

NMR metabolome and stool consistency in IBS: a layer-wise multi-omics comparison

Yi Sui

gisui777@163.com

Chengdu No.7 High School, International Department <https://orcid.org/0009-0006-3612-9633>

Research Article

Keywords: irritable bowel syndrome, layer-wise multi-omics, Bristol stool consistency, short-chain fatty acids, subject-grouped cross-validation, label-blind transductive stability selection, held-out zero-shot validation, Mendelian randomization

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11496080/v1>

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

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

NMR metabolome and stool consistency in IBS: a layer-wise multi-omics comparison

Yi Sui

Chengdu No.7 High School, International Department

Email: ysisui777@163.com

Abstract

Background Irritable bowel syndrome (IBS) is markedly heterogeneous [1], and existing biological subtyping based on taxonomy (16S/metagenomic species) has limited predictive ability under subject-level validation. Using real clinical labels (IBS subtype, IBS-SSS severity, and Bristol stool consistency), this study performs a unified, subject-grouped cross-validated comparison across multiple omics layers and completes held-out zero-shot validation and label-blind transductive cross-cohort stability selection in an external cohort (zero contact with target-domain labels). The core comparison task is Bristol stool consistency regression; the three-class task serves only as an auxiliary control.

Methods We used the Cell 2020 longitudinal multi-omics cohort [2] (75 subjects, 464 clinically labeled stool samples: IBS-D 169, IBS-C 142, healthy 153; layer-specific availability differed: genus/species 457, KO/KEGG 459, NMR 445). The Bristol score constructed from stool diaries (441 samples, 67 subjects; IBS-D 4.92 > healthy 3.94 > IBS-C 2.93, direction validated) served as the core quantitative phenotype. Eight feature layers were compared: genus/species taxonomy, SHOGUN KEGG Orthology (KO) genes, KEGG pathways (375 aggregated), NMR metabolome (19), bile acids (12), tryptophan metabolites (12), and NMR+taxonomy combined (163 dimensions). All modeling used subject-grouped 5-fold cross-validation (StratifiedGroupKFold/GroupKFold) with nested tuning; Ridge/ElasticNet for regression (evaluated by OOF R^2 and Spearman ρ) and L2 logistic regression for classification. Cross-layer consistency was assessed by Spearman correlation between functional potential and metabolite pools (n=423). Cell 2020-trained models were frozen and blindly tested zero-shot on the Jacobs 2023 cohort [3] (368 samples $\times$ 601 metabolites). Finally, source-domain closed-loop stability selection was performed (π_{uni_cell} using only source-domain grouped bootstrap, B=300; platform-shift KS using only target-domain unlabeled features; two sensitivity levels, pre-screening nested and fully nested, with gates recomputed within resampling). After freezing the selector, a fully held-out zero-shot evaluation was performed on Jacobs (2000 $\times$ paired bootstrap) to define the cross-platform stable core (Iron-Core). Two-sample Mendelian randomization (MR) was used to test potential causal effects of 27 candidate genera on IBS risk between the MiBioGen microbiome GWAS [4] and the IBS GWAS (GCST90016564) [5] (IVW/weighted median/MR-Egger plus Cochran Q and leave-one-out sensitivity, with global BH-FDR correction).

Results ① The taxonomic layer showed limited signal for the IBS-D/C/healthy three-class task in the **full sample** (genus 0.506 ± 0.077 , 95% CI [0.421, 0.595], permutation $p=0.005$; species 0.520 ± 0.134 , 95% CI [0.434, 0.611]; chance=0.333), rising to 0.556/0.559 on the common complete subset and clustering with NMR (see ②); no reliable signal for IBS-SSS severity (genus balanced accuracy=0.500, 95% CI [0.420, 0.578], permutation $p=0.458$; only 2/70 features $q < 0.05$). ② For Bristol regression, the NMR metabolome had the highest point-estimate OOF R^2 : full-sample OOF $R^2=0.196$, $\rho=0.468$ (16/19 significant), and it remained highest on the common complete subset (0.193 vs 0.196). Its OOF R^2 point estimate exceeded those of taxonomy (genus 0.089/subset 0.037; species 0.043/subset -0.023), bile acids (0.053/0.073), tryptophan (0.023/0.015), and functional gene layers (KO 0.005/-0.039; KEGG 0.036/-0.008); the unfiltered NMR+genus combined model was 0.171/0.165, lower than NMR; after filtering at the same prevalence=0.10, the combined model was 0.204 (19 NMR+68 genera, n=424), not significantly different from NMR alone ($\Delta R^2=+0.006$, 95% CI -0.081-+0.095, $p=0.454$)—the earlier

“combined dilution” was mainly caused by unfiltered low-prevalence genera. The ordering of mid-tier layers (genus, species, bile acids, tryptophan) was sensitive to sample composition (genus and bile acids swapped on the subset) and does not constitute a strong ranking claim. In the auxiliary three-class task, full-sample NMR balanced accuracy=0.558±0.085, but on the common subset NMR (0.556) clustered with genus (0.556), species (0.559), and combined (0.565), indistinguishable from one another—layer ordering in the classification task is sensitive to sample composition and does not provide evidence that “NMR is optimal for classification.” Within-NMR Shapley decomposition showed that **the SCFA group's marginal contribution was $\phi=0.106$, 2.4 times the amino-acid group's $\phi=0.045$** (remove SCFA $\Delta R^2=0.076$, $p=0.002$; remove AA $\Delta R^2=0.016$, $p=0.085$, not significant)—upgrading “SCFA dominance” from a qualitative ordering to a reproducible decomposition conclusion. ③ The markers most strongly positively correlated with Bristol were acetate ($\rho=0.414$), butyrate (0.379), lysine (0.374), tyrosine (0.311), propionate (0.297), succinate (0.273), and branched-chain amino acids (0.19–0.23); the branched short-chain fatty acid 2-methylbutyrate was the only significant negative correlate (−0.185). ④ Within subtypes (exploratory): NMR retained signal within IBS-C ($\rho=0.288$, subject-cluster bootstrap 95% CI [+0.074, +0.483]; R^2 CI crossed zero); no robust signal was detected within IBS-D ($\rho=-0.263$, 95% CI [−0.573, +0.128] crossing zero, $P(\rho \geq 0)=0.083$; $R^2=-0.048$, 95% CI [−0.226, −0.001] on the negative side, indicating no predictive ability)—under the power constraint of $n \approx 150$, this is only an exploratory hypothesis and is insufficient to support unsupervised subgrouping within subtypes. ⑤ Cross-layer validation: functional potential and metabolite pools were generally weakly coupled, but **the butyrate-producing community genomic abundance marker bcd (K00248, butyryl-CoA dehydrogenase gene, non-rate-limiting enzyme) was significantly positively correlated with the fecal butyrate pool ($\rho=0.265$, $q=8 \times 10^{-6}$)**; BSH/bai showed no association with the bile acid pool—“potential $\neq$ activity.” ⑥ Zero-shot cross-cohort validation (Cell 2020 → Jacobs 2023): the 11 shared-metabolite model achieved Diarrhea vs Constipation+Mixed AUC=0.647 (D vs C sensitivity 0.676), $\rho=+0.278$, with directional replication in 11/11 (one-sided sign test $p=4.9 \times 10^{-4}$; features are not independent, so this p value is descriptive only); the random 11-feature baseline AUC=0.506 (95% CI 0.434–0.589, $p=0.005$), but a demographics-only (age/BMI/sex) baseline already reached AUC=0.593 (95% CI 0.509–0.675), and R11's increment over demographics was +0.053 (95% CI −0.045–+0.153, $p=0.148$), not significant; the full-feature model including unmeasurable SCFAs collapsed to AUC=0.558—feature transferability is the first constraint on cross-cohort generalization. ⑦ **Source-domain closed-loop stability selection (label-blind transductive: $\pi_{\text{uni_cell}} \geq 0.8$ + platform-shift $KS < 0.15$, zero contact with target-domain labels) identified a 5-item conditionally stable core (Iron-Core: lysine, tyrosine, uracil, glucose, leucine; all reproduced under pre-screening nesting); under fully nested gate recomputation, unconditional robustness concentrated on lysine (0.96) and tyrosine (0.79), while the other three (leucine 0.36, glucose 0.28, uracil 0.23) were conditionally stable and require external replication**; after freezing, zero-shot AUC on the fully held-out target cohort was 0.634 (95% CI 0.559–0.705), $\rho=+0.243$ —not significantly different from the 11-shared-feature set (R11, 0.647; paired $\Delta AUC=-0.013$, $p=0.805$), and significantly better than the full 16-feature set including imputed SCFAs (F16, 0.558; $\Delta AUC=+0.075$, $p=0.013$). ⑧ Two-sample MR (genus→IBS causal test independent of the metabolite layer; 313 pruned SNPs → 259 harmonized SNPs): under the current instrumental-variable power and approximate distance-pruning framework, **no robust potential causal effect was detected** (minimum FDR=0.835); the strongest nominal signals were Bifidobacterium

