Main Manuscript for

Bioenergetic mapping of 'healthy microbiomes' via compound processing potential imprinted in gut and soil metagenomes

Craig Liddicoat[a,1], Robert A. Edwards[a], Michael Roach[a], Jake M. Robinson[a], Kiri Joy Wallace[b], Andrew D. Barnes[b], Joel Brame[a], Anna Heintz-Buschart[c], Timothy R. Cavagnaro[a], Elizabeth A. Dinsdale[a], Michael P. Doane[a], Nico Eisenhauer[d,e], Grace Mitchell[b,f], Bibishan Rai[b], Sunita Ramesh[a], Martin F. Breed[a]

[a]College of Science and Engineering, Flinders University, Bedford Park, South Australia, Australia
[b]Environmental Research Institute, University of Waikato, Hamilton, Aotearoa New Zealand
[c]Swammerdam Institute for Life Sciences, University of Amsterdam, 1098 XH Amsterdam, The Netherlands
[d]German Centre for Integrative Biodiversity Research (iDiv), 04103 Leipzig, Germany
[e]Institute of Biology, Leipzig University, 04103 Leipzig, Germany
[f]Manaaki Whenua – Landcare Research, Hamilton, Aotearoa New Zealand

[1]Corresponding author: Craig Liddicoat; craig.liddicoat@flinders.edu.au

ORCIDs:

CL: 0000-0002-4812-7524, RAE: 0000-0001-8383-8949, MR: 0000-0003-1488-5148, JMR: 0000-0001-8108-3271, KJW: 0000-0003-1236-6846, ADB: 0000-0002-6499-381X, JB: 0000-0002-1852-9987, AHB: 0000-0002-9780-1933, TRC: 0000-0002-9922-5677, EAD: 0000-0002-2177-203X, MPD: 0000-0001-9820-2193, NE: 0000-0002-0371-6720, GM: 0000-0002-2839-7981, SR: 0000-0003-2230-4737, MFB: 0000-0001-7810-9696

Author Contributions: CL, RAE, MFB designed research; CL, MR, JMR, KJW, ADB, AHB, MPD, NE, GM, BR performed research; CL, RAE, JMR, KJW, ADB, JB, AHB, TRC, EAD, MPD, NE, SR, MFB analyzed data; CL drafted the manuscript; all authors reviewed and edited the final version.

Preprint Server: bioRxiv.org

Competing Interest Statement: The authors declare no competing interest.

Classification: BIOLOGICAL SCIENCES; Applied Biological Sciences

Keywords: Metagenomics, compound processing potential, healthy microbiome, human health, bioenergetic mapping

This PDF file includes:

Main Text
Figures 1 to 7
Tables 1 (only)

Abstract

Microbiomes are critical to the health and functioning of humans and ecosystems. Defining 'healthy microbiomes', however, remains elusive. More advanced knowledge exists on health associations for the compounds used or produced by microbes. Because microbes, their feedstocks and micro-environments interact synchronously, using functional genes to facilitate chemical transformations, this presents an intriguing opportunity to examine microbiomes through their potential to process compounds associated with human health. There is also growing interest in environmental microbiota that might be efficient at processing health-associated compounds because these microbes may readily transfer to humans and environmental interventions could modulate our exposure to them. Here we propose a bioenergetic mapping approach to microbiome assessments that examines the compound processing potential imprinted in human gut and environmental soil metagenomes. From shotgun metagenomics functional profiling, we derive quantitative measures of compound processing potential for human health-associated compound classes (e.g., lipids, carbohydrates) and selected biomolecules of interest (e.g., vitamins, short-chain fatty acids). We mapped microbial functions to compounds using the complexity-reducing van Krevelen bioenergetic mapping framework, based on carbon-hydrogen-oxygen stoichiometry and principal axes that explain variation in microbial distribution and chemical speciation. We found differences in compound processing potential within gut metagenomes comparing health- and disease-associated samples, including atherosclerotic cardiovascular disease, colorectal cancer, type 2 diabetes and anxious-depressive behaviors. Patterns of compound processing potential in soil metagenomes were linked with ecosystem maturity. Assessment of compound processing potential offers a new lens to explore mechanisms of microbiome-mediated human health including connections to health-promoting environmental microbiomes.

Significance Statement

Despite mounting evidence of their importance, the definition and measurement of 'healthy microbiomes' remain unclear. Knowledge gaps hinder development of microbiota-oriented approaches in human health, including potential for environmental interventions. By integrating interdisciplinary knowledge frameworks

including functional genomics and biochemistry, we derive summary measures of potential for human gut and environmental soil metagenomes to process major compound classes and biomolecules linked to human health. Measures of compound processing potential were linked with states of human health and disease; and displayed seemingly predictable shifts along gradients of ecological disturbance in plant-soil systems. Compound processing potential offers a simplifying approach for applying powerful and otherwise complex metagenomics in ongoing efforts to understand and quantify the role of microbiota in human- and environmental-health.

Main Text

Introduction

Microbial communities (microbiota), their feedstocks (substrates, nutrients) and environmental conditions (e.g., pH, redox potential, temperature, moisture, salinity) work in concert to drive microbially-mediated reactions essential to fueling life on Earth (1). Microbiomes (i.e., microbiota, genetic material and metabolites) are intimately linked to human health and disease (2-4), as well as the functioning of ecosystems (5, 6). Microbial functional capacity supports the transformation and exchange of chemicals, molecules, and energy, benefiting microbiota members, host organisms, and wider ecological networks (1, 5, 7, 8). Many microbes are often highly specialized and efficient at performing a particular suite of reactions. Accordingly, microbiota are shaped by the resources they utilize and the environments they inhabit (3, 9).

Microbes typically operate as a community (10) where many taxa lack the functional capacity for stand-alone survival (11). Complex cross-feeding and resource sharing in the extracellular space (7) suggest that community-scale functional profiles (rather than specific microbial taxa) underpin the health-supporting capacity of microbiota. However, community-scale complexity has hindered progress towards clear definitions of a 'healthy microbiome' (12). Nevertheless, researchers want to better understand the assembly and structure of health-promoting microbiomes to improve the course of microbiome-associated diseases. Disease-associated microbiota are often characterized by a loss of diversity and dominance by opportunistic pathogens (4, 13), but it may be unclear whether they represent facilitators or followers of disease. In contrast to direct microbiota-health links, our knowledge of health associations for various biomolecules and other chemical compounds (linked to microbiomes) is comparatively well advanced. Because microbiota-mediated reactions fundamentally involve transformations between different chemical compounds, this creates the intriguing possibility of examining microbiomes through their potential to process (i.e., convert or produce) compounds associated with human health.

Additionally, the involvement of environmental microbiomes in processing human health-associated compounds is of interest. Transfer of environmental microbiota to humans may help supplement important functional capacity, protective microbiota, and immune-signaling agents, particularly in infants, but also in adults who have depleted microbiota due to antibiotic use, poor diet, lifestyle or other health incidents (14, 15). If the functional composition of microbiota varies predictably along environmental gradients, then through design, management, and behavior we should be able to modulate our exposure to health-promoting versus disease-associated microbes. Soils, in particular, can represent a rich source of microbial diversity with potential to support human health (16). Microbiota in plant-soil systems are shaped by macro-scale factors including climate, soil characteristics, vegetation composition, diversity, land use and management (9, 17). With the prospect of cost-effectively encouraging health-promoting microbes, it is frequently asked, "What type of environment is best?" Yet, the attributes of health-promoting environmental microbiomes, including potential functional overlaps with human microbiomes, remain understudied.

Many microbiome-associated diseases are linked to bioenergetic mechanisms (7), with oxidation-reduction (redox) potential recognized as a key factor shaping microbial communities. The healthy anaerobic gut favours obligate anaerobes, whereas dysbiosis is often accompanied by increased oxygenation of the colonic epithelium and expansion of oxygen-tolerant facultative anaerobic bacteria (18, 19). Oxygen is a highly electronegative element important in shaping electrochemical gradients, biochemical reactions, and gene expression (20). Oxygen content varies in different types of organic matter (i.e., microbial feedstocks). Yet, the interrelationship between bioenergetic drivers, compounds, microbial environments and microbiota development receives little attention. In soils, redox potential varies with weather, vegetation, land use, management, drainage, organic-content, vicinity to roots, soil characteristics, and microbial activity (21, 22). At the molecular level, redox potential shapes what kind of molecules can be made and how energy is stored. Therefore, a compound-oriented examination of healthy microbiomes might capitalize on available knowledge linking compounds with human health, while also considering deterministic influences of bioenergetic (or electrochemical) energy gradients.

88 For example, in gut microbiome bioenergetics and chronic metabolic diseases, Daisley*, et al.* (7;
89 their Figs. 3-4) highlight key human health-associated biomolecules found within extracellular resources
90 shared by microbes. Short-chain fatty acids (SCFAs) acetate, propionate, and butyrate benefit host
91 metabolism, intestinal barrier function, systemic anti-inflammatory effects, and contribute up to 10% of
92 daily energy requirements (23). B group vitamins: riboflavin (B2), cobalamin (B12), pyridoxal 5'-phosphate
93 (B6), and folate (B9) are critical in electron transport and represent precursors to a variety of enzyme
94 cofactors essential to the tricarboxylic acid (TCA) cycle, fatty acid oxidation, and other metabolic
95 pathways (7). Menaquinone (Vitamin K2) is a critical electron carrier in bacteria and considered essential
96 in humans for calcium regulation (7). Other keystone health-linked biomolecules include glutamate and
97 pyruvate. Glutamate is the major excitatory neurotransmitter of the healthy mammalian brain, and an
98 abundant free amino acid important in multiple metabolic pathways, which requires regulation at optimal
99 levels in extracellular fluids (24). Glutamate is sensed luminally in the intestinal mucosa, triggering vagus
100 nerve (gut-brain axis) activity (25). Pyruvate is a critical intermediate involved in human energy
101 metabolism, where dysregulation is associated with cancer, heart failure, and neurodegeneration (26).

102 Here, we examine the functional potential of gut and soil microbiota from a compound processing
103 potential (CPP) viewpoint, to assess patterns in human health and disease, and with gradients of
104 ecosystem maturity. Such an approach might discern health- versus disease-promoting microbiotas from
105 the types of biochemical compounds they are attuned to consuming or producing (reflecting microbial
106 feedstocks and metabolites respectively). Previous metabolome prediction frameworks (e.g., 27, 28-30)
107 rely variously on supplementary metabolome training datasets, microorganism-specific genome-scale
108 metabolic models, taxonomic abundance estimates, and modeled or assumed environmental conditions
109 (e.g., human gut). In this work, we wanted to exploit community-scale compound-oriented information that
110 might be embedded within metagenomes (regardless of taxa present). As functional potential profiling
111 from whole genome sequencing (or shotgun metagenomics) does not directly measure functions
112 performed, we characterized microbiota-linked CPP from DNA sequencing, without direct measurement
113 of compounds. We premised that the ease of transformation between health-associated compounds and

114 other compounds that closely resemble them will depend on stoichiometric and energetic similarities,
115 microbiota functional diversity, and environmental conditions.

116 We utilized a framework that integrates information about compounds, bioenergetics, and
117 environmental conditions. The complexity-reducing van Krevelen (vK) coordinate space offers a simplified
118 and intuitive bioenergetic framework for approximate mapping of compounds based on their carbon (C),
119 oxygen (O) and hydrogen (H) content, while also reflecting energy density and principal axes that explain
120 microbial distribution and chemical speciation (Fig. 1; SI Appendix, Fig. S1) (31, 32). Compounds are
121 mapped into vK space using their O:C and H:C molar ratios (x- and y-axis respectively). We surmised this
122 framework could offer an exhaustive and intuitive mapping space to summarize the nature of microbiota-
123 mediated functional reactions in a way that reflects mean or dominant compound properties, reaction
124 stoichiometries, and potential overlaps between dietary or environmental substrates, and key health-
125 associated biomolecules.

