at equal
900 abundance and 6 with increased abundance of each species). We chose this sample
901 size to approximately represent the sample sizes of published cross-sectional
902 microbiome-metabolome association studies (14, 16). We set the initial inflow
903 concentrations to the amount that would dilute in over one hour under the calculated
904 inflow rates.

905

906 ***Comparison with Shapley values***

907 We implemented an approximate Shapley value algorithm (41) as an alternative
908 strategy to calculate contributions for the simulated dataset. Briefly, 15,000 random
909 orderings of the 10 species were randomly generated. For each ordering, the variance
910 in metabolite activity is calculated for subsets of size 1 to 10, adding in species
911 according to the specified ordering. The difference in variance as a given species is
912 added to the subset, denoting the *marginal* contribution of that species to variation, is

913 recorded. The average marginal contribution across all orderings for each species is
914 then defined as its contribution to variance.

915

916 ***Species-metabolite correlation analysis***

917 We calculated Spearman correlations between absolute species abundances
918 (quantified as total biomass) and concentrations of variable metabolites. We used
919 absolute abundances in order to evaluate the relationships between species and
920 metabolites under the hypothetically best possible measurements of both data types.
921 We also compared correlation results using relative abundances and found very
922 minimal differences in the main simulation dataset: only 7 species-metabolite pairs
923 (1.3%) are significantly correlated using absolute abundances but not relative, and only
924 4 pairs (0.8%) are correlated using relative abundances but not absolute.

925

926 We used a *p*-value threshold of 0.01 to classify "significant" associations for binary
927 comparisons. For interpretability, we refer to *p*-values not corrected for multiple
928 hypothesis testing, since the number of tests remained constant across nearly all of our
929 analyses (520 possible species-metabolite pairs). The 0.01 threshold we use to define
930 significantly correlated pairs is equivalent to a Benjamini-Hochberg corrected q-value
931 threshold of 0.08, calculated using the R package *qvalue* (75).

932

933 ***Logistic regression modeling of correlation outcomes***

934 We used logistic regression models to identify factors that can be used to predict

935 whether a non-contributing species-metabolite pair displays a significant correlation
936 (false positive), and whether a key contributor species-metabolite pair fails to be
937 correlated (false negative). We used the *glm* function in R to fit models of the log odds
938 of whether a non-contributing species is correlated with its corresponding metabolite
939 (false positive or true negative), using as predictors grouped indicator values for species
940 and metabolite identities. We separately fit another set of logistic regression models to
941 predict whether a key contributor species is correlated (true positive or false negative),
942 with the same predictors. Models were compared using likelihood ratio tests using the
943 *anova* function in R.

944

945 ***Simulations with varied inflow quantities***

946 We ran 8 additional sets of simulations with the same set of 61 different initial species
947 compositions but with varying degrees of inflow fluctuations. Specifically, the nutrient
948 inflow quantities were sampled independently from a normal distribution, with a mean of
949 the original inflow concentration and the standard deviation equal to a set percent of the
950 mean. The 8 levels of deviation were 0.5%, 1%, 2%, 3%, 4%, 5%, 8%, or 10%. In the
951 comparison of correlation results across samples, we evaluated the same set of 52
952 variable metabolites as for the original dataset for consistency, although given the
953 added noise, additional metabolites met the same variance cutoff we used to define
954 variable metabolites.

955

956 To evaluate correlation performance as a function of increasing environmental
957 contribution, we binned the 38 analyzed inflow metabolites across the 8 datasets based

958 on the size of the environmental contribution to variance for the metabolite in that
959 dataset. In other words, metabolites in any dataset with an environmental contribution
960 greater than 0 but less than 10% of the total positive variance contributions were binned
961 into a single category, those with an environmental contribution between 10% and 20%
962 were binned into the next category, and so on. We analyzed the 52 metabolites in the
963 original constant-environment dataset as a separate category, and did the same for the
964 14 non-inflow metabolites in each of the 8 environmentally-varying datasets.
965
966 Confidence intervals for AUC values were calculated using the *pROC* package in R
967 (76), using a bootstrap method with 500 resamplings.
968

969 ***Application of MIMOSA to simulated data and comparison with correlation***
970 ***analysis***

971 We applied MIMOSA v1.0.2 (github.com/borenstein-lab/MIMOSA) (56) to the obtained
972 set of metabolite and species abundances. To construct the community metabolic
973 network model required by MIMOSA, we merged the 10 species-level models used in
974 the simulations into a single stoichiometric matrix. If a reversible reaction only ever
975 proceeded in a single direction in any simulation, we encoded it as non-reversible. To
976 apply the KEGG-based version of MIMOSA, we converted the model metabolite IDs to
977 KEGG IDs (55), downloaded KEGG Orthology gene annotations for the 10 modeled
978 species from the IMG/M database (77), and ran a MIMOSA analysis using the KEGG
979 metabolic network model encoded in *reaction_mapformula.lst* (KEGG version
980 downloaded 2-2018).

981

982 ***Code and data availability***

983 Code for all the analyses presented in this study is available online in the form of R
984 notebooks at https://github.com/borenstein-lab/microbiome-metabolome-evaluation. All
985 data generated and analyzed in this study and displayed in the figures are included in
986 Supplementary Data 1 through 3.

987

988 # Author contributions

989 C.N. and E.B. designed the study and wrote the paper. C.N. performed the analysis.
990 H.C.C. and C.P.M. contributed to the multi-species metabolic modeling simulations. All
991 authors read and approved the paper.

992

993 # Acknowledgements

994 C.N. was supported in part by a National Science Foundation (NSF) IGERT DGE-
995 1258485 fellowship. C.P.M. was funded by NHGRI grant T32 HG000035. This work was
996 supported in part by NIH New Innovator Award DP2 AT007802–01 and NIH grant
997 1R01GM124312–01 to E.B.

998

References

1. Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D'haeseleer P, Holman H-YN, Osman S, Lu Z, Van Nostrand JD, Deng Y, Zhou J, Mason OU. 2010. Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. Science 330:204–208.

2. Shi W, Moon C, Leahy S, Kang D, Froula J, Kittelmann S, Fan C, Deutsch S, Gagic D, Seedorf H, Kelly W, Atua R, Sang C, Soni P, Li D, Pinares-Patiño C, McEwan J, Janssen P, Chen F, Visel A, Wang Z, Attwood G, Rubin E. 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res gr.168245.113.

3. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L, Smith JD, DiDonato JA, Chen J, Li H, Wu GD, Lewis JD, Warrier M, Brown JM, Krauss RM, Tang WHW, Bushman FD, Lusis AJ, Hazen SL. 2013. Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 19:576–585.

4. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK. 2013. Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders. Cell 155:1451–1463.

1020 5. Dumas M-E, Barton RH, Toye A, Cloarec O, Blancher C, Rothwell A, Fearnside J,
1021 Tatoud R, Blanc V, Lindon JC, Mitchell SC, Holmes E, McCarthy MI, Scott J,
1022 Gauguier D, Nicholson JK. 2006. Metabolic profiling reveals a contribution of gut
1023 microbiota to fatty liver phenotype in insulin-resistant mice. Proc Natl Acad Sci
1024 103:12511–12516.

1025 6. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial metabolites and
1026 colorectal cancer. Nat Rev Microbiol 12:661–672.

1027 7. Wlodarska M, Luo C, Kolde R, d'Hennezel E, Annand JW, Heim CE, Krastel P,
1028 Schmitt EK, Omar AS, Creasey EA, Garner AL, Mohammadi S, O'Connell DJ,
1029 Abubucker S, Arthur TD, Franzosa EA, Huttenhower C, Murphy LO, Haiser HJ,
1030 Vlamakis H, Porter JA, Xavier RJ. 2017. Indoleacrylic Acid Produced by
1031 Commensal Peptostreptococcus Species Suppresses Inflammation. Cell Host
1032 Microbe 22:25-37.e6.

1033 8. Ferreyra JA, Wu KJ, Hryckowian AJ, Bouley DM, Weimer BC, Sonnenburg JL. 2014.
1034 Gut Microbiota-Produced Succinate Promotes C. difficile Infection after Antibiotic
1035 Treatment or Motility Disturbance. Cell Host Microbe 16:770–777.

1036 9. Rath S, Heidrich B, Pieper DH, Vital M. 2017. Uncovering the trimethylamine-
1037 producing bacteria of the human gut microbiota. Microbiome 5.

1038 10. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling
1039 AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ.

1040 2013. Diet rapidly and reproducibly alters the human gut microbiome. Nature
1041 505:559–563.

1042 11. De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Storia A, Laghi L, Serrazanetti
1043 DI, Di Cagno R, Ferrocino I, Lazzi C, Turroni S, Cocolin L, Brigidi P, Neviani E,
1044 Gobbetti M, O'Toole PW, Ercolini D. 2015. High-level adherence to a
1045 Mediterranean diet beneficially impacts the gut microbiota and associated
1046 metabolome. Gut gutjnl-2015-309957.

1047 12. Snijders AM, Langley SA, Kim Y-M, Brislawn CJ, Noecker C, Zink EM, Fansler SJ,
1048 Casey CP, Miller DR, Huang Y, Karpen GH, Celniker SE, Brown JB, Borenstein E,
1049 Jansson JK, Metz TO, Mao J-H. 2016. Influence of early life exposure, host
1050 genetics and diet on the mouse gut microbiome and metabolome. Nat Microbiol
1051 2:16221.

1052 13. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. 2017.
1053 Microbiome and metabolome data integration provides insight into health and
1054 disease. Transl Res.

1055 14. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, Raftery D,
1056 Marrazzo JM, Fredricks DN. 2015. Metabolic Signatures of Bacterial Vaginosis.
1057 mBio 6:e00204-15.

1058 15. Theriot CM, Koenigsknecht MJ, Carlson Jr PE, Hatton GE, Nelson AM, Li B,
1059 Huffnagle GB, Z. Li J, Young VB. 2014. Antibiotic-induced shifts in the mouse gut

1060 microbiome and metabolome increase susceptibility to Clostridium difficile
1061 infection. Nat Commun 5.

1062 16. Califf KJ, Schwarzberg-Lipson K, Garg N, Gibbons SM, Caporaso JG, Slots J,
1063 Cohen C, Dorrestein PC, Kelley ST. 2017. Multi-omics Analysis of Periodontal
1064 Pocket Microbial Communities Pre- and Posttreatment. mSystems 2:e00016-17.

1065 17. Garg N, Wang M, Hyde E, da Silva RR, Melnik AV, Protsyuk I, Bouslimani A, Lim
1066 YW, Wong R, Humphrey G, Ackermann G, Spivey T, Brouha SS, Bandeira N, Lin
1067 GY, Rohwer F, Conrad DJ, Alexandrov T, Knight R, Dorrestein PC. 2017. Three-
1068 Dimensional Microbiome and Metabolome Cartography of a Diseased Human
1069 Lung. Cell Host Microbe.

1070 18. Antharam VC, McEwen DC, Garrett TJ, Dossey AT, Li EC, Kozlov AN, Mesbah Z,
1071 Wang GP. 2016. An Integrated Metabolomic and Microbiome Analysis Identified
1072 Specific Gut Microbiota Associated with Fecal Cholesterol and Coprostanol in
1073 Clostridium difficile Infection. PLoS ONE 11:e0148824.

1074 19. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach
1075 L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. 2016. Integrated multi-omics
1076 of the human gut microbiome in a case study of familial type 1 diabetes. Nat
1077 Microbiol 2:16180.

1078 20. Hua C, Tian J, Tian P, Cong R, Luo Y, Geng Y, Tao S, Ni Y, Zhao R. 2017.
1079 Feeding a High Concentration Diet Induces Unhealthy Alterations in the

1080 Composition and Metabolism of Ruminal Microbiota and Host Response in a Goat
1081 Model. Front Microbiol 8.

1082 21. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, McDonald DT,
1083 Kusebauch U, Moss CL, Zhou Y, Qin S, Moritz RL, Brogaard K, Omenn GS,
1084 Lovejoy JC, Hood L. 2017. A wellness study of 108 individuals using personal,
1085 dense, dynamic data clouds. Nat Biotechnol.

1086 22. Vandeputte D, Falony G, Vieira-Silva S, Wang J, Sailer M, Theis S, Verbeke K,
1087 Raes J. 2017. Prebiotic inulin-type fructans induce specific changes in the human
1088 gut microbiota. Gut 66:1968–1974.

1089 23. Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ,
1090 Cotter PD. 2016. Microbial Succession and Flavor Production in the Fermented
1091 Dairy Beverage Kefir. mSystems 1:e00052-16.

1092 24. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. 2013. Stool
1093 Microbiome and Metabolome Differences between Colorectal Cancer Patients and
1094 Healthy Adults. PLoS ONE 8:e70803.