(OR=1.070 [1.004–1.141], $p=0.039$) and *Eubacterium hallii* group (OR=1.074, $p=0.062$), both with low heterogeneity (Cochran Q $p>0.7$), no evidence of horizontal pleiotropy (Egger intercept $p>0.85$), and stable leave-one-out estimates, but observational and causal directions did not converge; *Roseburia* showed a nominal protective trend (OR=0.930, $p=0.219$), opposite to its positive observational correlation ($\rho=+0.284$).

Conclusions Under subject-grouped cross-validation, **the layer with the highest point-estimate OOF R^2 for the Bristol continuous-trait regression task** was the metabolome (within NMR, dominated by the SCFA group: Shapley $\varphi=0.106$, 2.4 times $\varphi_{AA}=0.045$), not taxonomy or functional gene potential. The cross-cohort platform-robust core was data-driven and conditionally stable at 5 metabolites (lysine, tyrosine, uracil, glucose, leucine), achieving discrimination on the fully held-out target cohort that was not significantly different from the 11-shared-feature set with fewer features (0.634 vs 0.647), and significantly better than the 16-feature full model (0.558). The results support three mechanistic axes in terms of directional consistency, but with different levels of evidence: **SCFA→serotonin→intestinal motility** is supported mainly by within-NMR evidence (Shapley $\varphi=0.106$, consistent marker direction spectrum); the **proteolytic fermentation fingerprint** is supported by the directional consistency of BCAAs/aromatic amino acids; and the **bile acid–TGR5 secretomotor** axis is supported only by the directional positive correlation of the bile-acid layer Top6 with loose stools, lacking functional/causal validation (“potential $\neq$ activity”). In the single-layer comparison, the feature-layer ranking was direct metabolic measurement > taxonomy > functional potential—this ranking applies only to the single-layer comparison for the Bristol regression task and should not be extrapolated to subtype classification (in the three-class task, NMR and taxonomic layers were indistinguishable on the common subset and the ordering was not robust), nor to combined models (after the same filtering, the combined model was not significantly different from NMR). It also shows that predictive magnitude and cross-domain transferability are not equivalent—the cross-cohort stable core is not the strongest within-domain feature. Two-sample MR further indicates that, under the current instrumental-variable power and approximate LD-pruning framework, no robust potential causal evidence was detected between mainstream candidate genera and IBS risk (none significant after FDR correction), and causal interpretation of observational microbiome–phenotype associations requires larger samples, bidirectional designs, and stronger instruments.

Keywords: irritable bowel syndrome; layer-wise multi-omics; Bristol stool consistency; short-chain fatty acids; subject-grouped cross-validation; label-blind transductive stability selection; held-out zero-shot validation; Mendelian randomization

1 Introduction

1.1 Background

IBS is one of the most common disorders of gut–brain interaction [1], and its clinical subtypes (IBS-D, diarrhea-predominant / IBS-C, constipation-predominant) and symptom severity (IBS-SSS) are highly heterogeneous. Previous studies based on fecal microbiota have generally reported that IBS can be separated from healthy controls (AUC 0.7–0.9)

[1,3], but whether this “separability” reflects a generalizable biological mechanism or is dominated by batch effects, subject effects, and repeated sampling has long been debated. For “precision biological subtyping,” the core question is not “can we classify?” but **which omics layer carries the most phenotypic information that generalizes across subjects and cohorts.**

1.2 Methodological caveat: domain leakage from longitudinal repeated samples

Early analyses in this project used unsupervised DMM clusters [8] as pseudo-labels and obtained a high balanced accuracy=0.83; however, because the same subject contributed samples from multiple time points, random cross-validation placed samples from the same subject in both training and test sets, causing severe domain leakage. After switching to (i) real clinical labels and (ii) subject-grouped cross-validation, performance dropped to ≈ 0.51 . **This lesson shows that any individualized subtyping evaluation must partition validation sets by subject (domain)**—this is the methodological basis for the “domain generalization” objective of this project.

1.3 This study

We focused on Bristol stool consistency, a continuous phenotype with clear physiological meaning, performed a unified benchmark comparison of eight feature layers, further examined the coherence of the “functional potential–metabolite pool–phenotype” chain through cross-layer association analysis, and finally converged on transferable core features through external-cohort held-out zero-shot validation and label-blind transductive stability selection. Contributions:

1. Constructed a continuous Bristol target validated in the clinical direction (IBS-D>healthy>IBS-C), covering 441 samples/67 subjects;
2. Compared 8 omics layers (8 layers $\times$ 2 tasks) under a unified grouped cross-validation framework;
3. Found that in Bristol regression, the NMR metabolome had a higher point-estimate OOF R^2 than taxonomy and functional layers; combined modeling showed no gain when unfiltered, and after filtering at the same prevalence was not significantly different from NMR;
4. Through cross-layer association and within-subtype robustness tests, provided mechanistic interpretation and the methodological conclusion “potential $\neq$ activity”;

5. **Completed cross-cohort held-out zero-shot validation and label-blind transductive stability selection (source-domain closed-loop protocol)**, identifying 5 conditionally stable features (Iron-Core: lysine, tyrosine, uracil, glucose, leucine; under fully nested analysis, unconditional robustness concentrated on lysine/tyrosine)—not significantly different from the 11-shared-feature set in zero-shot discrimination, and significantly better than the 16-feature full model.