126 Specifically, we combined SUPER-FOCUS functional profiling (33), the comprehensive
127 ModelSEED (34) functional-biochemistry database system and vK coordinate mapping to assign
128 functional potential relative abundances from human gut and soil sample metagenomes to overall mean
129 reaction-level vK coordinates. This approach effectively mapped every SUPER-FOCUS function (where
130 feasible via available corresponding database information) to an abundance-weighted mean proxy
131 chemical compound (or reaction-level 'meta-compound') represented in the two-dimensional vK space (SI
132 Appendix, Fig. S2). Limitations to this simplified representation of functional profiles are discussed below.
133 We aimed to: 1) investigate measures of microbiota CPP imprinted in human gut and soil metagenomes;
134 and 2) test for differences in human health and disease, and in disturbed, restored and natural
135 ecosystems. We hypothesized this bioenergetic mapping approach might identify CPP profiles, and
136 overlaps in human and environmental datasets, that could inform the definition and future shaping of
137 'healthy microbiomes'. We were also keen to explore whether CPP measures might enhance the
138 interpretability and accessibility of metagenomics data to aid hypothesis building and prioritizing future
139 research.

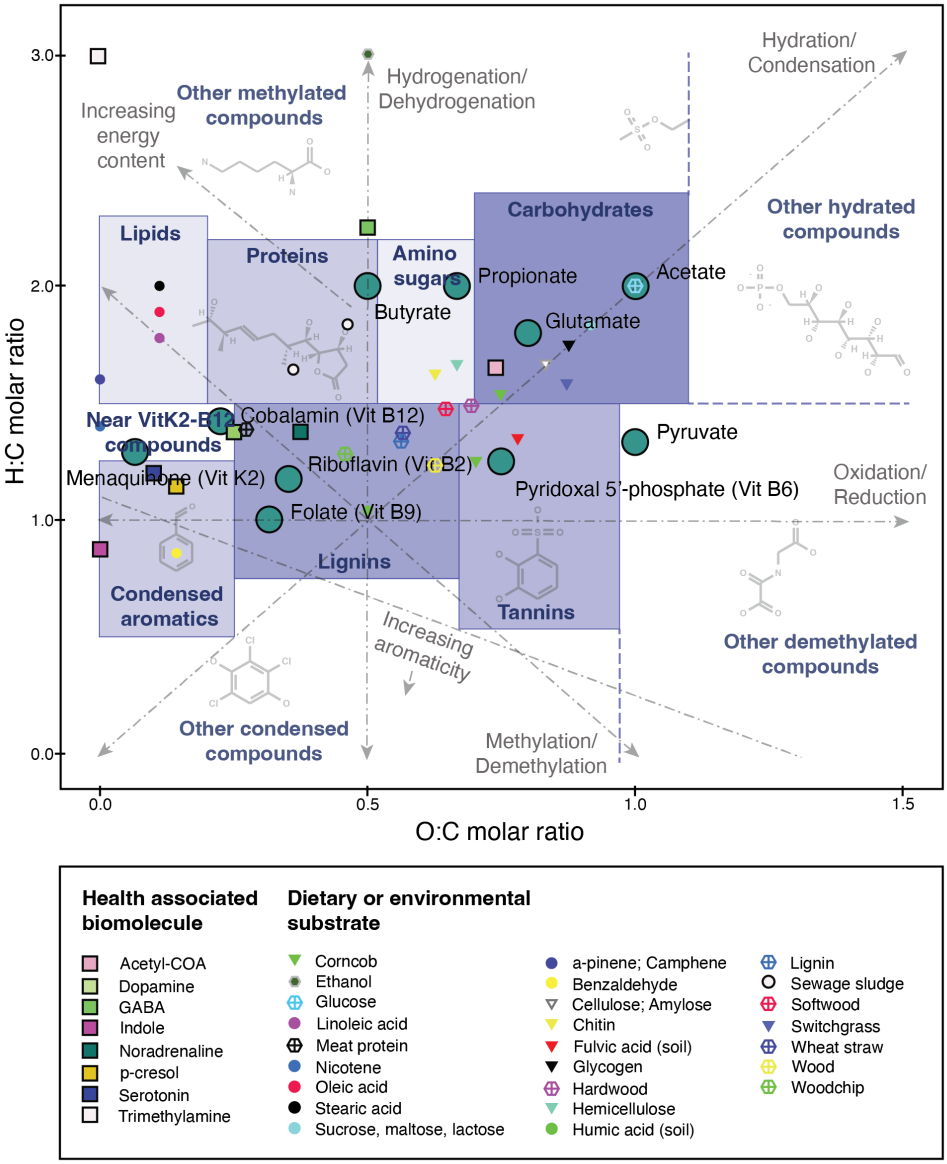


Fig. 1. Van Krevelen (vK) coordinate space (adapted from 31), displaying major compound classes (purple zones and text), key gradients (grey axes and text), focus biomolecules examined in this study (large dots), and additional example health-associated biomolecules, dietary or environmental substrates (legend). vK zones were adapted from (32) (see SI Appendix, Tables S1, S2). Key gradients include oxidation-reduction (x-axis), hydrogenation-dehydrogenation (y-axis), hydration-condensation (top-right to bottom-left), methylation-demethylation (top-left to bottom-right) and increasing energy content (towards top-left).

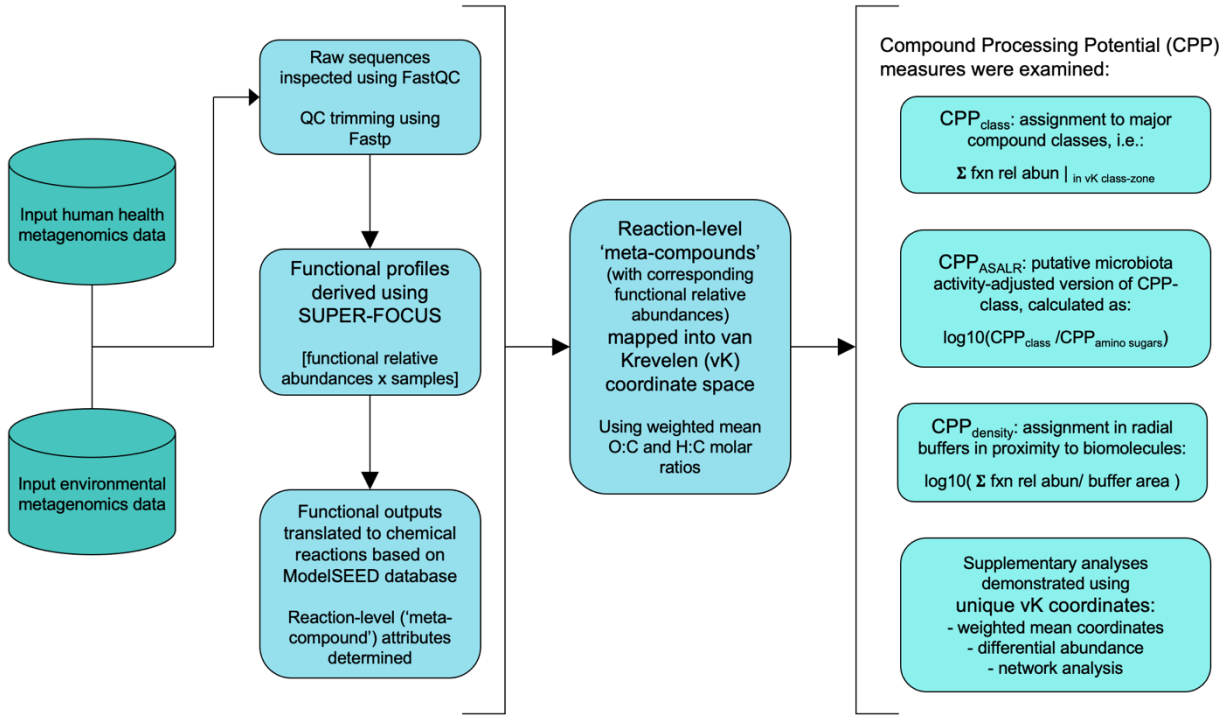
148 **Results**

149 We developed compound processing potential metrics to assess four human health and disease
150 datasets, comprising atherosclerotic cardiovascular disease (ACVD)(35), colorectal cancer (36), type 2
151 diabetes (T2D)(37) and problem (anxious-depressive) behaviors in children (38); and three environmental
152 soil datasets from ecological restoration and disturbed versus natural plant-soil systems (39-41) (Table 1).
153 Four CPP metrics were evaluated (Fig. 2; detailed in Methods): 1) CPP_{class} values summed functional
154 relative abundances mapping to major compound classes (Fig. 1); 2) CPP_{ASALR}: noting high variability in
155 CPP_{class} values across case studies, we implemented a first-pass normalization aiming to account for
156 microbial activity levels, based on CPP_{class} abundances assigned to amino sugars (42), here termed
157 amino sugar adjusted log ratio (ASALR) data; 3) $CPP_{density}$ captured the density of functional relative
158 abundances in close radial proximity to focus biomolecules (Fig. 1); and 4) compound-associated vK
159 coordinates: these data underpin the above measures (i.e., aggregated within major classes, or within
160 close radii of biomolecules) but were also used to consolidate functions with shared vK coordinates for
161 supplementary analyses in selected case studies described below (i.e., differential abundance, correlation
162 networks, and calculating weighted mean vK coordinates within major compound classes). We
163 successfully mapped most SUPER-FOCUS functional relative abundances to vK coordinates (sample
164 ranges 52-84%, means 55-67%; Table 1; SI Appendix Table S3).

165

166 **Human health and disease.** In overview, gut metagenome CPP_{class} (SI Appendix, Figs. S4, S7, S11,
167 S17, Table S3) and $CPP_{density}$ (SI Appendix, Figs. S5-S6, S8-S9, S12-S13, S18-S19, Table S4) data
168 produced strong associations in ACVD and colorectal cancer compared to normal subjects (detailed
169 below). Many patterns observed in CPP_{class} data were reinforced in the putative activity-normalized
170 CPP_{ASALR} measures (Fig. 3; SI Appendix Table S5), and this transformed data format showed stabilized
171 variance across case study datasets. Interestingly, across all CPP_{class}, $CPP_{density}$, and CPP_{ASALR}
172 measurements, when associations were found in both sexes they were always in the same direction (SI
173 Appendix Tables S3-5). In the T2D (female only) and problem behavior case studies, we observed far
174 fewer relationships in the coarse CPP_{class}, CPP_{ASALR} or biomolecule-focused $CPP_{density}$ data. Therefore,

175 we pursued network analyses and differential abundance analyses respectively in these case studies, as
176 illustrative examples of more detailed supplementary analyses. Weighted mean vK-coordinate analyses
177 produced striking associations in ACVD (Fig. 4; SI Appendix, Tables S6-S7), but weaker effects in other
178 case studies (SI Appendix, Figs. S10, S14, S20, Tables S8-S12). The statistical test results described
179 below are detailed in the SI Appendix Tables.
180
181



182
183 **Fig. 2.** Flow chart of compound processing potential (CPP) analyses.
184

185 **Table 1.** Description of case study metagenome datasets (further detail in SI Appendix, Supporting
186 Information).