1095 25. Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein
1096 PC, Gilbert JA, Green JL, Jansson JK, Knight R, Maxon ME, McFall-Ngai MJ, Miller
1097 JF, Pollard KS, Ruby EG, Taha SA, Unified Microbiome Initiative Consortium.
1098 2015. A unified initiative to harness Earth's microbiomes. Science 350:507–508.

1099 26. iHMP Research Network Consortium. 2014. The Integrative Human Microbiome
1100 Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of
1101 Human Health and Disease. Cell Host Microbe 16:276–289.

1102 27. Choo JM, Kanno T, Zain NMM, Leong LEX, Abell GCJ, Keeble JE, Bruce KD,
1103 Mason AJ, Rogers GB. 2017. Divergent Relationships between Fecal Microbiota
1104 and Metabolome following Distinct Antibiotic-Induced Disruptions. mSphere
1105 2:e00005-17.

1106 28. Kang D-W, Ilhan ZE, Isern NG, Hoyt DW, Howsmon DP, Shaffer M, Lozupone CA,
1107 Hahn J, Adams JB, Krajmalnik-Brown R. 2018. Differences in fecal microbial
1108 metabolites and microbiota of children with autism spectrum disorders. Anaerobe
1109 49:121–131.

1110 29. Lin Z, Ye W, Zu X, Xie H, Li H, Li Y, Zhang W. 2018. Integrative metabolic and
1111 microbial profiling on patients with Spleen-yang-deficiency syndrome. Sci Rep 8.

1112 30. Melnik AV, da Silva RR, Hyde ER, Aksenov AA, Vargas F, Bouslimani A, Protsyuk
1113 I, Jarmusch AK, Tripathi A, Alexandrov T, Knight R, Dorrestein PC. 2017. Coupling
1114 Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide
1115 Association Studies of Human Fecal Samples. Anal Chem 89:7549–7559.

1116 31. Chong J, Xia J. 2017. Computational Approaches for Integrative Analysis of the
1117 Metabolome and Microbiome. Metabolites 7:62.

1118 32. Hoyles L, Jiménez-Pranteda ML, Chilloux J, Brial F, Myridakis A, Aranias T,
1119 Magnan C, Gibson GR, Sanderson JD, Nicholson JK, Gauguier D, McCartney AL,

1120 Dumas M-E. 2018. Metabolic retroconversion of trimethylamine N-oxide and the
1121 gut microbiota. Microbiome 6.

1122 33. Biggs MB, Medlock GL, Moutinho TJ, Lees HJ, Swann JR, Kolling GL, Papin JA.
1123 2016. Systems-level metabolism of the altered Schaedler flora, a complete gut
1124 microbiota. ISME J.

1125 34. Kešnerová L, Mars RAT, Ellegaard KM, Troilo M, Sauer U, Engel P. 2017.
1126 Disentangling metabolic functions of bacteria in the honey bee gut. PLOS Biol
1127 15:e2003467.

1128 35. Bauer E, Zimmermann J, Baldini F, Thiele I, Kaleta C. 2017. BacArena: Individual-
1129 based metabolic modeling of heterogeneous microbes in complex communities.
1130 PLOS Comput Biol 13:e1005544.

1131 36. Garza DR, van Verk MC, Huynen MA, Dutilh BE. 2018. Towards predicting the
1132 environmental metabolome from metagenomics with a mechanistic model. Nat
1133 Microbiol.

1134 37. Heinken A, Thiele I. 2015. Anoxic conditions promote species-specific mutualism
1135 between gut microbes in silico. Appl Environ Microbiol AEM.00101-15.

1136 38. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A,
1137 Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RMT, Thiele I. 2016.
1138 Generation of genome-scale metabolic reconstructions for 773 members of the
1139 human gut microbiota. Nat Biotechnol.

1140 39. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot
1141 E, de Wouters T, Juste C, Rizkalla S, Chilloux J, Hoyles L, Nicholson JK, Dore J,
1142 Dumas ME, Clement K, Bäckhed F, Nielsen J. 2015. Quantifying Diet-Induced
1143 Metabolic Changes of the Human Gut Microbiome. Cell Metab 22:320–331.

1144 40. Shapley LS. 1953. 17. A Value for n-Person Games, p. . *In* Kuhn, HW, Tucker, AW
1145 (eds.), Contributions to the Theory of Games (AM-28), Volume II. Princeton
1146 University Press, Princeton.

1147 41. Manor O, Borenstein E. 2017. Systematic Characterization and Analysis of the
1148 Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host
1149 Microbe 21:254–267.

1150 42. Chiu H-C, Levy R, Borenstein E. 2014. Emergent Biosynthetic Capacity in Simple
1151 Microbial Communities. PLoS Comput Biol 10:e1003695.

1152 43. Manor O, Levy R, Borenstein E. 2014. Mapping the Inner Workings of the
1153 Microbiome: Genomic- and Metagenomic-Based Study of Metabolism and
1154 Metabolic Interactions in the Human Microbiome. Cell Metab 20:742–752.

1155 44. Varma A, Palsson BO. 1994. Metabolic Flux Balancing: Basic Concepts, Scientific
1156 and Practical Use. Bio/Technology 12:994–998.

1157 45. Faith JJ, McNulty NP, Rey FE, Gordon JI. 2011. Predicting a Human Gut
1158 Microbiota's Response to Diet in Gnotobiotic Mice. Science 333:101–104.

1159 46. Rivière A, Gagnon M, Weckx S, Roy D, De Vuyst L. 2015. Mutual Cross-Feeding
1160 Interactions between Bifidobacterium longum subsp. longum NCC2705 and
1161 Eubacterium rectale ATCC 33656 Explain the Bifidogenic and Butyrogenic Effects
1162 of Arabinoxylan Oligosaccharides. Appl Environ Microbiol 81:7767–7781.

1163 47. Unterseher M, Jumpponen A, öPik M, Tedersoo L, Moora M, Dormann CF,
1164 Schnittler M. 2011. Species abundance distributions and richness estimations in
1165 fungal metagenomics - lessons learned from community ecology: COMMUNITY
1166 ECOLOGY IN FUNGAL METAGENOMICS. Mol Ecol 20:275–285.

1167 48. Mee MT, Collins JJ, Church GM, Wang HH. 2014. Syntrophic exchange in
1168 synthetic microbial communities. Proc Natl Acad Sci 111:E2149–E2156.