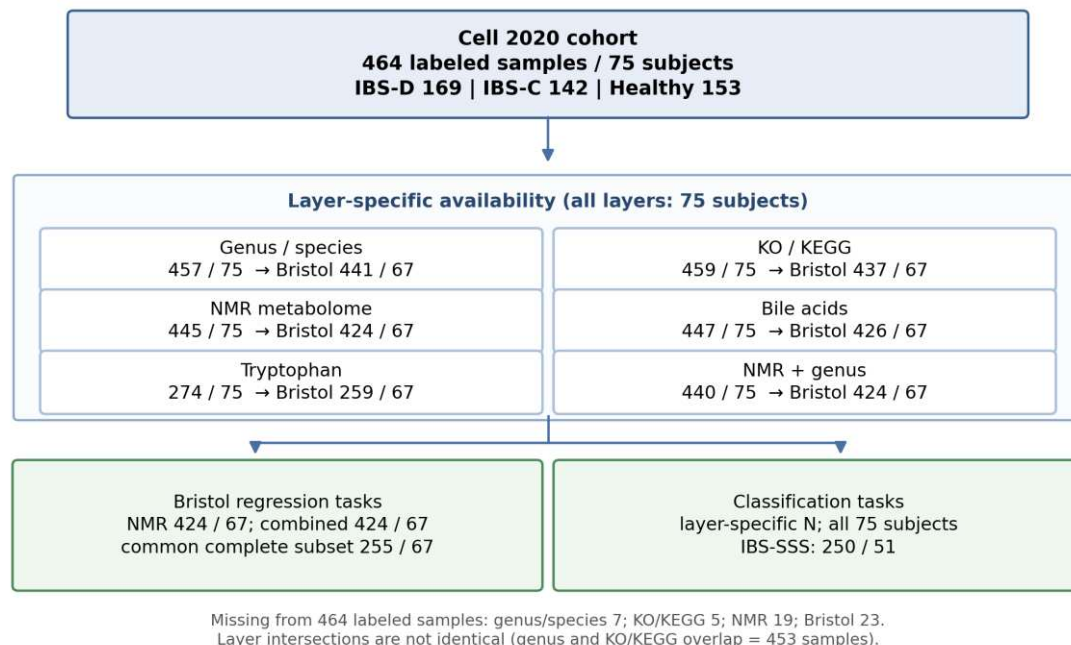
2 Materials and methods

2.1 Cohort and clinical labels

- **Source:** Cell 2020 longitudinal multi-omics IBS cohort [2] (Mars et al., Cell 2020; ENA PRJEB37924), with metagenomic abundances taken from GMrepo [11] processed data; sample–phenotype mapping was fully consistent with `disease_stage` (C↔IBS-C, D↔IBS-D, H↔healthy).
- **Scale:** 75 subjects and 464 clinically labeled stool samples (IBS-D 169, IBS-C 142, healthy 153). Layer-specific availability differed: genus/species 457 (D169/C137/H151), KO/KEGG 459 (D169/C140/H150), NMR 445 (D168/C133/H144), bile acids 447, tryptophan 274, and NMR+genus 440.
- **Severity:** IBS-SSS [7] available subset 250 samples/51 subjects, median 181.5 (binary threshold: median split).
- **Data hygiene:** the genus matrix removed the unclassified artifact (mean 21.4%, non-zero in 99.5% of samples); the species matrix removed the -1 artifact (mean 11.9%). All previous genus-level conclusions were rerun on this basis and remained unchanged.
- **External validation cohort:** Jacobs et al., Microbiome 2023 [3] (PRJNA812699), 368 samples × 601 untargeted LC-MS metabolites, BH subtypes (Constipation 89 / Diarrhea 84 / Mixed 37 / Unspecified 19), no SCFA measurements and no matched taxonomic matrix.
- **Sample flow:** 464 labeled samples → genus/species 457 (7 missing taxonomy), KO/KEGG 459 (5 missing the functional layer), and NMR 445 (19 missing NMR); Bristol was available for 441 (67 subjects) → $\text{NMR} \cap \text{Bristol}$ 424 (67 subjects). The 17 Bristol samples without NMR comprised w3 7, w7 4, and subject-mean fallback 6; another 21 NMR samples had no Bristol, and 2 samples had neither. KO/KEGG and genus/species are not the same subset (intersection 453); the three-class task does not

require Bristol, so all 459 KO/KEGG samples were usable for classification. Jacobs 368 → main evaluation 210 (C 89/D 84/M 37), D vs C sensitivity 173 (C 89/D 84). See Figure 15.

A Cell 2020 discovery cohort



B Jacobs 2023 external validation

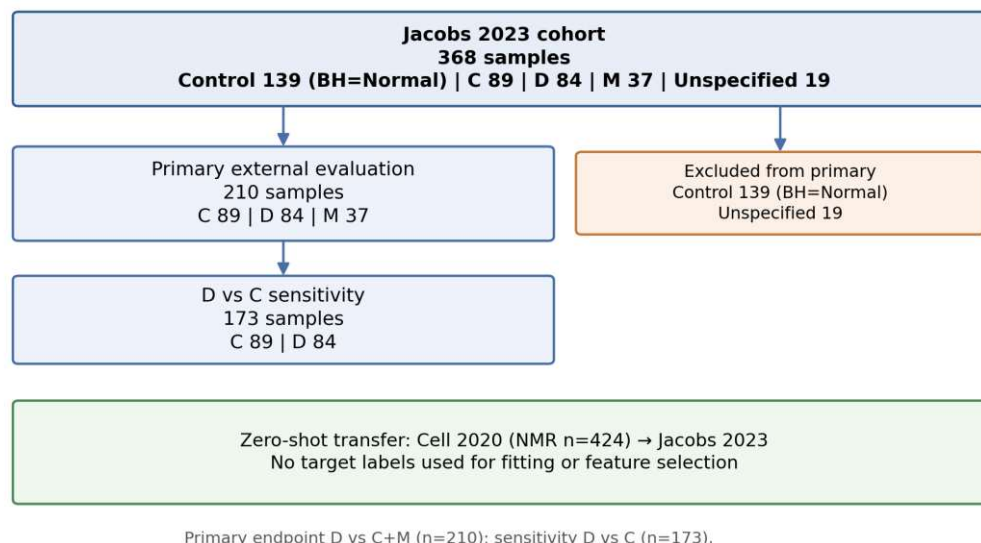


Figure 15. Sample flow and layer-specific availability. A: Cell 2020 discovery cohort (464 labeled samples; genus/species 457, KO/KEGG 459, NMR 445; Bristol 441/67, $NMR \cap Bristol$ 424/67; Bristol-missing 23 = 21 NMR samples without Bristol + 2 with neither); B: Jacobs 2023 external validation (368→210 main evaluation; 173 D vs C sensitivity).

2.2 Bristol target construction

Using stool diary collection dates (Excel serial format) for alignment, we constructed the Bristol score [6]: ± 3 -day window (110 samples) → ± 7 -day window (136 samples) →

subject-specific mean fallback (195 samples), totaling 441 samples/67 subjects (strictly counted by sample-source hierarchy; stratified sensitivity in §3.4/Figure 9). Direction validation: IBS-D 4.92 (SD 0.79) > healthy 3.94 > IBS-C 2.93 (SD 1.10), fully consistent with clinical priors.

2.3 Omics layers

Layer	Dimensions	Samples	Preprocessing
Genus taxonomy	144 (67–69 after filtering)	457	CLR (pseudo-count method consistent with the original pipeline)
Species taxonomy	328 (143–145 after filtering)	457	CLR
SHOGUN[9] KO	7332 (5298–5310 after $\text{prev} \geq 0.3$)	459	CLR
KEGG pathways	375 (aggregated; 327 after $\text{prev} \geq 0.5$)	459	KO assigned via <code>rest.kegg.jp/link/pathway</code> , summed, CLR
NMR metabolome	19	445	log1p
Bile acids	12	447	log1p (30.5% missing, min/2 imputation)
Tryptophan metabolites	12	274	log1p ('ND' treated as missing)
NMR+genus combined	163 (unfiltered) / 87 (same filtering, sensitivity)	440 / 424	pre-transform concatenation (transform=none); sensitivity filtered the genus layer at prevalence=0.10 (68 genera+19 NMR)

Functional layer data came from Mendeley repository [24] (SHOGUN KO table, NMR, bile acids, tryptophan, cytokines, and other processed matrices).

2.4 Modeling and validation

- **Regression (Bristol):** Ridge/ElasticNet; outer **5-fold subject-grouped** (GroupKFold), inner tuning ($\alpha \in \{0.1, \dots, 10^4\}$); metrics: OOF R^2 , OOF Spearman ρ ; univariate Spearman+BH-FDR [13].
- **Classification (IBS-D/C/healthy):** L2 logistic regression (C tuning); **StratifiedGroupKFold (5 folds, by subject)**; metrics: balanced accuracy, macro-F1; univariate tests+BH-FDR.
- **Classification uncertainty:** for each layer's OOF predictions, computed subject-cluster $2000 \times$ bootstrap 95% CI, subject-level label permutation p (NMR/genus/species/bile acids/tryptophan 200 permutations; KEGG 100; KO 20, with C fixed at 0.01 to control high-dimensional permutation cost), and 3×3 confusion matrices; the permutation p resolution floor is $1/(B+1)$ (≈ 0.005 for 200, ≈ 0.010 for 100, ≈ 0.048 for 20), and values

at the floor should not be interpreted as exact p values; IBS-SSS was genus-only with the same 200-permutation suite.