Case study focus	Main comparison variable or diagnosis groups (and sample/subject numbers*)	Metagenome functional profile characteristics (sample mean ± s.d.)†	Source data reference; Country of origin
Human gut			
Atherosclerotic cardiovascular disease (ACVD)	ACVD (n_F = 53, n_M = 157) or normal healthy (n_F 101, n_M = 69). Total n = 380.	Functions^0 n = 21,117 (9301 ± 3523) Functions^1 n = 10,829 (5051 ± 1843) Total fxn rel abun^1 = 63.2 ± 3.5 % vK coordinates n = 2535 (1509 ± 502)	(35, SRA accession PRJEB21528); China
Colorectal cancer	Colorectal cancer (n_F = 24, n_M = 29) or normal healthy (n_F = 33, n_M = 27). Total n = 113.	Functions^0 n = 12,896 (2809 ± 1924) Functions^1 n = 6932 (1750 ± 1117) Total fxn rel abun^1 = 66.7 ± 2.7 % vK coordinates n = 1954 (711 ± 362)	(36, SRA accession PRJEB6070); France
Type 2 diabetes (T2D) and impaired glucose tolerance (IGT)	T2D with no Metformin treatment (T2D Met-, n = 33), T2D with Metformin (T2D Met+, n = 20), IGT (n = 49), or normal healthy (n = 43). Subjects are females only. Total n = 145.	Functions^0 n = 19,099 (9916 ± 2393) Functions^1 n = 9916 (4490 ± 874) Total fxn rel abun^1 = 64.4 ± 2.2 % vK coordinates n = 2393 (1431 ± 202)	(37, SRA accession PRJEB1786); Sweden
Problem behaviors in children	First principal component (PC1) of anxious-depressive problem behaviors, examined as either numeric scores (n_F = 20, n_M = 17); or high/low PC1 groups‡ (high PC1 n_F = 10, n_M = 8; Low PC1 n_F = 10, n_M = 9). Total n = 37.	Functions^0 n = 20,599 (10282 ± 2481) Functions^1 n = 10,322 (5507 ± 1224) Total fxn rel abun^1 = 54.9 ± 1.3 % vK coordinates n = 2459 (1645 ± 281)	(38, SRA accession PRJNA496479); United States
Soils			
People Cities and Nature (PCaN) forest ecosystem restoration	Soil samples spanned young to old revegetation age, and remnant sites (treated as ordinal variables). Data were separated into pH-based groups: strongly acidic (pH < 4.5, 10-40 yr old, remnant, n = 8); and acidic-neutral soils (4.5 < pH < 7, 11-48 yr old, n = 10). Total n = 18.	Functions^0 n = 36,324 (27,969 ± 1011) Functions^1 n = 18,197 (14,690 ± 396) Total fxn rel abun^1 = 61.7 ± 0.3 % vK coordinates n = 3302 (2965 ± 49)	(39, Aotearoa Genomic Data Repository project AGDR00045); Aotearoa New Zealand
Post-mining forest ecosystem restoration	Soil samples spanned revegetation ages of 6, 12, 22, 31 years, and unmined (UM) samples (treated as ordinal variables). Comprising five age-based groups, each with three replicates. Total n = 15.	Functions^0 n = 30,125 (20,328 ± 767) Functions^1 n = 15,576 (11,051 ± 334) Total fxn rel abun^1 = 61.9 ± 0.3 % vK coordinates n = 3076 (2537 ± 54)	(40, MG-RAST project mgp16379); United States
Australian Microbiome Initiative (AMI) disturbed versus natural	Disturbed (n = 29) or natural (n = 55) soils, comprising temperate climate zone, surface (0-10cm) with 7.5–45% clay content (i.e., avoiding very sandy and very clayey soils). Total n = 84.	Functions^0 n = 37,335 (22,959 ± 1172) Functions^1 n = 18,551 (12,212 ± 553) Total fxn rel abun^1 = 60.8 ± 1.0 % vK coordinates n = 3326 (2700 ± 69)	(41, AMI Data Portal [Data accessed Sep 2022])§; Australia

187 *F = females, M = males. †Number of functions^0 from initial SUPER-FOCUS profiling, versus functions^1
188 (fxn) with available compound information for mapping to vK coordinates. ‡PC1 of problem behaviors
189 were analyzed as numeric values for visualizing and assessing CPP_{class} and $CPP_{density}$ data; and high vs.
190 low PC1 groups for CPP_{ASALR} data. High PC1 values represent more problematic behavior. §AMI Data
191 portal URL: https://data.bioplatforms.com/organization/australian-microbiome.

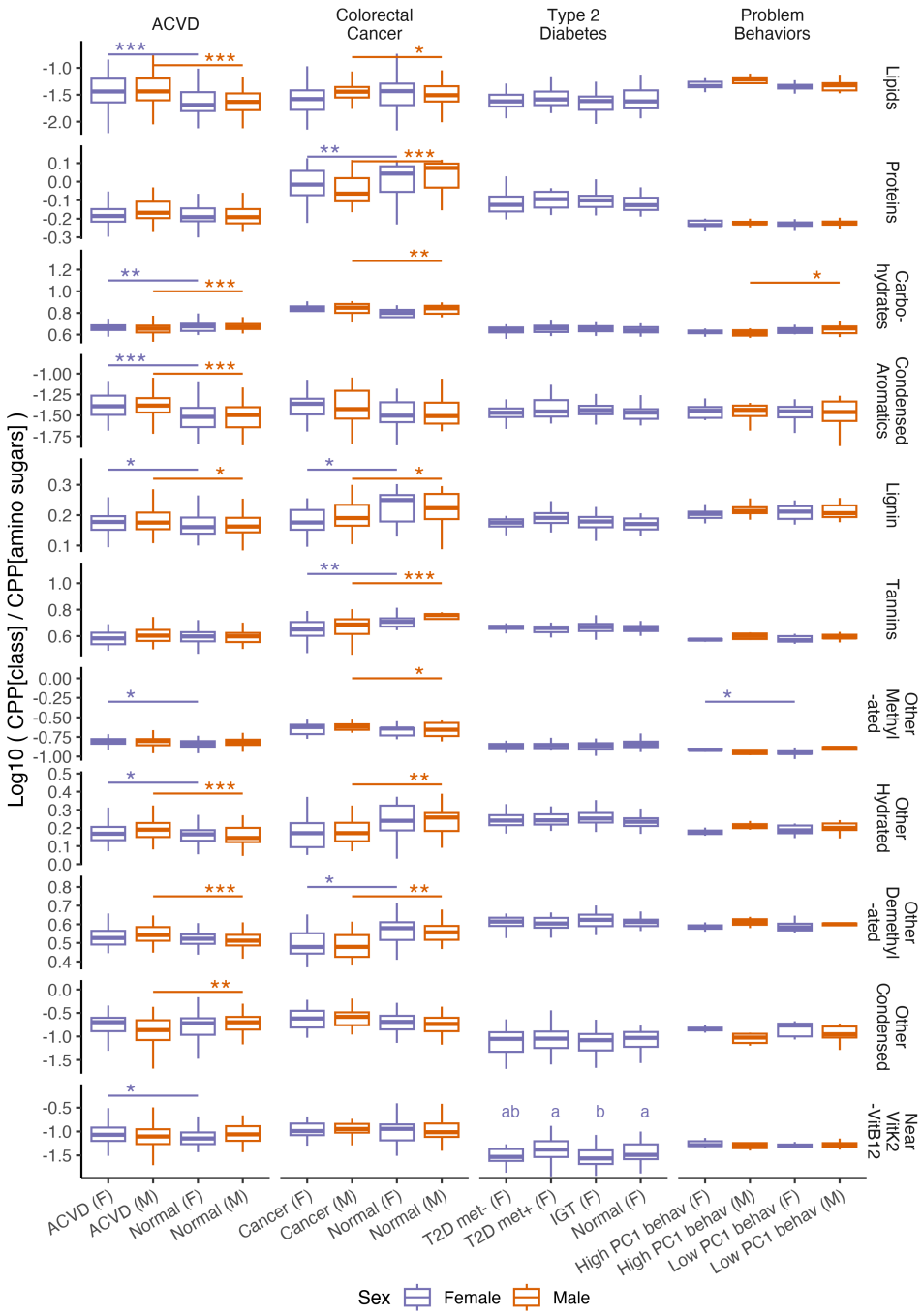
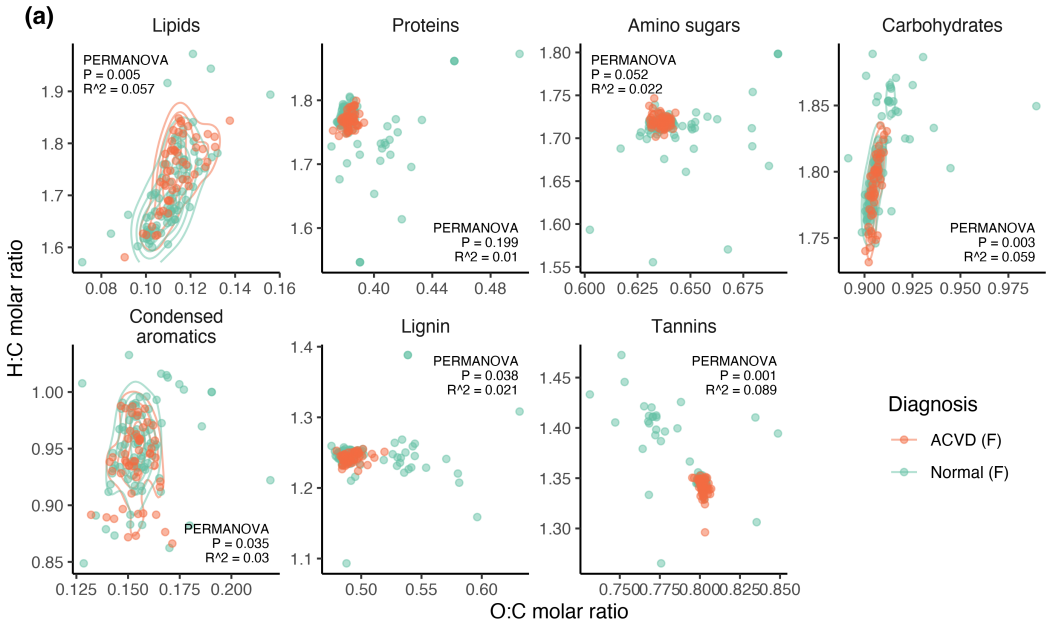
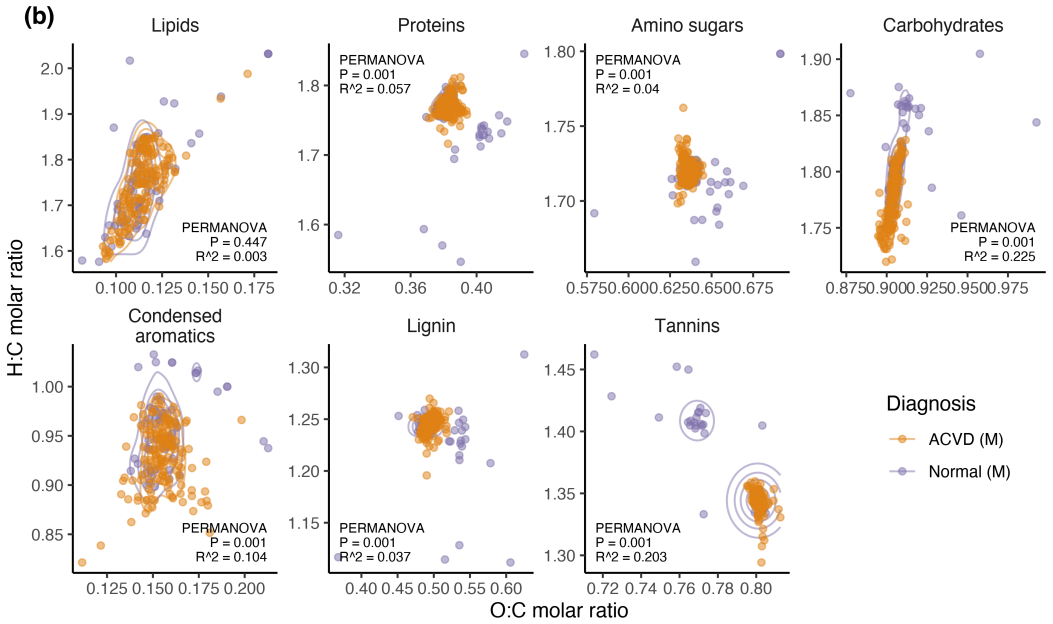


Fig. 3. Amino sugar-adjusted log ratio compound processing potential (CPP_{ASALR}), representing putative microbiota activity-normalized values, in normal healthy and diseased female (F) and male (M) subjects for atherosclerotic cardiovascular disease (ACVD), colorectal cancer, type 2 diabetes (T2D) with and without Metformin treatment (met +/-), impaired glucose tolerance (IGT), and high and low first principal component (PC1) problem behavior values. For visualization purposes outlying values are not shown. However, statistical tests were based on all data (SI Appendix Table S5). Sample sizes are detailed in Table 1. Tests for differences are performed within a single sex. In T2D data, groups not sharing a letter are different.



201



202

203 **Fig. 4.** Weighted mean vK coordinates within compound classes in ACVD and normal subjects in (a)
204 females (ACVD n = 53, normal n = 101), and (b) males (ACVD n = 157, normal n = 69). PERMANOVA
205 and beta-dispersion results are in SI Appendix Tables S6-7. Contour lines indicate the probability density
206 of data points (SI Appendix Supporting Information).