1169 49. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber
1170 TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ,
1171 Borneman J. 2013. Integrative analysis of the microbiome and metabolome of the
1172 human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome
1173 1:17.

1174 50. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu
1175 ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Sun F, Zhou
1176 J, Knight R. 2016. Correlation detection strategies in microbial data sets vary
1177 widely in sensitivity and precision. ISME J 10:1669–1681.

1178 51. Werhli AV, Grzegorczyk M, Husmeier D. 2006. Comparative evaluation of reverse
1179 engineering gene regulatory networks with relevance networks, graphical gaussian
1180 models and bayesian networks. Bioinformatics 22:2523–2531.

1181 52. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein
1182 PC, Knight R. 2016. Microbiome-wide association studies link dynamic microbial
1183 consortia to disease. Nature 535:94–103.

1184 53. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA,
1185 Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower
1186 C. 2013. Predictive functional profiling of microbial communities using 16S rRNA
1187 marker gene sequences. Nat Biotechnol 31:814–821.

1188 54. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA,
1189 Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong
1190 Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. 2014.
1191 The MetaCyc database of metabolic pathways and enzymes and the BioCyc
1192 collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459–D471.

1193 55. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes.
1194 Nucleic Acids Res 28:27–30.

1195 56. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, Fredricks
1196 DN, Borenstein E. 2016. Metabolic Model-Based Integration of Microbiome
1197 Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between
1198 Ecological and Metabolic Variation. mSystems 1:e00013-15.

1199 57. Casero D, Gill K, Sridharan V, Koturbash I, Nelson G, Hauer-Jensen M, Boerma M,
1200 Braun J, Cheema AK. 2017. Space-type radiation induces multimodal responses in
1201 the mouse gut microbiome and metabolome. Microbiome 5.

1202 58. Stewart CJ, Mansbach JM, Wong MC, Ajami NJ, Petrosino JF, Camargo CA,
1203 Hasegawa K. 2017. Associations of Nasopharyngeal Metabolome and Microbiome
1204 with Severity among Infants with Bronchiolitis. A Multiomic Analysis. Am J Respir
1205 Crit Care Med 196:882–891.

1206 59. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G,
1207 Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C.
1208 2014. Relating the metatranscriptome and metagenome of the human gut. Proc
1209 Natl Acad Sci 111:E2329–E2338.

1210 60. Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, DeSantis
1211 TZ. 2016. Piphillin: Improved Prediction of Metagenomic Content by Direct
1212 Inference from Human Microbiomes. PLOS ONE 11:e0166104.

1213 61. Magnúsdóttir S, Thiele I. 2018. Modeling metabolism of the human gut microbiome.
1214 Curr Opin Biotechnol 51:90–96.

1215 62. Doledec S, Chessel D. 1994. Co-inertia analysis: an alternative method for
1216 studying species-environment relationships. Freshw Biol 31:277–294.

1217 63. Randolph TW, Zhao S, Copeland W, Hullar M, Shojaie A. 2015. Kernel-Penalized
1218 Regression for Analysis of Microbiome Data. ArXiv151100297 Stat.

1219 64. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE,
1220 Schadt EE. 2012. Stitching together Multiple Data Dimensions Reveals Interacting
1221 Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. PLoS
1222 Biol 10:e1001301.

1223 65. Berry D, Stecher B, Schintlmeister A, Reichert J, Brugiroux S, Wild B, Wanek W,
1224 Richter A, Rauch I, Decker T, Loy A, Wagner M. 2013. Host-compound foraging by
1225 intestinal microbiota revealed by single-cell stable isotope probing. Proc Natl Acad
1226 Sci 110:4720–4725.

1227 66. Kurczy ME, Forsberg EM, Thorgersen MP, Poole FL, Benton HP, Ivanisevic J, Tran
1228 ML, Wall JD, Elias DA, Adams MWW, Siuzdak G. 2016. Global Isotope
1229 Metabolomics Reveals Adaptive Strategies for Nitrogen Assimilation. ACS Chem
1230 Biol 11:1677–1685.

1231 67. Turnbaugh PJ, Gordon JI. 2008. An Invitation to the Marriage of Metagenomics and
1232 Metabolomics. Cell 134:708–713.

1233 68. Collins WD, Bitz CM, Blackmon ML, Bonan GB, Bretherton CS, Carton JA, Chang
1234 P, Doney SC, Hack JJ, Henderson TB, Kiehl JT, Large WG, McKenna DS, Santer
1235 BD, Smith RD. 2006. The Community Climate System Model Version 3 (CCSM3). J
1236 Clim 19:2122–2143.

1237 69. Connolly AJ, Angeli GZ, Chandrasekharan S, Claver CF, Cook K, Ivezic Z, Jones
1238 RL, Krughoff KS, Peng E-H, Peterson J, Petry C, Rasmussen AP, Ridgway ST,
1239 Saha A, Sembroski G, vanderPlas J, Yoachim P. 2014. An end-to-end simulation

1240 framework for the Large Synoptic Survey Telescope, p. 915014. *In* Angeli, GZ,
1241 Dierickx, P (eds.), .

1242 70. McNally CP, Borenstein E. 2018. Metabolic model-based analysis of the
1243 emergence of bacterial cross-feeding via extensive gene loss. BMC Syst Biol 12.

1244 71. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC,
1245 Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. 2011.
1246 Quantitative prediction of cellular metabolism with constraint-based models: the
1247 COBRA Toolbox v2.0. Nat Protoc 6:1290–1307.

1248 72. Casteleyn C, Rekecki A, Van der Aa A, Simoens P, Van den Broeck W. 2010.
1249 Surface area assessment of the murine intestinal tract as a prerequisite for oral
1250 dose translation from mouse to man. Lab Anim 44:176–183.

1251 73. Holzhütter H-G. 2004. The principle of flux minimization and its application to
1252 estimate stationary fluxes in metabolic networks: Flux minimization. Eur J Biochem
1253 271:2905–2922.

1254 74. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, Pudlo N a,
1255 Muegge BD, Henrissat B, Hettich RL, Gordon JI. 2013. Effects of diet on resource
1256 utilization by a model human gut microbiota containing Bacteroides cellulosilyticus
1257 WH2, a symbiont with an extensive glycobiome. PLoS Biol 11:e1001637.

1258 75. Dabney A, Storey JD. 2015. qvalue: Q-value estimation for false discovery rate
1259 control.