- **Within-subtype:** repeated Bristol regression separately within IBS-D (n=154) and IBS-C (n=128).
- **Bristol source sensitivity:** recomputed the main model stratified by target source ($\pm 3d$ / $\pm 7d$ / measured union / mean fallback) to test whether conclusions were driven by fallback samples (§3.4, Figure 9).
- **Cluster-level bootstrap:** within IBS-D and IBS-C, ρ and R^2 were used to construct 95% CIs with subject-cluster 2000 $\times$ bootstrap (§3.5, Figure 10).
- **External baselines:** trained on Cell 2020 and evaluated held-out on Jacobs 2023, comparing R11, 200 random 11-feature sets, and age/BMI/sex demographic models; the primary AUC endpoint was D vs C+M, with D vs C as sensitivity; paired bootstrap 2000 $\times$ tested Δ AUC.
- **Cross-layer association:** 11 functional features (butyrate synthesis gene set, bcd, adhE, ko00650 butyrate metabolism, ko00280 BCAA degradation, ko00350 tyrosine metabolism, ko00310 lysine degradation, ko00380 tryptophan metabolism, ko00120/ko00121 bile acid synthesis, BSH K01442) were Spearman-correlated with 16 significant NMR markers + 6 bile acid markers + Bristol (n=423), BH-FDR (denominator 253 tests).
- **Fair Comparison Protocol:** cross-layer comparisons were always run under a unified supervised protocol (same preprocessing $\rightarrow$ same GroupKFold $\rightarrow$ same Ridge/Logistic tuning, output OOF R^2 /balanced accuracy; all §3.3 numbers come from this protocol); unsupervised structural metrics (silhouette, Calinski-Harabasz, Davies-Bouldin, Gap, ARI) were used only for cluster-structure diagnostics and **must not be compared across scales with supervised predictive metrics**; any cross-layer comparison citing unsupervised scripts for genus/species matrices must be recomputed through the unified fair comparison protocol with default alignment parameters (prevalence=0.10, CLR, no additional denoising) to obtain OOF R^2 before entering the main text.
- **Software:** Python 3.13.13, NumPy 2.4.6, pandas 3.0.3, SciPy 1.17.1, scikit-learn 1.8.0, Matplotlib 3.10.9; metagenomic annotation MetaPhlAn 3.0.14 [10] (GMrepo [11] processing pipeline).

Notation: ρ (rho) denotes the Spearman rank correlation coefficient, φ (phi) denotes the Shapley contribution, and p denotes the P value.

2.5 Source-domain closed-loop stability selection and held-out zero-shot evaluation (label-blind transductive)

Protocol overview (engineering hard isolation): the cross-cohort step was split into three unidirectional data flows: selector (reads only Cell features/labels and Jacobs **unlabeled** X, outputs a frozen selector) → held-out evaluation (reads only the frozen selector and Jacobs X; BH labels introduced only at the end of evaluation) → two-level nested sensitivity (pre-screening nested and fully nested). Target-domain labels had zero contact with any selection step (label-blind transductive feature selection).

Selection gates (prespecified): ① source-domain univariate stability $\pi_{\text{uni_cell}} \geq 0.8$ (Cell 2020 grouped bootstrap, B=300, 50% subject sampling without replacement; frequency of Spearman $p < 0.05$ with direction consistent with the full sample; π_{multi} is the Lasso non-zero selection frequency with fixed α within each grouped bootstrap resample, a secondary metric); ② platform shift KS < 0.15 (two-sample KS statistic between self-standardized log1p distributions of Cell and Jacobs; only X used, no labels. Note that domain shift is a **composite shift**—platform (NMR vs LC-MS), feature availability (Jacobs has no SCFA/no matched taxonomy), and sample processing/study protocol (longitudinal multiple time points vs cross-sectional, collection/storage/extraction differences) superimposed; KS can only provide an unlabeled diagnosis of composite magnitude, not decompose it into a single source, and “platform robustness” should be understood as robustness to composite shift); ③ the two groups were jointly measurable under the prespecified names. Features passing all three gates constituted the Iron-Core; the selector used GroupKFold(5) within Cell to select α and fit StandardScaler+Ridge, and the scaling parameters and coefficients were frozen and exported together with the feature set. The old protocol's iron-core determination using target-domain labels ($\pi_{\text{uni_jacobs}}$) was deprecated (Appendix S1).

Held-out zero-shot evaluation: the frozen selector was applied directly to 368 Jacobs samples (main evaluation subset n=210: C 89 / M 37 / D 84); metrics were AUC for Diarrhea vs C+M (same definition as previous analyses) and BH ordinal Spearman; sample-level paired bootstrap (B=2000) constructed 95% CIs for each model's AUC and the ΔAUC distribution, with p taken as the proportion of $\Delta\text{AUC} \leq 0$; control models R11 (11 shared features) and F16 (16 features including imputed SCFAs) were both frozen from Cell 2020, without any target-domain tuning.

Two-level nested sensitivity: ① pre-screening nested: the full-data BH $q < 0.05$ pre-screening was moved to the inner layer—within each grouped bootstrap resample, univariate screening was rerun on all 19 NMR features ($q < 0.05$ and consistent direction), with gate values ($\pi_{\text{uni_cell}}$, KS, measurability) fixed, outputting the nested core (§3.9); ② fully nested: gates were also recomputed within resampling—outer subject-level full bootstrap ($B_{\text{out}}=100$), inner re-estimation of $\pi_{\text{uni_cell}}$ using the original π protocol (50% subject subsampling $\times B_{\text{in}}=100$), KS recomputed on the Cell subset of that resample, and a composite half-sample variant (50% $\times$ 50%) provided a conservative lower bound. Prespecified Plan A/B triggering criteria: ① closed-loop core $\cap \{\text{Uracil, Glucose}\} \geq 1$ (observed 2, met); ② held-out zero-shot AUC $\geq R11$ and paired $p < 0.10$ (observed $0.634 < 0.647$ but no significant difference, not met); ③ nested core non-empty (met)—narrative predefined: Plan A (if ② met) = “the 5-item Iron-Core outperforms R11 on the held-out cohort”; Plan B (② not met, actually used) = “the 5 items are not significantly different from R11, maintaining discrimination with fewer features.” The results use the Plan B narrative: source-domain stability shrinkage repairs F16 migration poisoning. This study was not prospectively registered; the gates, endpoints, and Plan A/B were fixed before cross-cohort evaluation and recorded in a versioned analysis log. The text therefore uses “prespecified” rather than “pre-registered.”

2.6 Two-sample Mendelian randomization (genus $\rightarrow$ IBS risk)

- **Exposure (instrumental variables):** MiBioGen microbiome meta-analysis [4] (Kurilshikov et al., Nat Genet 2021; 16–24 cohorts) full hit table MBG.allHits.p1e4.txt (122,110 rows / 211 taxa). We selected 27 candidate genera (covering study markers and GTDB homologous genera), with instrument threshold $p < 10^{-5}$: 1,755 SNPs (F statistic 100% > 10); because no individual-level LD panel was available, ± 1 Mb distance pruning was used to approximate LD removal (retaining the lowest-p SNP within each window), yielding 313 independent SNPs.
- **Outcome:** harmonized summary statistics from IBS GWAS GCST90016564 [5] (Eijbouts et al. 2021; 486,601 European samples, 53,400 cases / 433,201 controls), matched by `hm_rsid` (305/313); allele harmonization included strand flipping (complementary pairs), and palindromic SNPs were conservatively removed; finally 259 SNPs entered estimation (4–16 per genus).