207 ACVD associated with CPP_{class} values, compared to normal subjects, in the form of increased
208 potential metabolism of lipids (in females – F, and in males – M), proteins (M), condensed aromatics (F,
209 M) lignin (F, M), tannins (M), other hydrated compounds (F, M), other demethylated compounds (M), and
210 near vitamin K2-vitamin B12 compounds (F); contrasting with decreased potential metabolism of
211 carbohydrates (F, M) and other condensed compounds (M). Despite all samples initially summing to
212 100% total functional relative abundances, there was less complete conversion to identifiable reactions
213 and total sum CPP_{class} data in females only for ACVD compared to normal cases. Many of these patterns
214 were reinforced in the CPP_{ASALR} data, with ACVD notably associated in both sexes with increased
215 potential metabolism of lipids, condensed aromatics, lignin, and other hydrated compounds; but
216 decreased potential metabolism of carbohydrates. For $CPP_{density}$ measures within 0.05 vK unit radii,
217 ACVD associated with increased potential metabolism of propionate (F), vitamin B12 (F), vitamin B6 (M),
218 vitamin B9 (F, M), vitamin K2 (F, M) and pyruvate (M); but decreased potential metabolism of acetate (M)
219 and glutamate (F, M).
220 Colorectal cancer also associated prominently with CPP_{class} values, specifically increased
221 potential metabolism of lipids (M), amino sugars (F, M), condensed aromatics (M), lignin (M), other
222 condensed compounds (F, M) and near vitamin K2-vitamin B12 compounds (M); but decreased potential
223 metabolism of proteins (F, M) and tannins (F, M). Total sum CPP_{class} data were reduced in male colorectal
224 cancer cases compared to normal. CPP_{ASALR} values in colorectal cancer subjects of both sexes were
225 associated with decreased potential metabolism of proteins, lignin, tannins, and other demethylated
226 compounds. For $CPP_{density}$ measures within 0.05 vK unit radii, colorectal cancer associated with increased
227 potential metabolism of vitamin B2 (M), vitamin B12 (F, M), vitamin B9 (F, M), and pyruvate (M). Weighted
228 mean vK-coordinates varied between colorectal cancer and normal subjects with different centroids in
229 condensed aromatics (F), lignin (M), and tannins (M); and different beta-dispersion or spread in proteins
230 (F, M) and condensed aromatics (F, M).
231 In the female-only T2D case study results, we found CPP_{class} values decreased for other
232 methylated compounds in IGT compared to normal subjects and decreased for near vitamin K2-vitamin
233 B12 compounds in IGT compared to normal, with a reduced total sum CPP_{class} in normal compared to

234 other subjects. No associations were found with $CPP_{density}$ values. CPP_{ASALR} values were decreased for
235 near vitamin K2-vitamin B12 compounds in IGT compared to normal and T2D Met+ subjects. From
236 analysis of weighted mean vK-coordinates among diagnosis groups, a difference in centroids was found
237 only in the compound class of proteins. With such weak results, further investigation was performed via
238 network analysis (SI Appendix Supporting Information) on 20 subjects in each diagnosis group based on
239 inferred correlations between functional relative abundances mapped to unique vK-coordinates. Network
240 analyses were based on commonly observed vK coordinates (present in at least 60% of samples and
241 minimum 2% sum functional relative abundance across samples). Network diagrams (Fig. 5) and
242 structure dendograms (SI Appendix, Fig. S15) display a transition in their complexity, number of nodes
243 and fraction of negative edges (interactions) from simplest in normal and IGT subjects to most complex in
244 T2D Met+ and T2D Met- subjects. Comparing network characteristics for the four groups (normal, IGT,
245 T2D Met-, T2D Met+; n = groups of 20) to a bootstrapped (B = 1000) density distribution of randomly
246 resampled networks (n = 20, drawn from the same pool of 80 subjects) (SI Appendix, Fig. S16, Table
247 S17) we found normal healthy subjects had the lowest fraction of negative edges and the highest degree
248 centralization. Untreated disease, T2D Met-, had the lowest closeness centralization (graph-level inverse
249 of average geodesic distance between nodes); and borderline significant results for the highest fraction of
250 negative edges (negative correlations between vK-coordinates), lowest betweenness centralization
251 (graph-level centrality based on broker positions connecting others), and lowest mean distance (average
252 path length between nodes). In short, normal subjects appear to have far less correlations between vK-
253 coordinates (fewer nodes / vertices), and for the nodes and links that are present they are largely
254 positively correlated and highly interlinked. Whereas T2D Met- (untreated disease) is characterized by a
255 much larger number of negatively correlated vK-coordinate nodes, which on average have shorter links,
256 and are less well connected-up across the whole network.

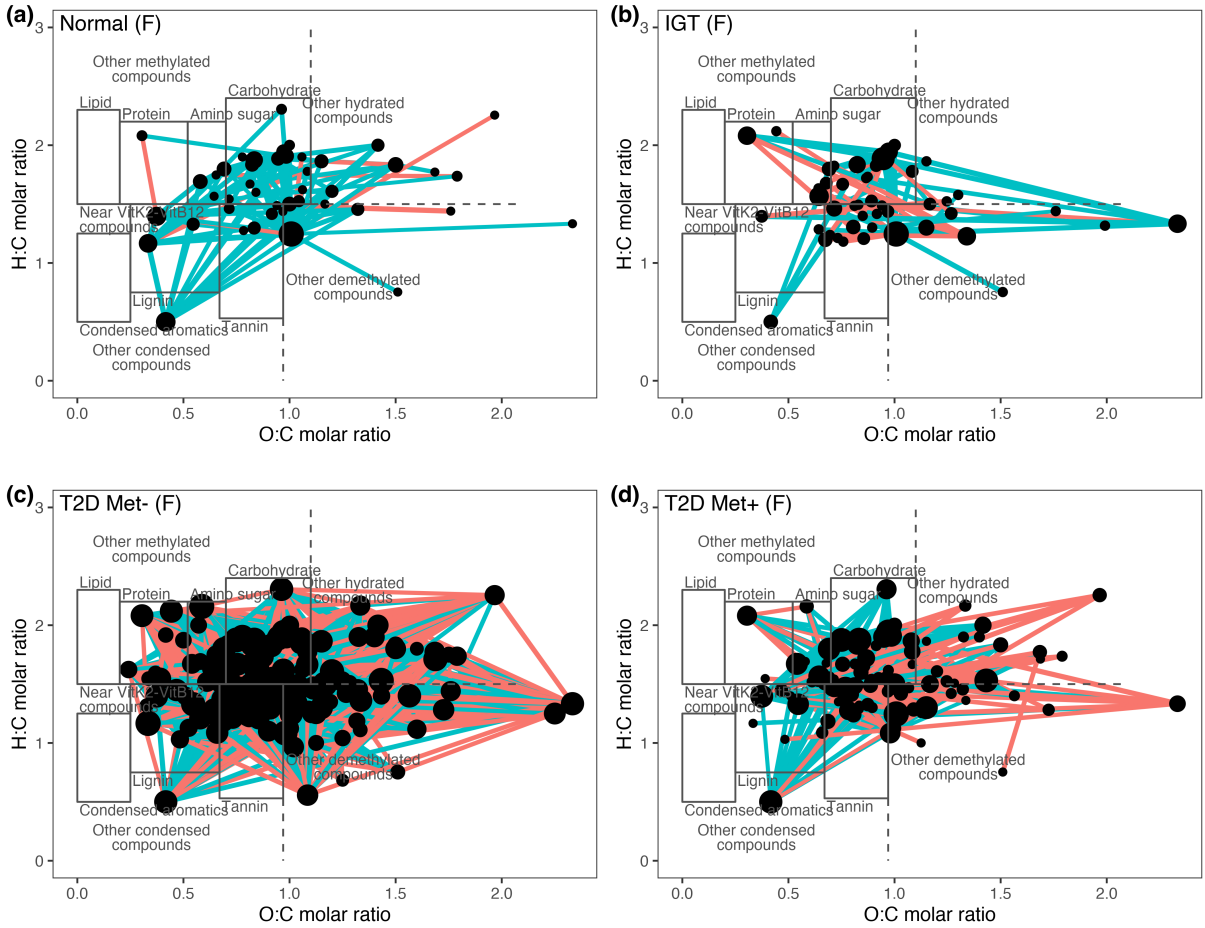


Fig. 5. Network diagrams based on commonly observed vK coordinates for female subjects (n = groups of 20) with diagnoses: (a) normal, (b) impaired glucose tolerance (IGT), (c) type 2 diabetes without Metformin (T2D Met-), and (d) type 2 diabetes with Metformin (T2D Met+). Nodes are located according to unique compound-associated vK coordinates, with size reflecting node degree (number of linked significant correlations). Links between nodes display positive (aqua color) and negative (red color) correlations ($p \leq 0.05$).

265 Problem behaviors displayed few significant associations in CPP_{class} values: increasing PC1 of
266 problem behaviors associated with increased potential metabolism of lipids (M), amino sugars (F), other
267 demethylated compounds (M), and near vitamin K2-vitamin B12 compounds (F); but decreased potential
268 metabolism of carbohydrates (M). For $CPP_{density}$ values, patterns were found in males only: increasing
269 PC1 of problem behaviors associated with increased potential metabolism of vitamin B6 (M) and vitamin
270 B9 (M); but decreased potential metabolism of acetate (M), vitamin B12 (M) and glutamate (M). For
271 CPP_{ASALR} values, the comparison was made between groups of high PC1 versus low PC1 of problem
272 behaviors (for consistent display with other case studies in Fig. 3). High PC1 values associated with
273 increased potential metabolism of other methylated compounds (F); and decreased potential metabolism
274 of carbohydrates (M). Weighted mean vK-coordinates between high PC1 and low PC1 of problem
275 behaviors showed a difference in beta-dispersion within condensed aromatics for females only. To
276 explore this dataset in more detail, we identified differentially abundant functions and vK-coordinates (i.e.,
277 aggregated functional relative abundances via bioenergetic mapping) in high PC1 versus low PC1
278 subjects, separately within each sex. In females, 22 differentially abundant functions, compared to only 2-
279 vK coordinates (with 3 corresponding functions), were identified (SI Appendix, Figs. S21, S23, Table S18,
280 S20). In males, 6 functions compared to 2 vK-coordinates (with 8 corresponding functions) were identified
281 (SI Appendix, Figs. S22, S24, Table S19, S21). Not all functions could be mapped into vK coordinate
282 space. Interestingly, there was no overlap in functions identified directly versus indirectly (from
283 aggregation into vK-coordinates). This means that differential abundance analysis using vK-coordinates
284 can provide entirely different foci for investigation compared to the standard function-level analysis. From
285 the vK-coordinate level analysis, high PC1 (compared to low PC1) females exhibited increased potential
286 metabolism of Uridine phosphorylase (EC 2.4.2.3) (fxn_14491) involved in pyrimidine conversions; and
287 decreased aldehyde lyases dihydroneopterin phosphate phosphatase and dihydroneopterin aldolase (EC
288 4.1.2.25) (fxn_12938; fxn_12942). High PC1 males exhibited increased phosphoenolpyruvate
289 carboxykinase (GTP) (EC 4.1.1.32) (fxn_2926) associated with pyruvate metabolism; and also increases
290 in multiple functions (fxn_821; fxn_2958; fxn_2973; fxn_12703; fxn_12705; fxn_12786; fxn_12788) all
291 involving alcohol dehydrogenase (EC 1.1.1.1) and acetaldehyde dehydrogenase (EC 1.2.1.10), with or

292 without pyruvate-formate-lyase deactivase – involved in degradation of aromatic compounds, biphenyl,
293 tryptophan, and pyruvate metabolism. Results and visualizations for the standard function-level analysis
294 are included for comparison but are not discussed further.
295
296 **Plant-soil systems.** We found remarkable consistency in many observed patterns across the
297 environmental soil case studies. Results reported here are for relative trends with increasing ecosystem
298 maturity (i.e., older revegetation and natural samples) in CPP_{class} (SI Appendix, Figs. S25, S29, S32,
299 Table S3), $CPP_{density}$ (SI Appendix, Figs. S26-S27, S30-S31, S33-34, Table S4), and CPP_{ASALR} (Fig. 6; SI
300 Appendix Table S5). Below, we highlight trends found in at least two of three case studies.