1260 76. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011.
1261 pROC: an open-source package for R and S+ to analyze and compare ROC
1262 curves. BMC Bioinformatics 12:77.

1263 77. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A,
1264 Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova
1265 NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and
1266 comparative analysis system. Nucleic Acids Res 40:D115–D122.
1267

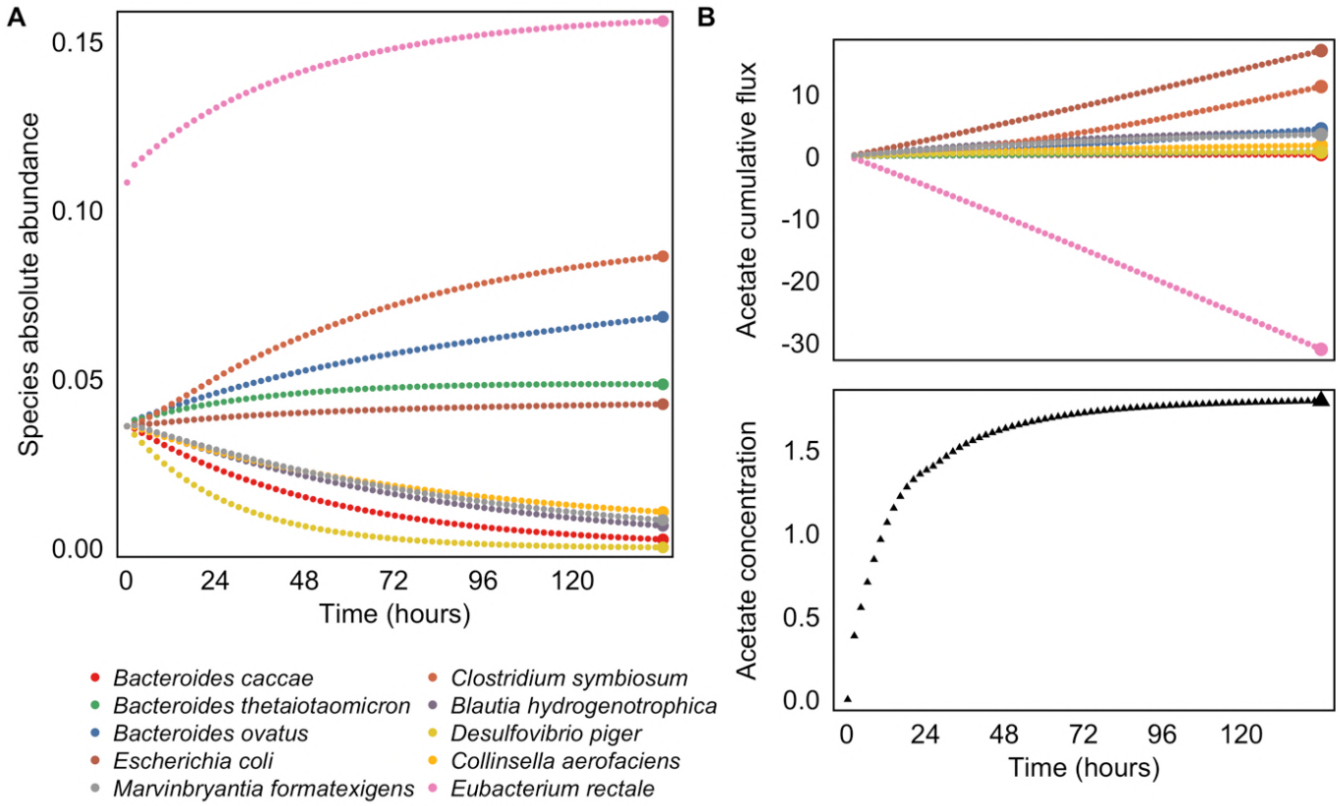

Figure 1. Simulating multi-omic data with a dynamic multi-species genome-scale framework. (A) Community species abundances throughout a single simulation run. Abundances were quantified in units of microbial biomass. In this simulation, community composition was initialized with a high relative abundance of *Eubacterium rectale*. For visual clarity, only every eighth time step is illustrated. Species abundances at the final time point (highlighted with larger colored circles) were used for calculating species-metabolite correlations. **(B)** Cumulative secretion and uptake of acetate by each community member, throughout the same simulation run illustrated in panel A. Acetate was synthesized by several species and consumed by *E. rectale* over the course of the simulation. Total cumulative fluxes (highlighted with larger colored circles) were used for calculating species contributions to metabolite variation. The bottom plot illustrates the resulting environmental concentration of acetate at each time point. The metabolite concentration at the final time point (highlighted with a larger black triangle) was used for calculating species-metabolite correlations.