- **Estimation and sensitivity:** primary analysis IVW [23] (random effects, delta-method SE); weighted median [22] (bootstrap 2000) and MR-Egger [21] (intercept test for horizontal pleiotropy) as sensitivity analyses; Cochran Q test for heterogeneity, leave-one-out robustness; BH-FDR on IVW p values for 27 genera. Separately, observational Spearman ρ between each genus and Bristol in this cohort/external cohort was summarized as a directional control.
- **STROBE-MR reporting:** reported data sources, instrument selection, harmonization, primary analysis, sensitivity, and multiple correction according to the STROBE-MR [12] checklist; complete 27-genus results and item mapping in Appendix S2.
- Data sources, result tables, and analysis workflow in §5.

3 Results

3.1 Clinical phenotype

The Bristol target aligned completely with subtypes (IBS-D 4.92 > healthy 3.94 > IBS-C 2.93), forming a credible continuous phenotype axis used as the unified target for all subsequent regression tasks.

3.2 Taxonomic layer: weak subtype signal, severity indistinguishable from chance

Task	Features	Result
IBS-D/C/healthy classification (genus)	69	balanced acc = 0.506±0.077 ; 95% CI [0.421, 0.595]; permutation p=0.005; macro-F1=0.508 (52/69 significant)
IBS-D/C/healthy classification (species)	145	balanced acc = 0.520±0.134 ; 95% CI [0.434, 0.611]; permutation p=0.005; macro-F1=0.519 (103/145 significant)
IBS-SSS binary (genus)	70	balanced acc = 0.500 ; 95% CI [0.420, 0.578]; permutation p=0.458; AUC=0.552; only 2/70 q<0.05 (Dorea, Romboutsia)

The taxonomic layer showed some signal relative to the chance baseline (0.333), but the magnitude was limited; the severity task was indistinguishable from chance (permutation p=0.458). Permutation resolution should be noted: with 200 permutations the p floor is 0.005, and the genus/species p=0.005 is at that floor; the KO p=0.095 is based on 20 permutations (floor $\approx$ 0.048) and is exploratory only. Eight-layer OOF performance, CIs, permutation p values, and confusion matrices are shown in Appendix S3.

3.3 Cross-layer comparison: Bristol regression is dominated by the metabolome (Figure 4)

Feature layer	Dimensions	Bristol OOF R ²	OOF ρ	Significant/total	Subtype balanced acc
NMR (19)	19	0.196	0.468	16/19	0.558±0.085
Genus (144)	67	0.089	0.309	22/67	0.506±0.077
Species (328)	143	0.043	0.261	54/143	0.520±0.134
Bile acids (12)	12	0.053	0.276	9/12	0.482±0.034
Tryptophan (12)	12	0.023	0.259	6/12	0.486±0.109
KO (7332)	5298	0.005	0.188	2349/5298	0.384±0.087
KEGG pathways (375)	327	0.036	0.235	274/327	0.436±0.088
NMR+genus combined (163)	163	0.171	0.438	62/163	0.503±0.099
NMR+genus combined (same filter, 87)	87	0.204	0.470	—	—

Key points: (i) In Bristol regression, the metabolome (NMR), with only 19 features, achieved a higher point-estimate OOF R² than 144–328-dimensional taxonomy and ~5.3k-dimensional functional features (no cross-layer paired test was performed); (ii) the unfiltered combined model (19 NMR+144 genera, 163 dimensions) had OOF R²=0.171, lower than NMR; after filtering at the same prevalence=0.10 as the single layers (19 NMR+68 genera, 87 dimensions, n=424), it was **0.204**, not significantly different from NMR alone at 0.196 ($\Delta R^2=+0.006$, 95% CI -0.081 – $+0.095$, p=0.454; paired subject bootstrap 2000×)—the earlier “combined dilution” was mainly caused by unfiltered low-prevalence genera, rather than taxonomy itself lacking marginal value; (iii) in the regression task, “functional potential” layers such as KO/KEGG pathways performed worst under grouped CV, suggesting an increasing “phenotypic distance” across the three data types: functional potential < taxonomic structure < metabolite pool (for the classification task see the common-subset paragraph; cross-layer ordering was not robust).

Note: In the table, the cross-layer comparison numbers for genus/species layers come from fair-comparison-protocol recomputation (prevalence=0.10, CLR; §2.4), not from the raw outputs of the unsupervised subtyping scripts; the NMR+genus combined row is the unfiltered 163-dimensional primary analysis, and the 87-dimensional sensitivity after the same filtering is described in (ii) above. The denominator of significant/total ranges from 19 (NMR) to 5298 (KO); comparing the number of significant features across layers is meaningless (BH-FDR power also decreases as the number of tests increases); cross-layer comparisons should focus on OOF R² and balanced accuracy. The 87-dimensional same-filter combined model was re-run for regression only (n=424); its classification metrics were not recomputed in that sensitivity, hence the em dashes.

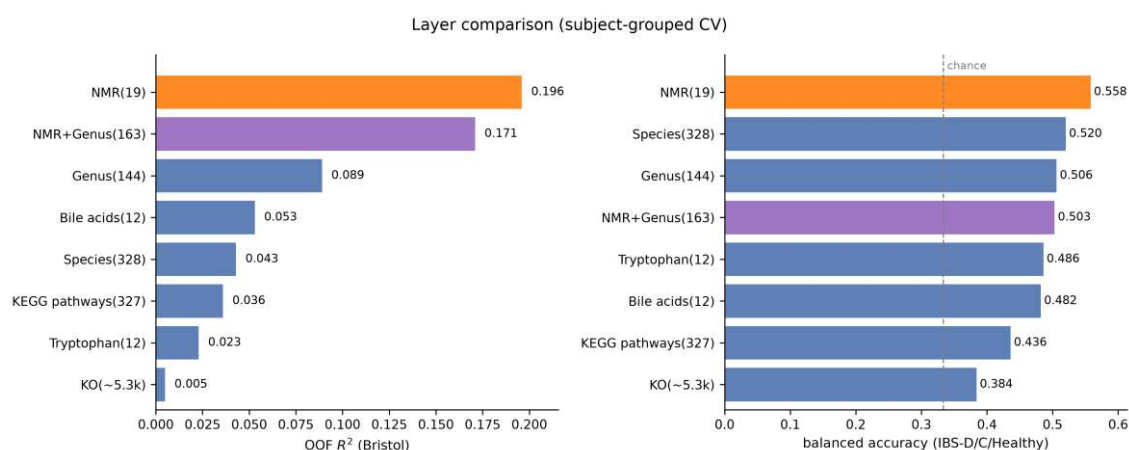


Figure 4. Cross-layer comparison of eight omics layers. Left: Bristol regression OOF R²; right: three-class balanced accuracy (dashed line=chance baseline 1/3); orange highlights the NMR layer. The 87-dimensional same-filter NMR+genus sensitivity (OOF R²=0.204) is not plotted; see the §3.3 table.

Common complete-subset sensitivity (Figure 14): All eight layers were rerun on the common complete sample subset under the same fair protocol (268 samples/75 subjects; regression n=255/67 subjects). **The first and last positions in regression and the rank of the combined model were unchanged:** NMR 0.193 (full sample 0.196) remained highest, NMR+genus combined 0.165 (0.171) second, and KO −0.039 (0.005) last (KEGG −0.008 and species −0.023 were also ≤0); mid-tier layers (bile acids 0.073, genus 0.037, tryptophan 0.015) were much lower in absolute magnitude than NMR, but the relative order of genus and bile acids swapped (genus 0.089→0.037, bile acids 0.053→0.073)—this magnitude difference is sensitive to sample composition, and no ranking claim is made in the text. In the classification task, the top four layers on the subset clustered together (combined 0.565, species 0.559, genus 0.556, NMR 0.556; fold SD 0.05–0.12) and were indistinguishable from one another; NMR balanced accuracy was essentially unchanged (0.558→0.556), while KO (0.399) and KEGG (0.425) remained lowest. **Conclusion: in the regression task, “NMR is the layer with the highest point-estimate OOF R² and functional potential is weakest” does not depend on sample-size differences across layers; the cross-layer ordering in the classification task is not robust;** small differences in mid-tier ordering and classification positions should not be overinterpreted.