301 In CPP_{class} data we observed: increased potential metabolism of lipids (post-mining, AMI) and
302 condensed aromatics (post-mining, AMI); but decreased potential metabolism of proteins (post-mining,
303 AMI), carbohydrates (post-mining, AMI, with marginal indications in both PCaN soil groups), lignin (post-
304 mining, AMI), other methylated compounds (post-mining, AMI), and other hydrated compounds (AMI,
305 PCaN acidic-neutral soils). Mixed or isolated results included potential metabolism: either decreased
306 (AMI) or increased (PCaN strongly acidic soils) for tannins; decreased for other demethylated compounds
307 (AMI); increased for other condensed compounds (AMI); and decreased (post-mining, PCaN acidic-
308 neutral soils) or increased (PCaN strongly acidic soils) for near vitamin K2-vitamin B12 compounds. Total
309 CPP_{class} compounds appeared to be less well characterized and mapped to functional reactions in AMI
310 natural compared to disturbed samples, but more well characterized in older revegetation (compared to
311 younger revegetation) within the PCaN strongly acidic soils.
312 CPP_{ASALR} results reinforced many patterns observed in the CPP_{class} data: we observed increased
313 potential metabolism of lipids (post-mining, AMI, marginal in PCaN strongly acidic soils) and condensed
314 aromatics (post-mining, AMI); but decreased potential metabolism of proteins (AMI, marginal in post-
315 mining), carbohydrates (post-mining, AMI, PCaN acidic-neutral soils), and other methylated compounds
316 (post-mining, AMI). Isolated results included, potential metabolism: increased for other condensed
317 compounds (AMI); but decreased for lignin (AMI), tannins (AMI), other hydrated compounds (AMI), and

other demethylated compounds (AMI). Differing from CPP_{class} results, for CPP_{ASALR} potential metabolism of near vitamin K2-vitamin B12 compounds increased (post-mining).

$CPP_{density}$ results were also quite consistent across case studies: we observed increased potential metabolism of vitamin B9 (AMI, marginal in post-mining and PCaN strongly acidic soils) and vitamin K2 (post-mining, AMI, marginal in PCaN strongly acidic); but decreased potential metabolism of acetate (post-mining, AMI, PCaN acidic-neutral soils), propionate (AMI), vitamin B2 (post-mining, marginal in PCaN strongly acidic soils), vitamin B12 (post-mining, AMI, PCaN strongly acidic soils), vitamin B6 (post-mining, AMI), glutamate (post-mining, AMI, PCaN acidic-neutral soils), and pyruvate (AMI). While butyrate $CPP_{density}$ values were low and indistinguishable across all samples using 0.05 vK unit radial buffers, for an example comparison we tested near butyrate $CPP_{density}$ using a larger 0.1 vK unit buffer in the AMI soils and found increased levels in natural compared to disturbed soils (SI Appendix, Fig. S35).

Post-mining and AMI samples displayed significant and mostly consistent directional shifts in weighted mean vK-coordinate centroids across all compound classes (vK mapping zones) considered, while PCaN acidic to neutral soils also displayed shifts within carbohydrates and condensed aromatics (Fig. 7; SI Appendix, Fig. S28, Tables S13-S16). Ecosystem maturity explained 60–92% of the variation in weighted mean vK-coordinates in post-mining soils (Fig. 7a, SI Appendix Table S15). The following general patterns emerged with increasing ecosystem maturity: lipids became more reduced (lower oxygen content), proteins became more hydrated, amino sugars became more dehydrogenated or condensed, carbohydrates became more condensed (post-mining, PCaN acidic-neutral) or reduced (AMI), condensed aromatics became more condensed, lignin showed mixed trends (dehydrogenation in post-mining, reduction in AMI), and tannins became more demethylated. Except for carbohydrates and amino sugars, these trends largely represented an outward extension of sample profile mapping into vK coordinate space with older ecosystems.

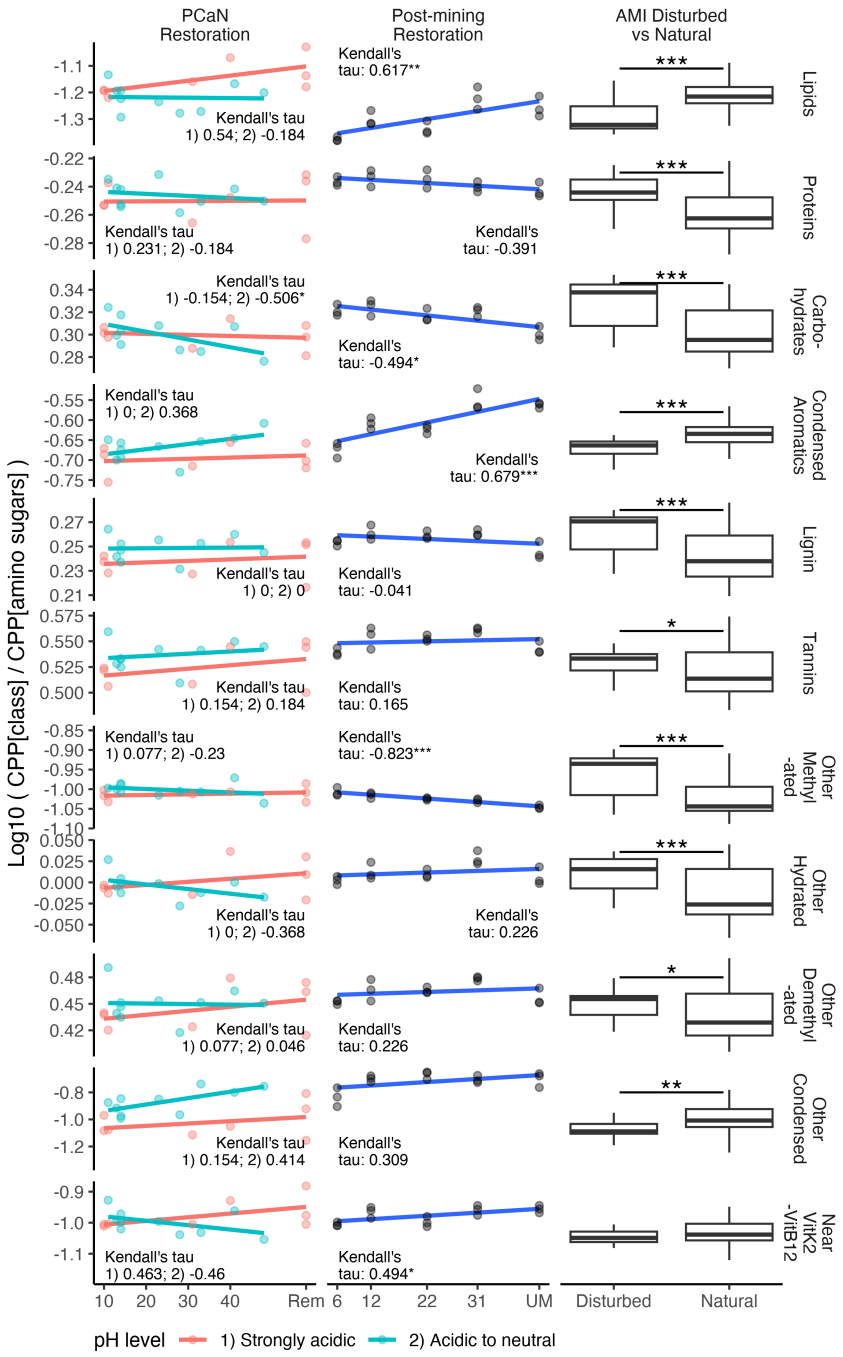


Fig. 6. Amino sugar-adjusted log ratio compound processing potential (CPP_{ASALR}), representing putative microbiota activity-normalized values, from People Cities and Nature (PCaN), post-mining restoration soils, and Australian Microbiome Initiative (AMI) disturbed versus natural soils. Linear trends were used for visualization purposes. However, Kendall's tau correlation tests (suited to ordinal data) were applied. Sample sizes are detailed in Table 1.

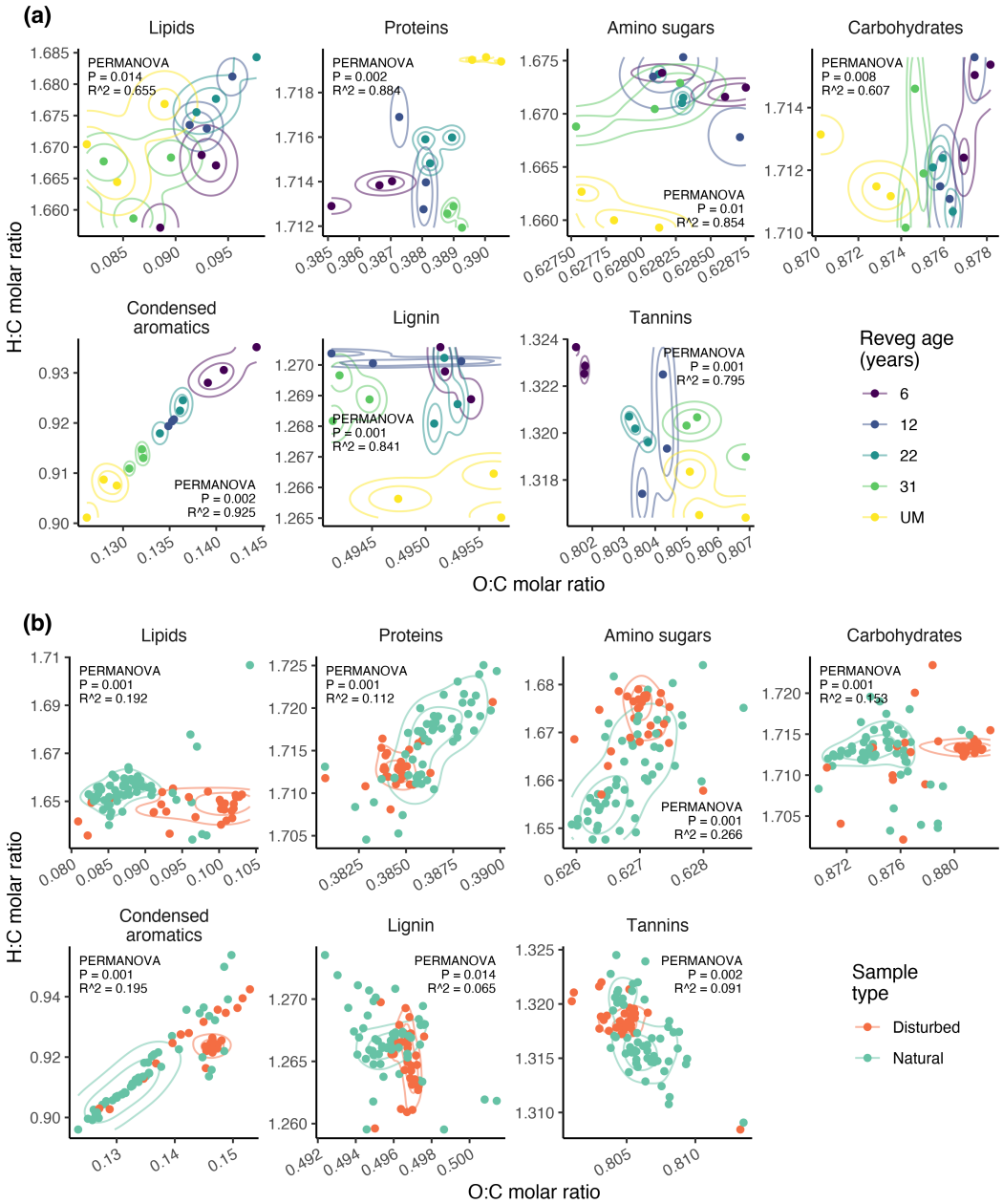


Fig. 7. Weighted-mean van Krevelen coordinates within select compound classes display significant shifts with maturity of plant-soil ecosystems. Patterns are from (a) post-mining forest ecosystem restoration soil samples (n = age-based groups of 3), and (b) AMI disturbed (n = 29) vs natural (n = 55) soil samples (b). PERMANOVA and beta-dispersion results are in SI Appendix, Table S15-S16.