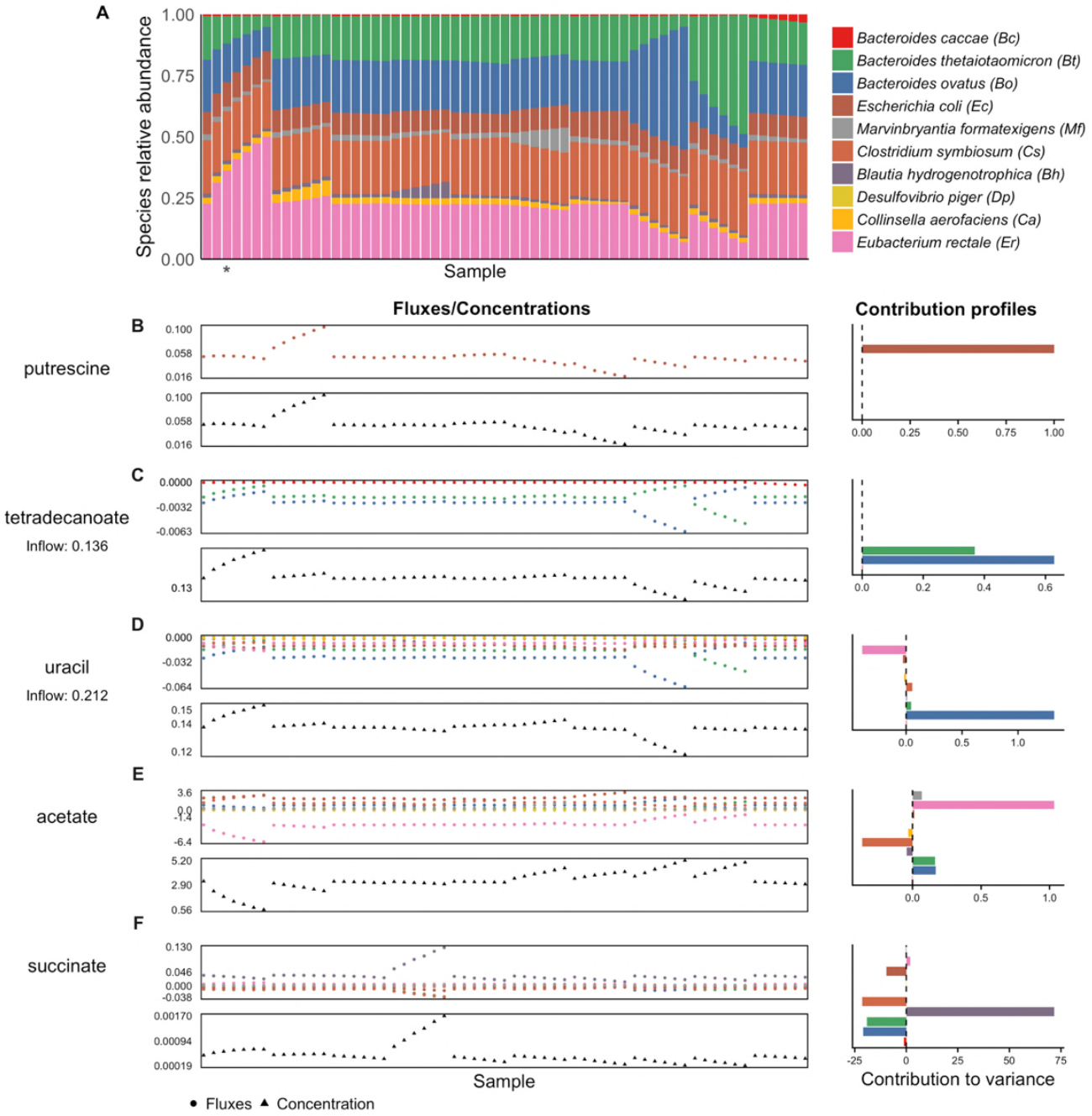

Figure 2. Species abundances, cumulative fluxes, and contributions to variance in metabolite concentrations in our simulated dataset. (A) The dataset of species abundances at the final time point of 61 simulation runs. Each bar represents a simulation run, with the colors indicating relative abundance of each species. The abundance profile from the simulation runs highlighted in Figure 1 is indicated with an asterisk. **(B-F)** For five example metabolites, the upper plot shows the total cumulative secretion or uptake of that metabolite by each species across all 61 simulation runs (or samples). The lower plot shows the corresponding environmental concentration at the final time point. The bar plot on the right shows the contribution values for each species and metabolite, calculated from the flux values and describing each species' linear contribution to the overall metabolite variance.

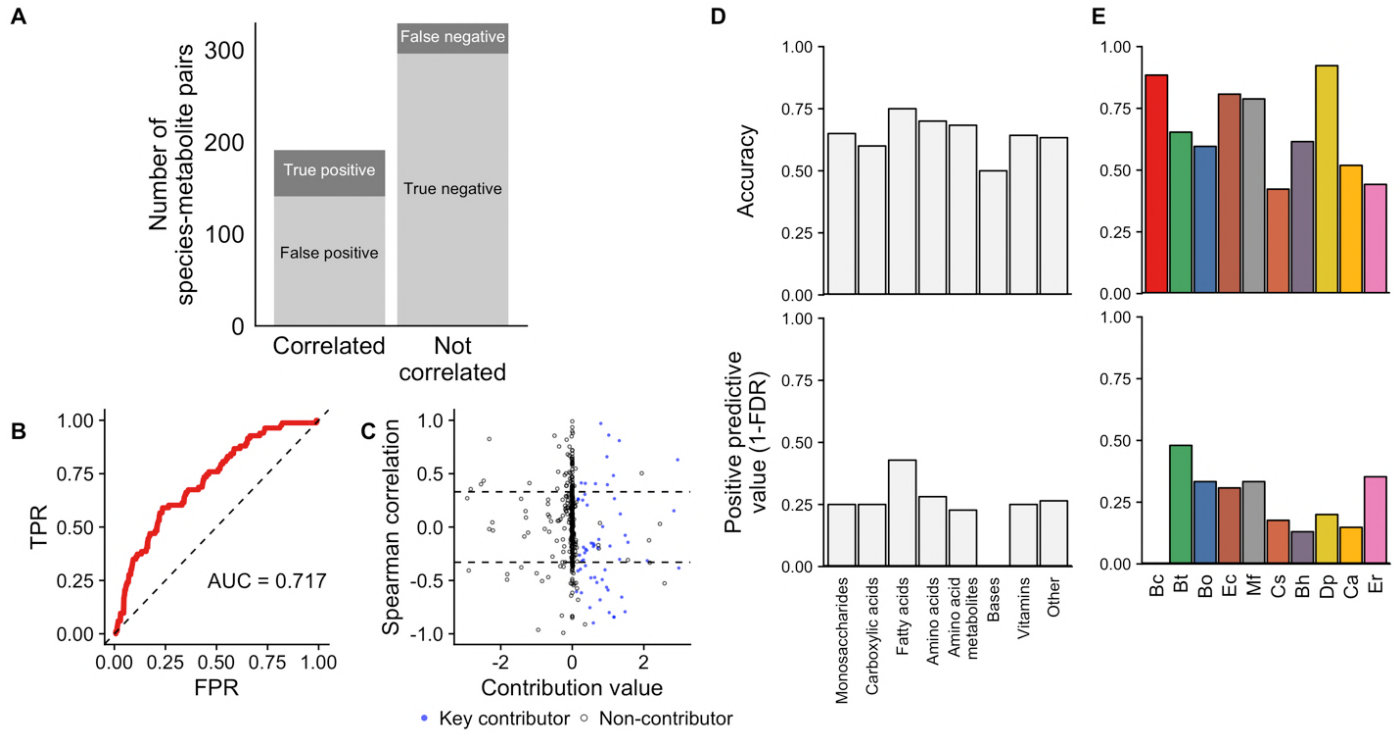

Figure 3. **Species-metabolite correlations poorly predict species contributions to metabolite variation. (A)** The number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (indicated by shade of gray). **(B)** Receiver operating characteristic (ROC) plot, showing the ability of absolute Spearman correlation values to classify key contributors among all species-metabolite pairs. **(C)** Scatter plot of species-metabolite pairs, showing the poor correspondence between true contribution values (x-axis) and Spearman correlation (y-axis). Key contributors are plotted as blue points, others as hollow circles. Dashed lines show significant correlations (p<0.01). There are 65 species-metabolite pairs with a contribution value greater than 3 in magnitude whose values are not shown. **(D-E)** Accuracy and positive predictive value of Spearman correlation analysis for detecting true key contributors across metabolite classes (Panel D) and for each of the 10 species (Panel E).