Explanation for regression stability and classification reversal: In the classification task, the taxonomic and combined models rose markedly on the common subset (genus 0.506→0.556, species 0.520→0.559, combined 0.503→0.565), whereas NMR was essentially unchanged (0.558→0.556); in the regression task, the opposite occurred: NMR was essentially unchanged (0.196→0.193), whereas taxonomic layers decreased (genus 0.089→0.037, species 0.043→−0.023). This indicates that **the NMR advantage in the**

regression task is more robust, and the full-sample cross-layer differences in the classification task are more affected by sample composition/repeated sampling; the classification task should not be used for strong ranking claims. In the subset regression, KO/KEGG/species showed negative OOF R^2 ($-0.039/-0.008/-0.023$), meaning that high-dimensional layers had held-out predictions worse than the mean on the $n=255$ subset, reflecting overfitting and instability of functional/high-dimensional taxonomic layers under limited sample size; this is consistent with the weak signals of KO 0.005 and KEGG 0.036 in the full sample and does not change the qualitative conclusion that “functional potential is weakest.” The regression subset was limited by Bristol availability (67 subjects), whereas the classification subset included all 75 subjects; the two subsets were constructed separately and are not the same samples, so cross-task comparisons require caution.

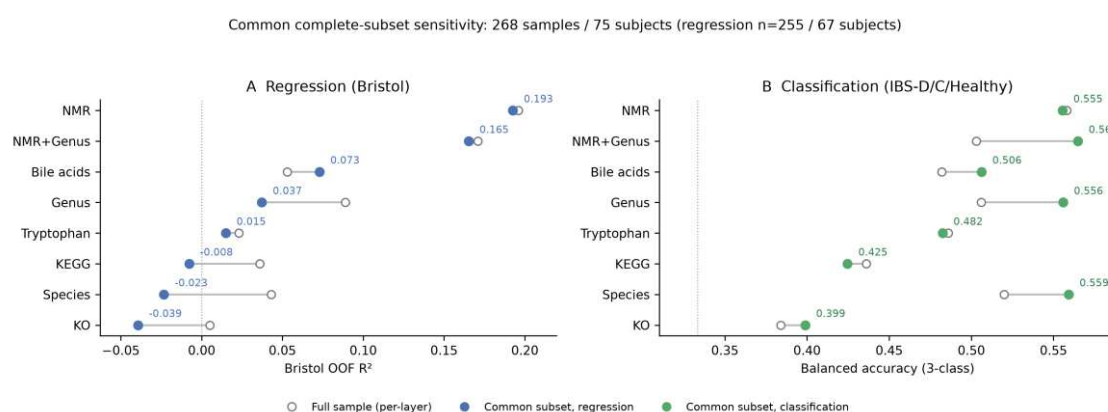


Figure 14. Common complete-subset sensitivity of the eight layers. A: Bristol regression OOF R^2 (open circles=full sample for each layer, n per layer in §2.3; filled circles=common complete subset $n=255/67$ subjects, limited by Bristol availability); B: three-class balanced accuracy (subset $n=268/75$ subjects, all subjects). On the subset, NMR regression remained highest (0.193 vs full-sample 0.196), and KO was lowest (-0.039); the top four classification layers clustered (0.556–0.565) and were indistinguishable. In subset regression, KO/KEGG/species showed negative OOF R^2 ($-0.039/-0.008/-0.023$), indicating that high-dimensional layers had held-out predictions worse than the mean at $n=255$: KO/KEGG were lowest, which does not change the qualitative conclusion that “functional layers are weakest”; the negative R^2 of the species layer reflects the instability of high-dimensional taxonomy under limited sample size, and the species layer should not be classified as a functional layer on this basis; the regression and classification subsets were constructed separately and are not the same samples.

3.4 NMR markers: SCFA and amino acid profiles (Figures 1 and 2)

Sixteen of 19 NMR metabolites were significantly correlated with Bristol (BH-FDR), with a highly consistent direction spectrum (closer to loose stools, higher concentrations; except branched-chain SCFAs):

Marker	ρ	q	Interpretation
Acetate	+0.414	<1e-16	SCFA
Butyrate	+0.379	<1e-14	SCFA/FFAR2-3 pathway core
Lysine	+0.374	<1e-13	Proteolytic fermentation

Marker	ρ	q	Interpretation
Tyrosine	+0.311	<1e-12	Aromatic AA/tyramine precursor
Propionate	+0.297	<1e-11	SCFA
Succinate	+0.273	<1e-9	Facultative anaerobic fermentation marker
Uracil	+0.236	<1e-7	Nucleic acid degradation
Glucose	+0.234	<1e-7	Carbon-source delivery/absorption imbalance
Leucine	+0.232	<1e-7	BCAA
Valine	+0.198	1e-4	BCAA
Isoleucine	+0.187	2e-4	BCAA
Alanine, hypoxanthine, lactate (3 items)	0.12–0.16	<0.05	—
Isovalerate	−0.122	0.014	Branched SCFA
2-Methylbutyrate	−0.185	2e-4	Branched SCFA (only significant strong negative signal)

Note: The row “alanine, hypoxanthine, lactate (3 items)” is a merged row (univariate ρ 0.123–0.162, all $q < 0.05$): alanine $\rho = 0.162$, $q = 0.0012$, assigned to the amino acid (AA) group in §3.4b (source of 6 significant features among 7 in that group); hypoxanthine $\rho = 0.156$, $q = 0.0017$ and lactate $\rho = 0.123$, $q = 0.014$ were assigned to the “other” group.

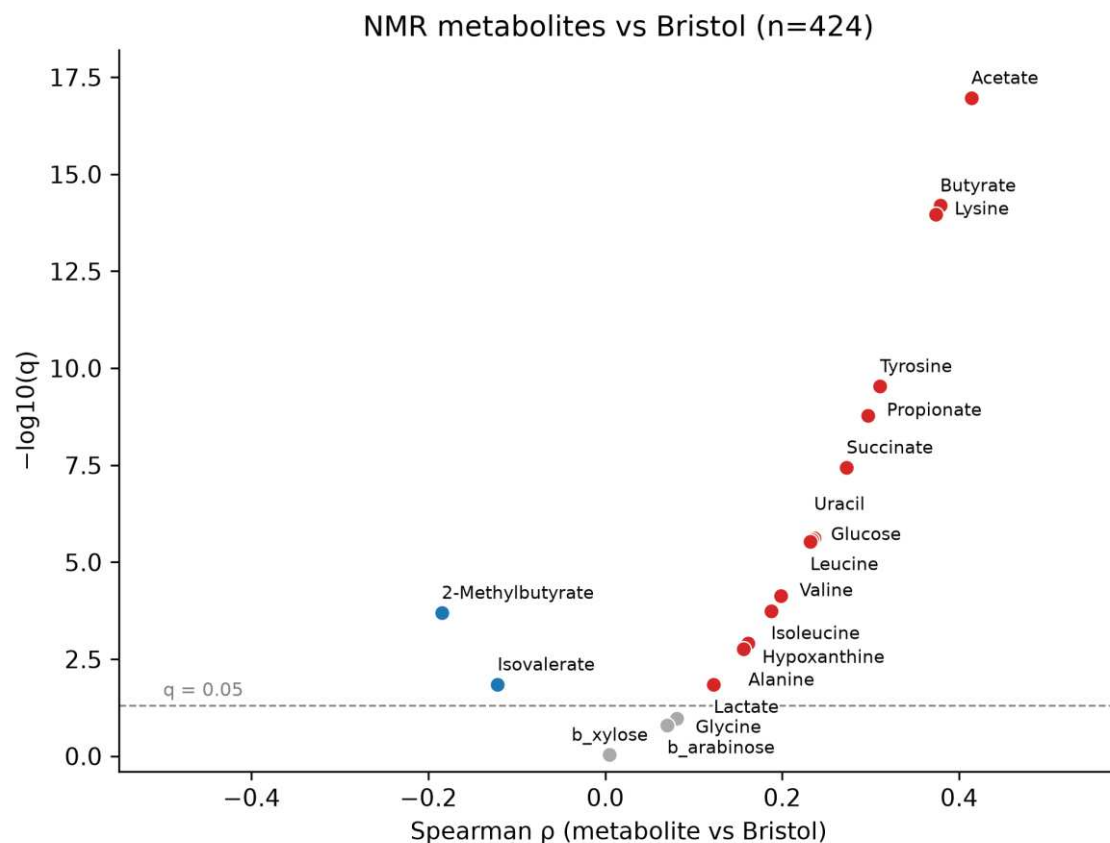


Figure 1. Volcano plot of NMR metabolite associations with Bristol. x-axis: Spearman ρ ; y-axis: $-\log_{10}(q)$; 16/19 metabolites had $q < 0.05$; dashed line: $q = 0.05$.