356 **Consistency in bioenergetic 'topography' mapping.** $CPP_{density}$ plots visualizing local polynomial
357 regression fitting ('loess') smoothed vK mapping profiles for increasing radii buffer areas displayed striking
358 consistency of form within each compound class, with alternative expressions in either gut (SI Appendix,
359 Figs. S5, S8, S12, S18) or soils (SI Appendix, Figs. S26, S30, S33).
360

361 **Discussion**

362 This work reveals a meaningful bioenergetic basis to the development and differentiation of community-
363 scale microbiomes. Our CPP metrics quantified putative shifts in a microbial community's relative
364 proficiency to process different types of compounds across gradients of human- and environmental-
365 health. Across the case studies we found significant patterns of CPP association at varying resolutions:
366 within major classes of compounds, near focus biomolecules, and for unique vK coordinates. In the
367 environmental soil metagenomes, surprising consistency in CPP profiles suggests they may link to plant-
368 soil system conditions in coherent and predictable ways. Our findings align with the notion that microbiota
369 are shaped by the bioenergetic status of prevailing substrates and micro-environments, and this
370 information is simultaneously recorded in their metagenomes. From a methodological perspective, our
371 compound-focused bioenergetic mapping approach demonstrates new pathways for assessing and
372 interpreting microbial systems, capable of supporting ongoing efforts to define healthy microbiomes. For
373 example, aggregating functions via vK-coordinates can provide entirely different foci for investigation
374 compared to standard function-level analyses.
375

376 **Patterns found in human health.** We found strong links in both sexes between ACVD and increased
377 potential metabolism of lipids, condensed aromatics, lignin, other hydrated compounds, vitamin B9 and
378 vitamin K2; and decreased potential metabolism of carbohydrates and glutamate. Links between high-fat
379 diets and ACVD are well established (43). Gut microbiota can contribute to ACVD by metabolizing the
380 dietary lipid phosphatidylcholine, with subsequent production of harmful trimethylamine oxide (44).
381 Polycyclic aromatic hydrocarbons are known risk factors in ACVD (45, 46). Lignin (different to lignan) is a
382 complex and ubiquitous structural plant polymer, considered predominantly insoluble fiber, with content

383 ranging from 1-2 g/100g in vegetables, fruits and cereals, up to 30-40 g/100g in nut shells and stone fruit
384 kernels (47). Lignin inhibits the enzymatic activity of α-glucosidase, delaying carbohydrate digestion and
385 absorption, with potential for low post-meal blood sugar levels (48). Acute hypoglycemia (low blood
386 sugar) can trigger cardiac events (49), with greater adverse risks in subjects with significant comorbidities
387 (e.g., T2D, ACVD). Our finding for decreased carbohydrates in ACVD might align with these impacts on
388 blood sugar. Alternatively, we speculate that result might be symptomatic of a more Westernized diet
389 (high in animal protein, sugar, starch, and fat – and lower in carbohydrate content than a plant-rich diet;
390 44). Excessive vitamin B9 (folate) is associated with ACVD risk via a non-linear u-shaped dose-response
391 relationship (50). Non-linear u-shaped dose-response relationships are common in biological systems
392 (i.e., hormesis, deficiency-sufficiency-toxicity) (51). Similarly, our results linking increased vitamin K2 with
393 ACVD appear contrary to recent opinion (52), although non-linear u-shaped dose-responses have also
394 been observed (53). Vitamin K2 is commonly found in fermented foods which are less common in
395 Western diets (54). Here, the ACVD case study was based on Chinese subjects whose diets potentially
396 contained higher quantities of fermented foods including vitamin K2. Possibly, these subjects were more
397 susceptible to adverse effects if excessive levels of vitamin K2 were reached. Our results linking reduced
398 glutamate with ACVD appear contrary to findings from large US cohort studies which found higher
399 glutamate levels, and lower glutamine:glutamate ratios, correlated with increased ACVD risk (55, 56).
400 Dietary proteins are a major source of glutamate (57). However, some ethnic populations may have
401 inadequate protein in their diets (58). Interestingly, in female adults from rural western China with
402 inadequate (and largely plant-derived) protein intake, increasing animal protein associated with reduced
403 risk of hypertension (58). Together, these observations suggest that a u-shaped dose-response may also
404 operate for animal-based proteins, glutamate, and ACVD risk.
405 In the colorectal cancer subjects (from France) we found consistent associations across CPP
406 metrics in both sexes with increased potential metabolism of amino sugars, vitamin B12 and B9; and
407 decreased potential metabolism of proteins and tannins. Colon cancer has been associated with low
408 dietary fiber, low fruit and vegetable consumption, and high red meat consumption (43). Our results were
409 consistent with reports for anticancer activity, including protective effects against colorectal cancer, from

410 some tannins or polyphenols (e.g., components in green and black tea, resveratrol in red wine and
411 grapes) (59). Unfortunately, our data did not distinguish between animal- and plant-based protein.
412 However, red meat is widely consumed in France, with 41% of males and 24% of females consuming
413 above guideline levels (60). Amino sugars are sugar molecules with at least one hydroxyl group
414 substituted by an amino group. In biological systems, they are formed by catalytic activity acting on amino
415 acids (glutamate, glutamine – building blocks of protein) to transfer an amino functionality to a sugar
416 phosphate or sugar nucleotide (61). Therefore, both glucose (sugar) and amino acids contribute to amino
417 sugar formation. Meanwhile, metabolism of both glucose and amino acids plays a key role in colorectal
418 cancer development (62). Possibly, our finding of increased potential metabolism of amino sugars with
419 colorectal cancer may reflect dysregulated activity of glucose and amino acids with the product of their
420 interaction (amino sugars) recorded by the gut microbiome. Consistent with our findings, vitamin B9
421 (folate or folic acid) and vitamin B12 supplementation have been associated with increased risk of
422 colorectal cancer (63).

423 In the T2D case study, our lack of clear findings linked to major compound classes or focus
424 biomolecules seems consistent with reports that T2D is a complex, multifaceted, highly heterogeneous
425 polygenic disease with uncertain etiology (64). We found untreated (Met-) T2D exhibited an anomalous
426 and complex CPP network, including a high number of negative correlations (indicating negative
427 feedbacks) ranging widely across vK coordinate space (i.e., covering a spectrum of compounds and
428 bioenergetic status). Diagnosis of T2D is based on elevated blood glucose, primarily arising from insulin
429 resistance and inadequate insulin secretion (37). However, a clear diagnostic test for T2D is lacking,
430 except by exclusion of other causes (64). A range of factors including genetics, dietary habits, sedentary
431 lifestyle, and gut microbiota are involved in disease development (37). Possibly, with more detailed
432 examination, diagnostic relationships (e.g., correlations, ratios) might be uncovered in relative abundance
433 patterns of compound-associated vK-coordinates underpinning the anomalous T2D Met- network.
434 In the problem behavior case study, different compound associations were observed in female
435 and male children. Increased problem behavior (higher PC1) in females associated with increased
436 potential metabolism of amino sugars, other methylated compounds, and near vitamin K2-vitamin B12

437 compounds. From differential abundance analysis, high PC1 females exhibited decreased
438 dihydroneopterin aldolase—an enzyme involved in converting dihydroneopterin (a molecule involved in
439 folate biosynthesis) into other compounds (65). Excessive serum levels of dihydroneopterin have been
440 associated with major depression (66). Possibly, in case study subjects, reduced levels of
441 dihydroneopterin-degrading enzyme have promoted accumulation of dihydroneopterin in association with
442 problem behaviors. High PC1 females also exhibited increased uridine phosphorylase, an enzyme
443 involved in pyrimidine metabolism that converts uridine to uracil (67), therefore possibly degrading uridine
444 levels in those subjects. Uridine is linked to energy metabolism and glutamate-mediated excitatory
445 neurotransmission in the brain, and supplemental uridine treatments have been used to reduce
446 depressive symptoms in adolescents (68). In males, increased PC1 associated with increased potential
447 metabolism of lipids, vitamins B6 and B9; and decreased potential metabolism of carbohydrates, acetate,
448 vitamin B12, and glutamate. Vitamins B12 (cobalamin) and B9 (folate) are recognized precursors involved
449 in forming key neurotransmitters dopamine, noradrenaline (norepinephrine), and serotonin (69). These
450 three neurotransmitters occur in the vicinity of vitamin B12 and K2 in vK coordinate space. Vitamin B12
451 deficiency has been associated with depressive disorders in older subjects (70). Glutamate's role as a
452 key neurotransmitter is described in earlier text. High PC1 males also exhibited increased
453 phosphoenolpyruvate carboxykinase—an enzyme involved in cataplerosis, or removal of intermediate 4-
454 and 5-carbon compounds from the TCA cycle (71, 72). These intermediates are removed because they
455 cannot be fully oxidized for energy metabolism within the TCA cycle, but are converted elsewhere to
456 glucose, fatty acids or amino acids (72). High PC1 males also exhibited increases in alcohol
457 dehydrogenase, acetaldehyde dehydrogenase, and pyruvate-formate-lyase deactivase, variously
458 involved in pyruvate metabolism, degradation of aromatics and biphenyl, and tryptophan catabolism. Key
459 processes of energy metabolism involving glucose, lipids, protein and the TCA cycle (via keystone
460 molecules pyruvate, acetyl-CoA, and glutamate) have been implicated in major depressive disorder,
461 although precise pathways of pathogenesis are still unclear (73).
462

463 **Patterns found in plant-soil systems.** Our findings point to generalizable patterns with older
464 ecosystems for increasing CPP associated with lipids, condensed aromatics, vitamin B9, and vitamin K2;
465 and decreasing CPP associated with proteins, carbohydrates, lignin, other methylated compounds, other
466 hydrated compounds, acetate, vitamin B12, vitamin B6, and glutamate. Drivers of shifting CPP in soils are
467 expected to include changing: 1) composition of biota and biotic materials including plants, organic debris,
468 and re-assembly of invertebrate and microbial communities, and 2) soil abiotic conditions due to plant-soil
469 feedbacks (e.g., pH, nutrients, organic carbon content, temperature, moisture regime) (74, 75). This
470 includes macro-environmental influences with development of vegetation structure and canopy cover
471 (e.g., shading, rainfall interception, altered drainage). CPP values also likely reflect a dynamic balance
472 between resource availability and use by microbiota. For example, we might expect greater accumulation
473 of lignin in soils of older ecosystems due to plant inputs such as dead roots, bark, leaf litter, and other
474 structural plant residues. However, we observed reduced CPP for lignin in these sample types. Fungi are
475 major lignin degraders (76) and fungal communities vary with ecosystem disturbance and abiotic
476 conditions (77). Interestingly, our results were counter to expectations for elevated fungal decomposition
477 of lignin in older ecosystems. Reforestation with native mixed-species can produce higher levels of
478 recalcitrant soil organic matter (78) (e.g., humic acid which is hard to decompose and maps to lignin in vK
479 space). Our CPP metrics are relative and compositional (based on functional relative abundances
480 summing to a maximum of 100%), so it may be that in relative terms, the metabolic foci of microbiota are
481 shifted to processing other materials. Or possibly, structural plant materials may be more accessible for
482 degradation in disturbed (e.g., agricultural) soil environments, depending on plant residue management,
483 nutrient availability and other factors.

484 We expect some CPP quantities are driven primarily by plant material inputs. For example, soils
485 from more mature ecosystems in temperate climates, represented in samples from AMI and post-mining
486 (in the Appalachian Plateau, southwestern Virginia USA; 40, 79), displayed a positive relationship with
487 CPP for lipids and condensed aromatics. These two compound classes are represented in plant-based
488 essential oils and volatile, aromatic organic compounds. Oils are found in high densities in much of the
489 fire-adapted Australian flora (unlike New Zealand flora) (80). Increased CPP for lipids might also arise

490 due to increased density of energy storage linked to primary production, or more active plant signaling in
491 response to abiotic stress (81). High levels of lipids and condensed aromatics in mature ecosystem soils
492 could also be a result of increasing plant investment into defensive mechanisms via antimicrobial
493 essential oils (82), and volatile and aromatic secondary defense compounds induced by herbivory
494 (typically by invertebrates) (83).