8

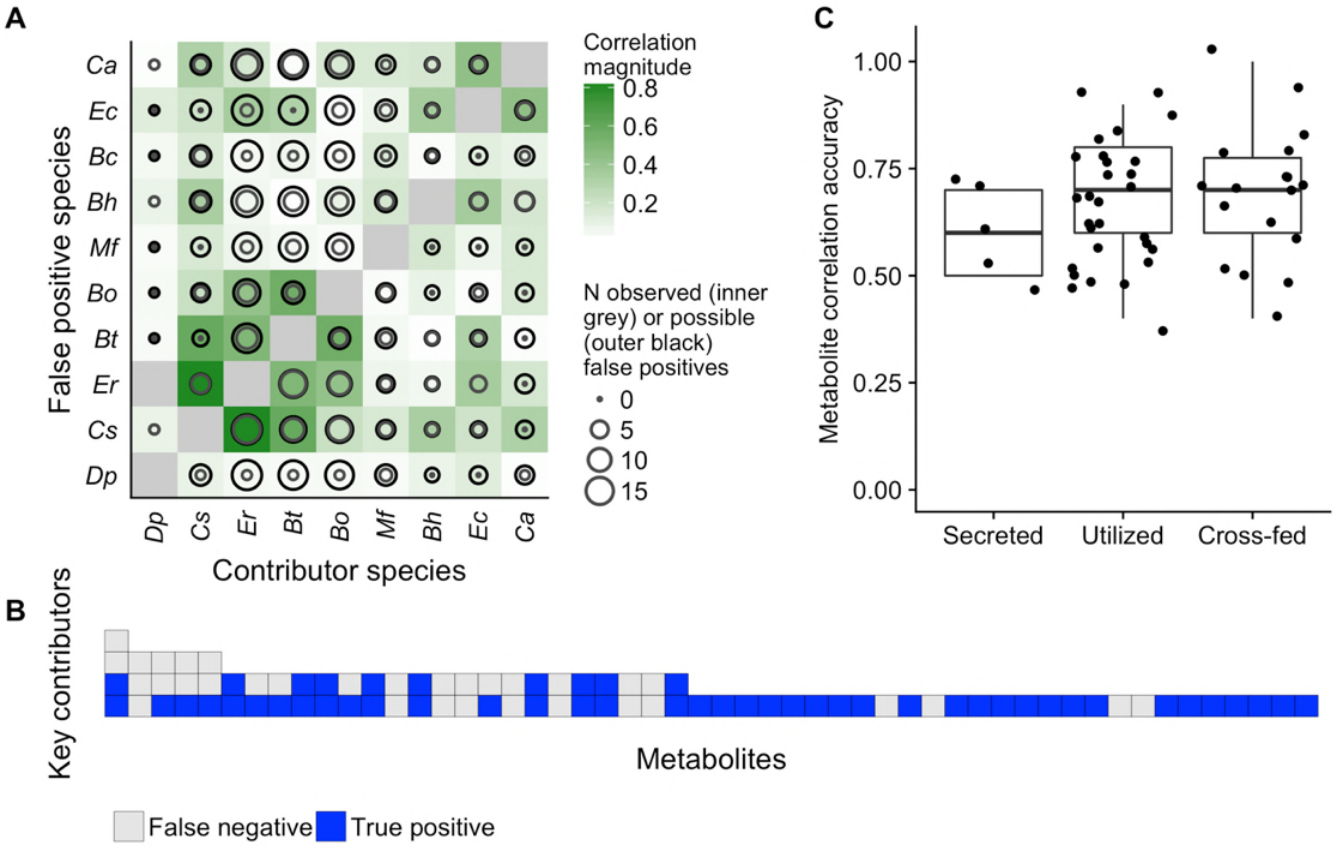

Figure 4. Metabolite and species properties explain correlation-contribution discrepancies. (A) Strongly correlated species pairs produced more false positive metabolite correlations. In this plot, the color of each tile indicates the strength of correlation in the abundances of each pair of species. The size of the outer black circle in each cell represents the number of metabolites for which the species on the x-axis is a key contributor and the species on the y-axis is not. The size of the inner circle represents the share of those metabolites for which a false positive is observed for the species on the y-axis. It can be seen that many false positive correlations involve the taxa with the strongest interspecies associations: *E. rectale, B. ovatus,* and *B. thetaiotaomicron.* **(B)** Metabolites with more microbial key contributors were more prone to false negative correlations. Each column represents an analyzed metabolite, ordered by its number of key microbial contributors, which are represented by each tile. The tiles are coded by the correlation outcome for each contributor. **(C)** Correlations detected key contributors equally accurately regardless of whether a metabolite is secreted, utilized, or cross-fed by the species. Each point represents the accuracy of correlations for a single metabolite across its comparisons with all 10 species.