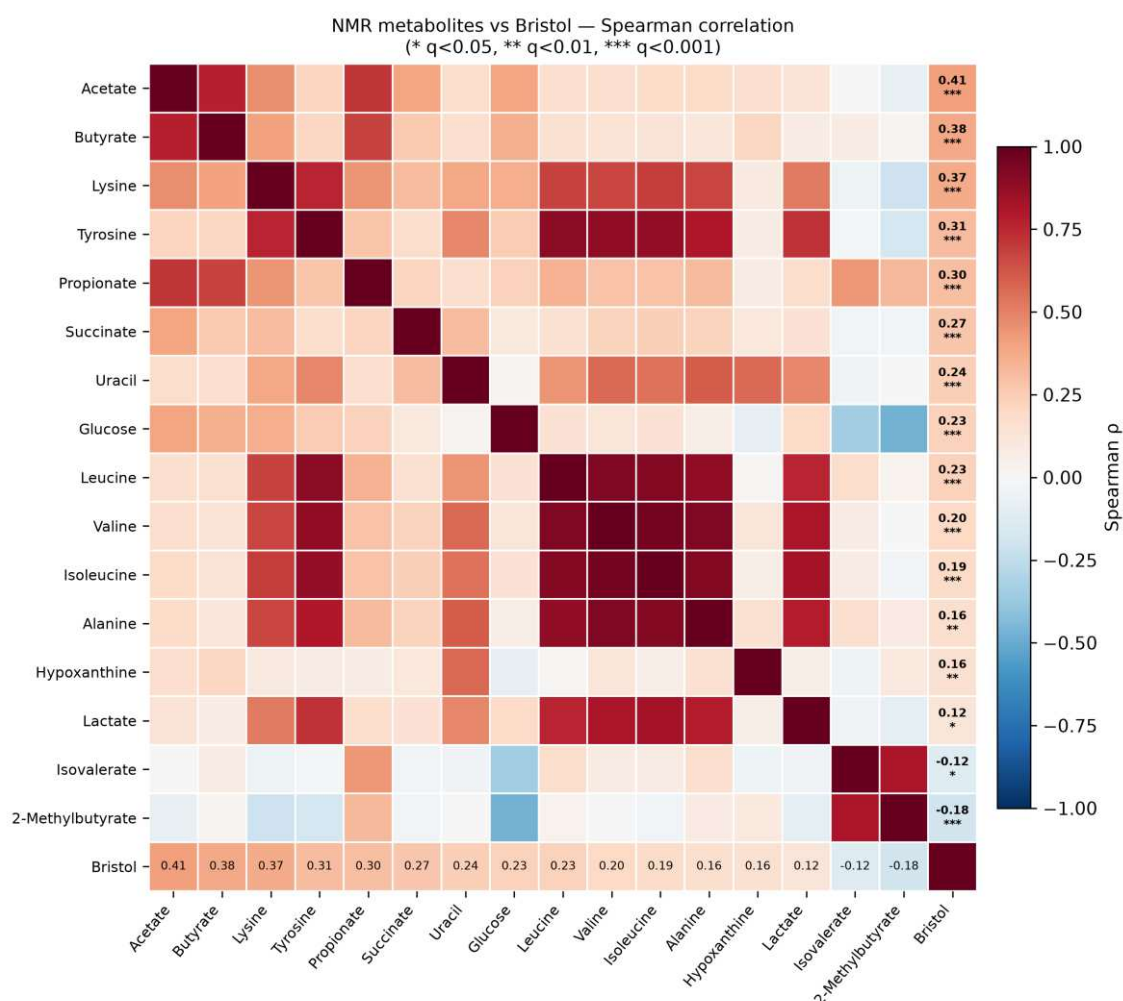


Figure 2. Spearman correlation heatmap of the 16 significant NMR markers × Bristol. Markers are ordered by correlation strength with Bristol; the Bristol column is annotated with ρ and significance ($q<0.05$, $q<0.01$, $q<0.001$).

Bile acid layer Top6 (all significant by BH-FDR): glycochenodeoxycholic acid GCDCA (+0.310), chenodeoxycholic acid CDCA (+0.309), deoxycholic acid-sulfate DCA-S (+0.305), cholic acid CA (+0.294), glycodeoxycholic acid GDCA (+0.272), and glycocholic acid GCA (+0.228)—**primary bile acids and glycine-conjugated bile acids were positively correlated with loose stools**, consistent with the continuous spectrum of bile acid diarrhea (BAD).

Bristol source sensitivity (Figure 9): The Bristol target was stratified by source (NMR complete analysis set $n=424$; raw source counts $441 = w3\ 110 + w7\ 136 + \text{fallback}\ 195$, three mutually exclusive layers with no overlap; the 17-case difference was entirely NMR missing—7 $w3 + 4\ w7 + 6\ \text{fallback}$, dispersed across 15/67 subjects (13 with 1 case each, 2 with 2 cases each), with no subject clustering, and all 67 subjects were retained in the NMR analysis set). Subject-cluster 2000× bootstrap (error bars=95% CI): measured

union $n=235$ ($R^2=0.261$ [0.112, 0.350], $\rho=0.500$ [0.352, 0.621]); $\pm 3d$ only $n=103$ (0.105 [−0.166, 0.253]); $\pm 7d$ only $n=132$ (0.147 [−0.094, 0.277]); fallback $n=189$ ($R^2=0.158$ [0.028, 0.247], $\rho=0.451$ [0.278, 0.585]). **Interpretation:** (i) rank correlations were consistent across all source layers and CI lower bounds all excluded zero (0.354–0.500), so the main conclusion was not driven by any single source; (ii) single-window (w3/w7) R^2 point estimates were lower than fallback, but CIs overlapped substantially (lower bounds negative), and the differences were within the noise range, so no strength claim is made; (iii) fallback targets were subject-mean smoothed (lower effective degrees of freedom), and their apparent R^2 may be inflated by smoothing, requiring conservative interpretation.