495 Shifting weighted mean vK coordinates across many compound classes (in AMI and post-mining)
496 suggests broad changes in the composition of microbial substrates with more mature ecosystems. The
497 changing composition of the microbiota itself may contribute to this. Carbohydrates are of interest due to
498 the potential contribution of plant-based material to human diet, and CPP_{class} values for carbohydrates
499 were consistently assigned the largest sum of functional relative abundances in the human gut samples.
500 With more mature ecosystems, CPP for carbohydrates decreased in relative terms, but weighted vK
501 coordinates suggest carbohydrate CPP shifts towards favoring processing materials with reduced oxygen
502 content per unit of carbon. There is likely to be global variation in environmental soil CPP driven by soil
503 abiotic factors and changing biota (vegetation and animals), previously outlined.
504

505 **Potential environment-human health links.** This work opens new avenues for investigating
506 environment-human health connections because environments will vary in their production of human
507 health-associated compounds. Moreover, varying environmental microbiota exposures may supply
508 modulating CPP profiles for colonizing or transient impacts to human microbiomes (e.g., skin, airway,
509 gut), which are intimately linked to our health. We show CPP patterns imprinted in environmental soil
510 metagenomes are linked with the maturity of plant-soil systems and abiotic factors such as soil pH. We
511 also show that CPP measures are significantly linked to human health and disease. However, we urge
512 caution in attempting to directly translate CPP trends in plant-soil environments (e.g., vitamins B12, B6,
513 B9, K2, glutamate) to infer possible implications for gut-associated human health. We stress that non-
514 linear, u-shaped dose-response relationships (51) are common and relevant in the context of
515 environmental exposure-human health links. Also, the gut represents a more tightly controlled micro-
516 environment (redox, pH, etc.) unsuited to many environmental microbes. Example evidence for potential

517 environment-human transfer of microbial CPP comes from Endomicrobia species found in oral microbiota
518 of indigenous peoples from central Australia (84). Endomicrobia species provide energetic advantage for
519 cellulose digestion in the guts of termites and wood-eating insects—and transfer to humans has occurred
520 likely through use of termites and termite mounds in traditional food and medicine (84). Speculatively, our
521 results suggest that if broad supplementation of human microbiota CPP capacity is required (spanning a
522 range of health-supporting biomolecules), this may require exposure to multiple types of environments.
523 However, certain environment types may provide more targeted microbiota CPP supplementation.
524
525 **Limitations.** There are important limitations in this study in addition to those already stated. CPP metrics
526 do not measure actual compounds; rather, they quantify conceptual 'meta-compounds' or assemblies of
527 elements based on functional reaction-level summary weighted mean O:C and H:C ratios consistent with
528 compounds of interest. Quantification occurred via mapping into vK space and aggregating functional
529 relative abundances into major compound classes, near focus biomolecules, or at unique vK coordinates,
530 to assess CPP structural profiles of metagenomes. Conceptually, mean reaction-level attributes (O:C and
531 H:C ratios) represent a mid-point of chemical transformation mediated by microbiota (i.e., interpreted as
532 'X% of functions were involved in processing compound/biomolecule type Y'). Stoichiometry rules
533 determine that reaction inputs and products will have balanced O, H, and C atomic counts. However,
534 different mean O:C and H:C ratios can arise due to uncounted O and H atoms in non-C containing
535 species (O:C, H:C values become undefined). For future work, the CPP mapping algorithm could be
536 readily adjusted to separately target reaction inputs, or products, or individual chemical species. Our
537 coarse compound classes did not distinguish (for example) plant versus animal proteins or high-fiber
538 versus low-fiber carbohydrates. Finer-resolution vK mapping zones would increase the precision of
539 results. We could not discern CPP differences for butyrate between sample types using 0.05 vK radii.
540 Butyrate is often present at low concentrations in the gut compared to other SCFAs, with rapid
541 consumption by colonocytes (23). The volatility of butyrate (85) may make it susceptible to loss from soils.
542 Our butyrate $CPP_{density}$ profiles spanning large to small vK radii may depict source-sink dynamics.
543 Compound mapping using O, H, and C content enable exhaustive and compartmented mapping within vK

544 coordinate space, however this represents a simplified, imperfect approach. Multi-element compound
545 mapping would offer increased precision (86) and may be developed to provide exhaustive and
546 compartmented mapping across a range of compound types. Mapping of SUPER-FOCUS functions was
547 incomplete (sample functional relative abundances ranged from 52-84%, with means 55-67%). This may
548 be improved with future algorithm refinement. CPP measures used here were relative, not absolute.
549 Further work is required to refine our first-pass ASALR normalization, examine CPP-disease links in wider
550 ethnic populations, and explore other potential explanatory variables not considered in our analyses. Like
551 many microbiome studies, our analyses do not permit causal insight to interpret whether increased or
552 decreased CPP may facilitate or follow disease. For example, excessive CPP may produce metabolites at
553 toxic levels, or degrade substrates leading to deficiency. Reduced CPP measures might correspond to
554 dietary deficiencies or suppression of functional pathways due to dysregulated environmental conditions.
555 Nonetheless, observed CPP trends may assist hypothesis-building and prioritizing mechanistic research.
556 We suggest that future work might address these limitations.
557
558

559 **Materials and Methods**

560

561 **Case study datasets.** Metagenomics samples used are summarized in Table 1 and further described in
562 SI Appendix Supporting Information.

563

564 **CPP mapping approach.** CPP values were derived via the following steps (Fig. 2; further details are in
565 SI Appendix Supporting Information)

566 1. Shotgun metagenomics raw sequences were accessed, and bioinformatic steps were run on Flinders
567 University DeepThought high-performance computing facility (87).
568 2. Raw sequence data were inspected using FastQC (v0.11.9; 88) and quality control trimming
569 performed using Fastp (v0.23.2; 89).
570 3. Functional potential profiles were derived from good quality read 1 sequences using SUPER-FOCUS
571 (33) software, linked to the Diamond sequence aligner (v0.9.19; 90) and version 2 100% identity-
572 clustered reference database (100_v2; https://github.com/metageni/SUPER-FOCUS/issues/66).
573 Where subjects/samples were represented by multiple sequence files, the combined SUPER-FOCUS
574 outputs were normalized so that the total functional relative abundances summed to 100% in each
575 subject/sample.
576 4. Every SUPER-FOCUS function (output row) was translated to one or more corresponding chemical
577 reaction(s) using a purpose-built R-script algorithm based on ModelSEED database lookup tables
578 (from https://github.com/ModelSEED/ModelSEEDDatabase; accessed 10 Aug 2022). The algorithm
579 sought matches based on either: full matching of functional hierarchies (using subsystem-class, -
580 subclass, -name and -role); detection of EC number; or matching of SUPER-FOCUS function name
581 within ModelSEED lookup tables for reactions (reaction name or alias), subsystems (role), or
582 reaction-pathways (external reaction name).
583 5. Every chemical reaction was converted to reaction-level mean vK coordinates (O:C and H:C molar
584 ratios), considering all C-containing reaction input and product compounds and weighted according to

585 reaction stoichiometry. Compounds not containing C were ignored due to undefined O:C and H:C
586 ratios. Data for compounds were based on Hill system chemical formulae in protonated form.
587 6. Overall mean vK coordinates were calculated for each functional output row of the SUPER-FOCUS
588 functional relative abundance table via averaging one or more associated chemical reactions. From
589 samples initially summing to 100% functional relative abundances, typically between 50–80% of
590 SUPER-FOCUS functions were identified and translated to weighted mean compound-associated vK
591 coordinates.
592 7. In vK coordinate space, we analyzed the spatial assignment of functional relative abundances to
593 derive the following CPP data types:
594 a. CPP_{class}: represented major compound classes based on pre-defined zones from (32) (see
595 Fig. 1; SI Appendix, Tables S1).
596 b. CPP_{ASALR}: to address large variation in CPP_{class} values (that impeded comparisons across
597 study groups) we considered normalization for microbial activity may be needed. Amino
598 sugars have previously been used as a biomarker of microbial residue turnover as they are
599 major components of bacterial and fungal cell walls (peptidoglycan and chitin) (28).
600 Therefore, we implemented preliminary putative 'activity-normalization' by dividing CPP_{class}
601 values for all other compound classes by the CPP_{class} value for amino sugars, followed by a
602 variance-stabilising log10-transformation. These data were denoted amino sugar-adjusted log
603 ratio (CPP_{ASALR}) values.
604 c. $CPP_{density}$: captured functional capacity within radial buffers of varying proximity (radii of 0.05,
605 0.1, 0.15, 0.2, 0.25 vK units) to focus biomolecules (Fig. 1; SI Appendix, Tables S2).
606 Functional relative abundances within radial buffers were summed then divided by the
607 respective area in vK units^2.
608 d. Unique compound-associated vK coordinates: Each SUPER-FOCUS functional row was
609 translated to corresponding vK coordinates (as used for the above spatial assignments).
610 Further supplementary analyses were undertaken at the vK coordinate level, including
611 network analyses (in the T2D case study), differential abundance analysis (in the problem

behavior case study), and weighted-mean vK coordinate analysis within major compound zones.

Data visualization and statistical analyses. Further detail of visualization and statistical testing using standard approaches are provided in SI Appendix Supporting Information.

Acknowledgments

We acknowledge the authors of previously published metagenomics datasets used here which enabled comparison of diverse samples spanning human and environmental health. We acknowledge First Nations Peoples and traditional custodians of the lands from which the case study datasets originate. This work was supported by the People, Cities and Nature research program (Aotearoa New Zealand Ministry of Business, Innovation and Employment, grant UOWX2101), Marsden Fund Fast-Start grant funding (Royal Society Te Apārangi), funding from the German Research Foundation iDiv ([DFG]–FZT 118, 202548816; Ei 862/29-1; Ei 862/31-1; Ei 862/27-1; HE 8266/4-1), and the Australian Research Council (grant numbers LP190100051, LP190100484). We acknowledge the contribution of the Australian Microbiome consortium in the generation of data used in this publication. The Australian Microbiome Initiative is supported by funding from Bioplatforms Australia and the Integrated Marine Observing System (IMOS) through the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS), Parks Australia through the Bush Blitz program funded by the Australian Government and BHP, and the CSIRO.

References

1. C. Averill *et al.*, Defending Earth's terrestrial microbiome. *Nature Microbiology* **7**, 1717-1725 (2022).
2. S. Banerjee, M. G. A. van der Heijden, Soil microbiomes and one health. *Nature Reviews Microbiology* 10.1038/s41579-022-00779-w (2022).
3. J. A. Gilbert *et al.*, Current understanding of the human microbiome. *Nature Medicine* **24**, 392 (2018).
4. J. L. Round, S. K. Mazmanian, The gut microbiome shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology* **9**, 313-323 (2009).
5. N. W. Sokol *et al.*, Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nature Reviews Microbiology* 10.1038/s41579-022-00695-z (2022).
6. M. Delgado-Baquerizo *et al.*, Multiple elements of soil biodiversity drive ecosystem functions across biomes. *Nature Ecology & Evolution* **4**, 210-220 (2020).
7. B. A. Daisley *et al.*, Emerging connections between gut microbiome bioenergetics and chronic metabolic diseases. *Cell Reports* **37**, 110087 (2021).
8. P. Trivedi, C. Mattupalli, K. Eversole, J. E. Leach, Enabling sustainable agriculture through understanding and enhancement of microbiomes. *New Phytol* **230**, 2129-2147 (2021).
9. N. Fierer, Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* **15**, 579-590 (2017).
10. L. L. Barton, D. E. Northup, *Microbial Ecology* (Wiley-Blackwell, 2011), https://doi.org/10.1002/9781118015841, pp. 407.
11. K. Zengler, L. S. Zaramela, The social network of microorganisms — how auxotrophies shape complex communities. *Nature Reviews Microbiology* **16**, 383-390 (2018).
12. M. Eisenstein, The hunt for a healthy microbiome. *Nature* **577**, S6-S8 (2020).
13. J. E. Belizario, M. Napolitano, Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. *Frontiers in Microbiology* **6** (2015).
14. L. Flandroy *et al.*, The impact of human activities and lifestyles on the interlinked microbiota and health of humans and of ecosystems. *Science of The Total Environment* **627**, 1018-1038 (2018).
15. M. I. Roslund *et al.*, A Placebo-controlled double-blinded test of the biodiversity hypothesis of immune-mediated diseases: Environmental microbial diversity elicits changes in cytokines and increase in T regulatory cells in young children. *Ecotoxicology and Environmental Safety* **242**, 113900 (2022).