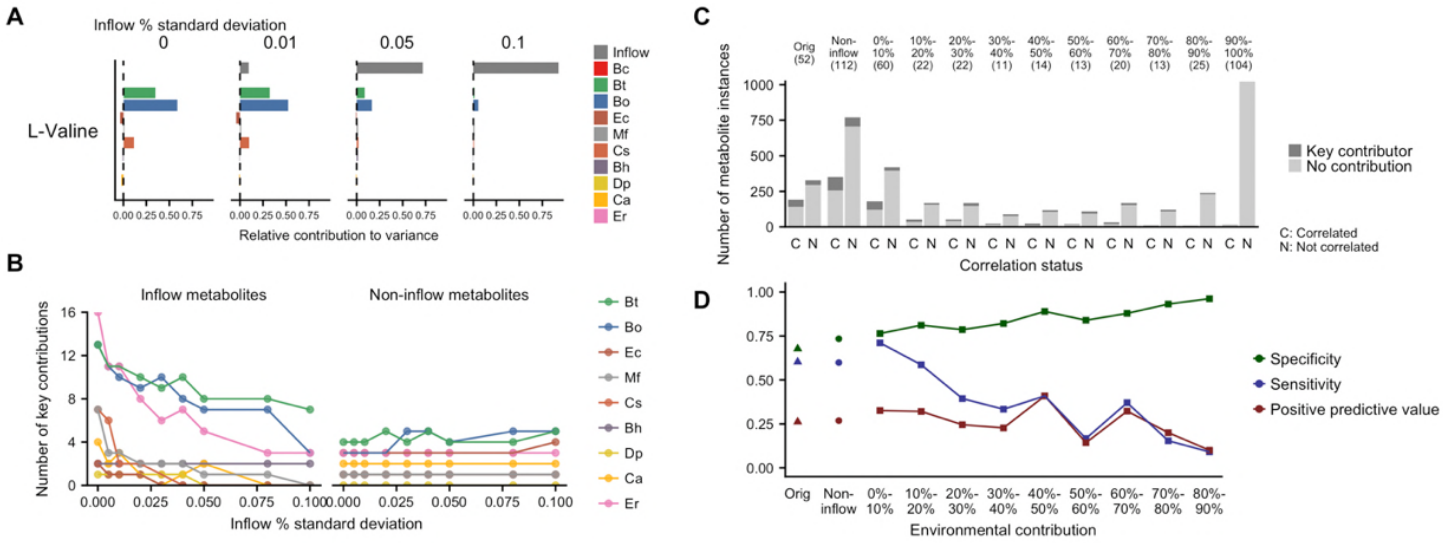

Figure 5. Environmental fluctuations impact correlation-contributor sensitivity and specificity. (A) Example set of contribution profiles for a single inflow metabolite, L-valine, with increasing fluctuations in its inflow. The relative contribution values for each species and for the inflow are shown for 4 sets simulation runs, each with a different degree of fluctuation. The label on each plot describes the relative standard deviation (coefficient of variation) of inflow metabolite concentrations for that set of simulations. The microbial contributions to variance in L-valine concentrations became relatively smaller with increasing variation from the external environment. **(B)** Shifts in key microbial contributors with increasing environmental inflow fluctuations. The number of key contributions of each species to the 52 analyzed metabolites is shown, separately for metabolites present in and absent from the nutrient inflow. Microbial contributors to inflow metabolites decreased as environmental contributions increased, but this effect varied between taxa. **(C)** Correlation analysis failed to detect key microbial contributors regardless of the size of contribution from external inflow variation. Across all sets of simulations, metabolites were binned based on the percent of total positive contribution from the external inflow. The bar plots shown have the same format as Figure 3A, showing the number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (indicated by shade of gray). The first two bars, labeled "Orig" describe the original set of simulations (replicating Figure 3A). The next two show the results for non-inflow metabolites across all levels of inflow fluctuations. The remaining bars show the results for metabolites with increasing levels of environmental contribution. **(D)** Correlation analysis detected key microbial contributors with increased specificity, decreased sensitivity, and generally consistent positive predictive value with increasing contribution from the external inflow. Sensitivity, specificity, and positive predictive value are shown for same environmental contribution bins as in Panel C.

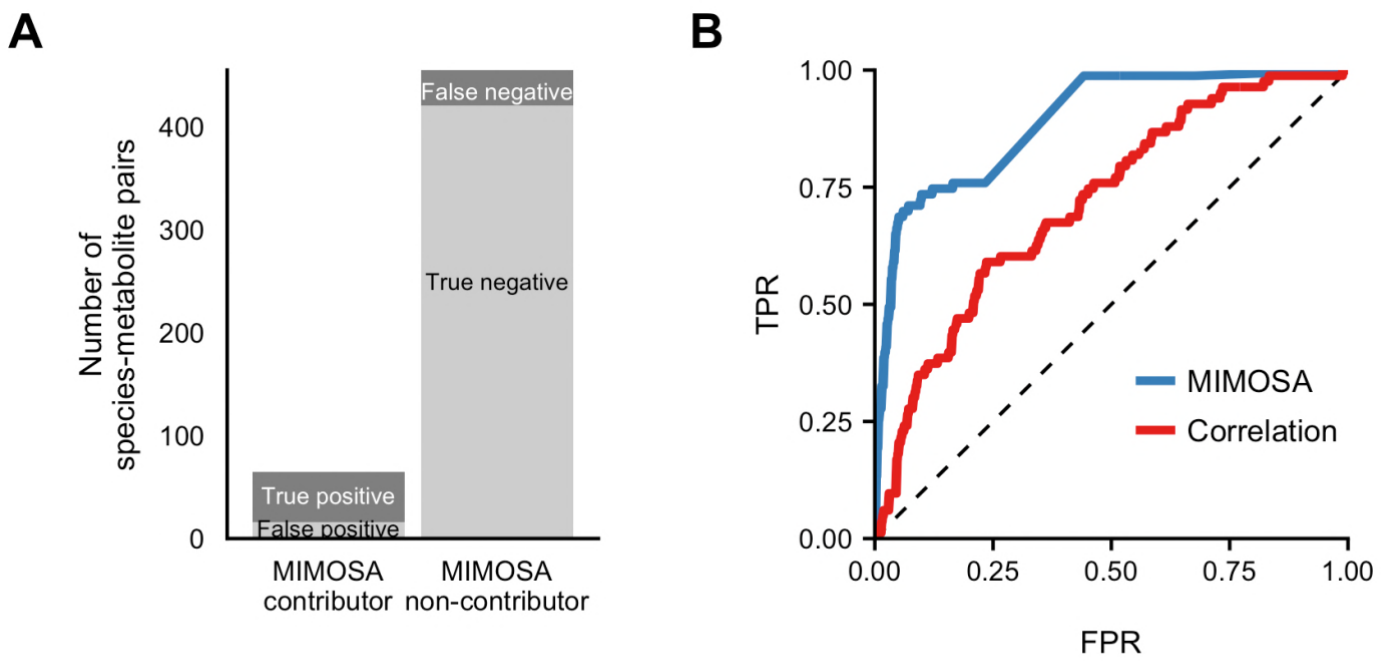

Figure 6. MIMOSA identified key microbial contributors more accurately than correlation analysis. (A) The number of species-metabolite pairs that were identified as potential contributors (left bar) or not (right bar) by MIMOSA, and its correspondence with true key contributors. **(B)** Receiver operating characteristic (ROC) plot, showing the ability of both MIMOSA and absolute Spearman correlation values to classify key contributors among all species-metabolite pairs.