16. X. Sun *et al.*, Harnessing soil biodiversity to promote human health in cities. *npj Urban Sustainability* **3**, 5 (2023).
17. M. Delgado-Baquerizo *et al.*, Ecological drivers of soil microbial diversity and soil biological networks in the Southern Hemisphere. *Ecology* **99**, 583-596 (2018).
18. Y. Litvak, M. X. Byndloss, A. J. Bäumler, Colonocyte metabolism shapes the gut microbiota. *Science* **362**, eaat9076 (2018).
19. F. Rivera-Chávez, C. A. Lopez, A. J. Bäumler, Oxygen as a driver of gut dysbiosis. *Free Radical Biology and Medicine* **105**, 93-101 (2017).
20. J. W. Wilson, D. Shakir, M. Batie, M. Frost, S. Rocha, Oxygen-sensing mechanisms in cells. *The FEBS Journal* **287**, 3888-3906 (2020).
21. P. Hinsinger, A. G. Bengough, D. Vetterlein, I. M. Young, Rhizosphere: biophysics, biogeochemistry and ecological relevance. *Plant and Soil* **321**, 117-152 (2009).
22. O. Husson, Redox potential (Eh) and pH as drivers of soil/plant/microorganism systems: a transdisciplinary overview pointing to integrative opportunities for agronomy. *Plant and Soil* **362**, 389-417 (2013).
23. G. den Besten *et al.*, The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research* **54**, 2325-2340 (2013).
24. Y. Zhou, N. C. Danbolt, Glutamate as a neurotransmitter in the healthy brain. *Journal of Neural Transmission* **121**, 799-817 (2014).
25. K. Torii, H. Uneyama, E. Nakamura, Physiological roles of dietary glutamate signaling via gut–brain axis due to efficient digestion and absorption. *Journal of Gastroenterology* **48**, 442-451 (2013).
26. L. R. Gray, S. C. Tompkins, E. B. Taylor, Regulation of pyruvate metabolism and human disease. *Cellular and Molecular Life Sciences* **71**, 2577-2604 (2014).
27. H. Mallick *et al.*, Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nature Communications* **10**, 3136 (2019).
28. C. Diener, S. M. Gibbons, O. Resendis-Antonio, MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota. *mSystems* **5**, 10.1128/msystems.00606-00619 (2020).
29. V. Neveu, G. Nicolas, A. Amara, R. M. Salek, A. Scalbert, The human microbial exposome: expanding the Exposome-Explorer database with gut microbial metabolites. *Scientific Reports* **13**, 1946 (2023).
30. D. R. Garza, M. C. van Verk, M. A. Huynen, B. E. Dutilh, Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nature Microbiology* **3**, 456-460 (2018).
31. J. D'Andrilli, W. T. Cooper, C. M. Foreman, A. G. Marshall, An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Communications in Mass Spectrometry* **29**, 2385-2401 (2015).
32. X. Wu *et al.*, Microbial Interactions With Dissolved Organic Matter Drive Carbon Dynamics and Community Succession. *Frontiers in Microbiology* **9** (2018).
33. G. G. Z. Silva, K. T. Green, B. E. Dutilh, R. A. Edwards, SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* **32**, 354-361 (2015).
34. S. M. D. Seaver *et al.*, The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Research* **49**, D575-D588 (2020).
35. Z. Jie *et al.*, The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications* **8**, 845 (2017).
36. G. Zeller *et al.*, Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10**, 766 (2014).
37. K. Forslund *et al.*, Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262-266 (2015).
38. J. E. Flannery *et al.*, Gut Feelings Begin in Childhood: the Gut Metagenome Correlates with Early Environment, Caregiving, and Behavior. *mBio* **11**, e02780-02719 (2020).
39. A. Barnes *et al.*, People Cities and Nature (PCaN) pilot metagenomics dataset.

40. S. Sun, B. D. Badgley, Changes in microbial functional genes within the soil metagenome during forest ecosystem restoration. *Soil Biology and Biochemistry* **135**, 163-172 (2019).
41. A. Bissett *et al.*, Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. *GigaScience* **5**, 21 (2016).
42. Y. Hu, Q. Zheng, S. Zhang, L. Noll, W. Wanek, Significant release and microbial utilization of amino sugars and d-amino acid enantiomers from microbial cell wall decomposition in soils. *Soil Biology and Biochemistry* **123**, 115-125 (2018).
43. M. Law, Dietary fat and adult diseases and the implications for childhood nutrition: an epidemiologic approach. *The American Journal of Clinical Nutrition* **72**, 1291s-1296s (2000).
44. L. G. Albenberg, G. D. Wu, Diet and the Intestinal Microbiome: Associations, Functions, and Implications for Health and Disease. *Gastroenterology* **146**, 1564-1572 (2014).
45. I. Burstyn *et al.*, Polycyclic Aromatic Hydrocarbons and Fatal Ischemic Heart Disease. *Epidemiology* **16**, 744-750 (2005).
46. M. A. Mallah *et al.*, Relationship Between Polycyclic Aromatic Hydrocarbons and Cardiovascular Diseases: A Systematic Review. *Frontiers in Public Health* **9** (2021).
47. J. Tao *et al.*, Lignin – An underutilized, renewable and valuable material for food industry. *Critical Reviews in Food Science and Nutrition* **60**, 2011-2033 (2020).
48. Y. Chen, Y. Liu, X. Li, J. Zhang, G. Li, Lignin Interacting with α-glucosidase and its Inhibitory Effect on the Enzymatic Activity. *Food Biophysics* **10**, 264-272 (2015).
49. B. M. Frier, G. Schernthaner, S. R. Heller, Hypoglycemia and Cardiovascular Risks. *Diabetes Care* **34**, S132-S137 (2011).
50. J. Xu *et al.*, Non-linear associations of serum and red blood cell folate with risk of cardiovascular and all-cause mortality in hypertensive adults. *Hypertension Research* **46**, 1504-1515 (2023).
51. E. J. Calabrese *et al.*, Biological stress response terminology: Integrating the concepts of adaptive response and preconditioning stress within a hormetic dose–response framework. *Toxicology and Applied Pharmacology* **222**, 122-128 (2007).
52. H. Essa *et al.*, Vitamin K2—a neglected player in cardiovascular health: a narrative review. *Open Heart* **8**, e001715 (2021).
53. J. W. Bellinge *et al.*, Vitamin K Intake and Atherosclerotic Cardiovascular Disease in the Danish Diet Cancer and Health Study. *Journal of the American Heart Association* **10**, e020551 (2021).
54. B. K. McFarlin, A. L. Henning, A. S. Venable, Oral Consumption of Vitamin K2 for 8 Weeks Associated With Increased Maximal Cardiac Output During Exercise. *Altern Ther Health Med* **23**, 26-32 (2017).
55. Y. Zheng *et al.*, Metabolites of Glutamate Metabolism Are Associated With Incident Cardiovascular Events in the PREDIMED PREvención con DIeta MEDiterránea (PREDIMED) Trial. *Journal of the American Heart Association* **5**, e003755 (2016).
56. W. Ma *et al.*, Dietary glutamine, glutamate and mortality: two large prospective studies in US men and women. *International Journal of Epidemiology* **47**, 311-320 (2017).
57. C. Loï, L. Cynober, Glutamate: A Safe Nutrient, Not Just a Simple Additive. *Annals of Nutrition and Metabolism* **78**, 133-146 (2022).
58. R. Liu *et al.*, Association between dietary protein intake and the risk of hypertension: a cross-sectional study from rural western China. *Hypertension Research* **36**, 972-979 (2013).
59. M. N. Alam, M. Almoyad, F. Huq, Polyphenols in Colorectal Cancer: Current State of Knowledge including Clinical Trials and Molecular Mechanism of Action. *BioMed Research International* **2018**, 4154185 (2018).
60. J. D. O. Mota, S. Guillou, F. Pierre, J.-M. Membré, Public health risk-benefit assessment of red meat in France: Current consumption and alternative scenarios. *Food and Chemical Toxicology* **149**, 111994 (2021).
61. K. Skarbek, M. J. Milewska, Biosynthetic and synthetic access to amino sugars. *Carbohydrate Research* **434**, 44-71 (2016).
62. J. Zhang, S. Zou, L. Fang, Metabolic reprogramming in colorectal cancer: regulatory networks and therapy. *Cell & Bioscience* **13**, 25 (2023).

63. S. Oliai Araghi *et al.*, Folic Acid and Vitamin B12 Supplementation and the Risk of Cancer: Long-term Follow-up of the B Vitamins for the Prevention of Osteoporotic Fractures (B-PROOF) Trial. *Cancer Epidemiology, Biomarkers & Prevention* **28**, 275-282 (2019).
64. E. R. Pearson, Type 2 diabetes: a multifaceted disease. *Diabetologia* **62**, 1107-1112 (2019).
65. H. Deng, R. Callender, G. E. Dale, A Vibrational Structure of 7,8-Dihydrobiopterin Bound to Dihydroneopterin Aldolase. *Journal of Biological Chemistry* **275**, 30139-30143 (2000).
66. K. Kusunoki *et al.*, Serum. Levels of Dihydroneopterin and Soluble Cytokine Receptors in Major Depression. *Pteridines* **10**, 24-26 (1999).
67. D. Cao, G. Pizzorno, Uridine phosophorylase: an important enzyme in pyrimidine metabolism and fluoropyrimidine activation. *Drugs of Today* **40**, 431-443 (2004).
68. Y.-H. S. Douglas G. Kondo, Tracy L. Hellem, Kristen K. Delmastro, Eun-Kee Jeong, Namkug Kim, Xianfeng Shi, and Perry F. Renshaw, Open-Label Uridine for Treatment of Depressed Adolescents with Bipolar Disorder. *Journal of Child and Adolescent Psychopharmacology* **21**, 171-175 (2011).
69. B. R. Hutto, Folate and cobalamin in psychiatric illness. *Comprehensive Psychiatry* **38**, 305-314 (1997).
70. Henning Tiemeier, M.D. , *et al.*, Vitamin B12, Folate, and Homocysteine in Depression: The Rotterdam Study. *American Journal of Psychiatry* **159**, 2099-2101 (2002).
71. J. Yang, S. C. Kalhan, R. W. Hanson, What is the metabolic role of phosphoenolpyruvate carboxykinase? *J Biol Chem* **284**, 27025-27029 (2009).
72. O. E. Owen, S. C. Kalhan, R. W. Hanson, The Key Role of Anaplerosis and Cataplerosis for Citric Acid Cycle Function. *Journal of Biological Chemistry* **277**, 30409-30412 (2002).
73. X. Gu *et al.*, Energy metabolism in major depressive disorder: Recent advances from omics technologies and imaging. *Biomedicine & Pharmacotherapy* **141**, 111869 (2021).
74. W. H. van der Putten, M. A. Bradford, E. Pernilla Brinkman, T. F. J. van de Voorde, G. F. Veen, Where, when and how plant–soil feedback matters in a changing world. *Functional Ecology* **30**, 1109-1121 (2016).
75. D. A. Wardle *et al.*, Ecological Linkages Between Aboveground and Belowground Biota. *Science* **304**, 1629-1633 (2004).
76. G. Janusz *et al.*, Lignin degradation: microorganisms, enzymes involved, genomes analysis and evolution. *FEMS Microbiology Reviews* **41**, 941-962 (2017).
77. J. C. Rodriguez-Ramos *et al.*, Changes in soil fungal community composition depend on functional group and forest disturbance type. *New Phytologist* **229**, 1105-1117 (2021).
78. S. C. Cunningham *et al.*, Reforestation with native mixed-species plantings in a temperate continental climate effectively sequesters and stabilizes carbon within decades. *Global Change Biology* **21**, 1552-1566 (2015).
79. B. N. Avera, B. D. Strahm, J. A. Burger, C. E. Zipper, Development of ecosystem structure and function on reforested surface-mined lands in the Central Appalachian Coal Basin of the United States. *New Forests* **46**, 683-702 (2015).
80. D. M. J. S. Bowman, B. P. Murphy, G. E. Burrows, M. D. Crisp, "Fire regimes and the evolution of the Australian biota". (CSIRO Publishing, Collingwood, Vic, 2012), pp. 27-47.
81. Q. Hou, G. Ufer, D. Bartels, Lipid signalling in plant responses to abiotic stress. *Plant, Cell & Environment* **39**, 1029-1048 (2016).
82. K. A. Hammer, C. F. Carson, T. V. Riley, Antimicrobial activity of essential oils and other plant extracts. *Journal of Applied Microbiology* **86**, 985-990 (1999).
83. M. Erb, D. J. Kliebenstein, Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. *Plant Physiology* **184**, 39-52 (2020).
84. M. Handsley-Davis *et al.*, Heritage-specific oral microbiota in Indigenous Australian dental calculus. *Evolution, Medicine, and Public Health* **10**, 352-362 (2022).
85. A. O. Wagner *et al.*, Sample preparation, preservation, and storage for volatile fatty acid quantification in biogas plants. *Engineering in Life Sciences* **17**, 132-139 (2017).
86. A. Rivas-Ubach *et al.*, Moving beyond the van Krevelen Diagram: A New Stoichiometric Approach for Compound Classification in Organisms. *Analytical Chemistry* **90**, 6152-6160 (2018).

87. Flinders_University (2021) DeepThought (HPC). Retrieved from https://doi.org/10.25957/FLINDERS.HPC.DEEPTHOUGHT.
88. S. Andrews (2018) FastQC - a quality control application for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
89. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
90. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368 (2021).