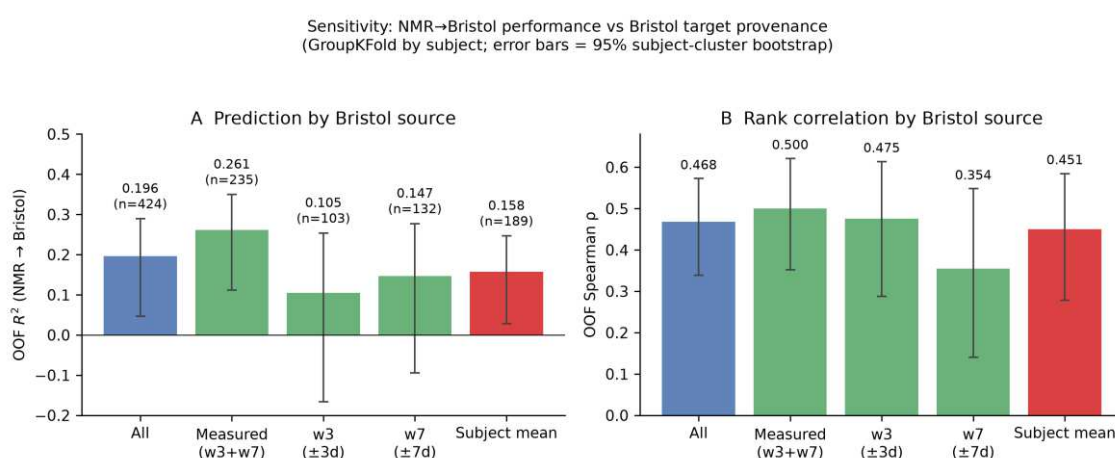


Figure 9. Bristol target source sensitivity. Stratified by source (raw counts 441 = 110/136/195, mutually exclusive; NMR complete analysis set 424 = 103/132/189, with the 17-case difference due to NMR missing): all ($R^2=0.196$ [0.047, 0.289], $\rho=0.468$ [0.339, 0.573]), measured union ($n=235$, 0.261 [0.112, 0.350]/0.500 [0.352, 0.621]), $\pm 3d$ only ($n=103$, 0.105 [−0.166, 0.253]), $\pm 7d$ only ($n=132$, 0.147 [−0.094, 0.277]), and mean fallback ($n=189$, $R^2=0.158$ [0.028, 0.247]/ $\rho=0.451$ [0.278, 0.585]); error bars=subject-cluster 2000× bootstrap 95% CI. Rank correlations were consistent across all source layers (CI lower bounds excluded zero), single-window R^2 point differences were within noise (overlapping CIs), and the main conclusion was not driven by fallback samples.

3.4b Internal decomposition of the NMR layer: Shapley contributions of SCFAs and amino acids (Figure 13)

Grouping and models: The 19 items were divided into three groups by chemical class—**SCFA 5 items** (acetate, propionate, butyrate, isovalerate, 2-methylbutyrate), **amino acids 7 items** (alanine, glycine, leucine, valine, isoleucine, lysine, tyrosine), and **other 7 items** (lactate, succinate, arabinose b_arabinose, xylose b_xylose, glucose, uracil, hypoxanthine; grouping determined by the actual data column names). Grouping was by chemical class, not by univariate significance (the AA group includes glycine, which was not univariately significant, to preserve class completeness); within-group feature-level decomposition (e.g., within BCAA, lysine/tyrosine relative) is planned next. Six OOF

models were run under the unified primary protocol (GroupKFold(5); Figure 13A): all 19 items **0.196**, SCFA-only 0.161, AA-only 0.102, other-only 0.045, no-SCFA 0.120, and no-AA 0.180.

Shapley decomposition (other group as constant baseline): $\varphi_{\text{SCFA}} = 0.5[(R^2(\text{SCFA} \cup \text{other}) - R^2(\text{other})) + (R^2(\text{all } 19) - R^2(\text{AA} \cup \text{other}))] = \mathbf{0.106}$; φ_{AA} symmetric = **0.045**. The sum of the two marginal contributions (0.151) equals the joint gain $R^2(\text{all } 19) - R^2(\text{other})$; the residual-defined interaction term is 0.000—in a two-group Shapley decomposition, the sum of the two components' Shapley values always equals the joint gain, and the interaction term is identically zero by algebraic definition, which does not constitute evidence of “no interaction”; separating the interaction term requires a full three-group (SCFA/AA/other) Shapley decomposition, planned next.

ΔR^2 subject-level permutation (1,000 each, shared permutation sequence): removing the SCFA group decreased the full-model R^2 by $\Delta=0.076$ ($p=0.002$); removing the amino acid group produced $\Delta=0.016$ ($p=0.085$, not significant) (Figure 13C). **“SCFA dominance” is upgraded from a qualitative ordering to a reproducible decomposition conclusion**; the association of the amino acid profile with Bristol (§3.4) and its marginal predictive increment relative to the SCFA group must be interpreted separately—the independent increment of BCAAs/lysine/tyrosine at this sample size did not reach permutation significance.

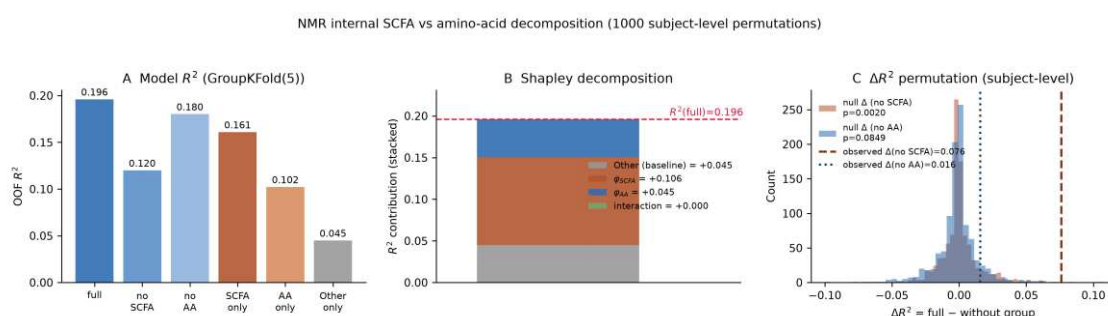


Figure 13. Between-group Shapley decomposition and ΔR^2 permutation in NMR. A: OOF R^2 of six models (full 0.196 / no-SCFA 0.120 / no-AA 0.180 / SCFA-only 0.161 / AA-only 0.102 / other-only 0.045); B: Shapley stacking (baseline 0.045 + φ_{SCFA} 0.106 + φ_{AA} 0.045, total= $R^2(\text{full})=0.196$; in the two-group decomposition the interaction term is identically zero by residual definition (algebraic identity), and this is not independent evidence of “no interaction”; see §3.4b); C: subject-level permutation null distribution of ΔR^2 (remove SCFA $\Delta=0.076$, $p=0.002$; remove AA $\Delta=0.016$, $p=0.085$). $M_{\text{scfa_only}}$ (0.161) includes the shared covariance of SCFAs with the AA/other groups and is not equal to the independent contribution of SCFAs (the independent contribution is given by Shapley decomposition, $\varphi_{\text{SCFA}}=0.106$); $M_{\text{no_scfa}}$ (0.120) is the performance of the AA+other groups after removing SCFAs, and its difference from $M_{\text{scfa_only}}$ reflects covariance structure and should not be directly interpreted by subtraction.

3.5 Predictive performance and within-subtype robustness (Figures 3 and 3b)

- The NMR layer contained 445 samples/75 subjects; Bristol regression used 424 samples/67 subjects. Overall OOF predictions: $R^2=0.196$, $\rho=0.468$ ($n=424$; including IBS-D 154, IBS-C 128, healthy 142).
- **Within IBS-C:** $R^2=0.061$, $\rho=0.288$ (5/19 significant; subject-cluster 2000× bootstrap: ρ 95% CI [+0.074, +0.483], $P(\rho \geq 0)=0.996$, R^2 95% CI [−0.180, +0.178])—the metabolite–Bristol coupling still held within a single subtype (rank correlation robust, magnitude CI wide).
- **Within IBS-D (exploratory):** $R^2=-0.048$, $\rho=-0.263$. Subject-cluster 2000× bootstrap ($n=154$, 23 subjects): ρ 95% CI [−0.573, +0.128] crossing zero ($P(\rho \geq 0)=0.083$), and R^2 95% CI [−0.226, −0.001] entirely negative. Thus **ρ was indistinguishable from zero, whereas the R^2 point estimate was negative with a CI on the negative side:** the model had no predictive ability within this subtype (rather than a robust negative correlation). It is therefore registered only as an exploratory hypothesis-generating observation, and no unsupervised subgrouping was performed within IBS-D (DMM/hierarchical clustering tends to pseudo-converge in small samples) (Figure 10).

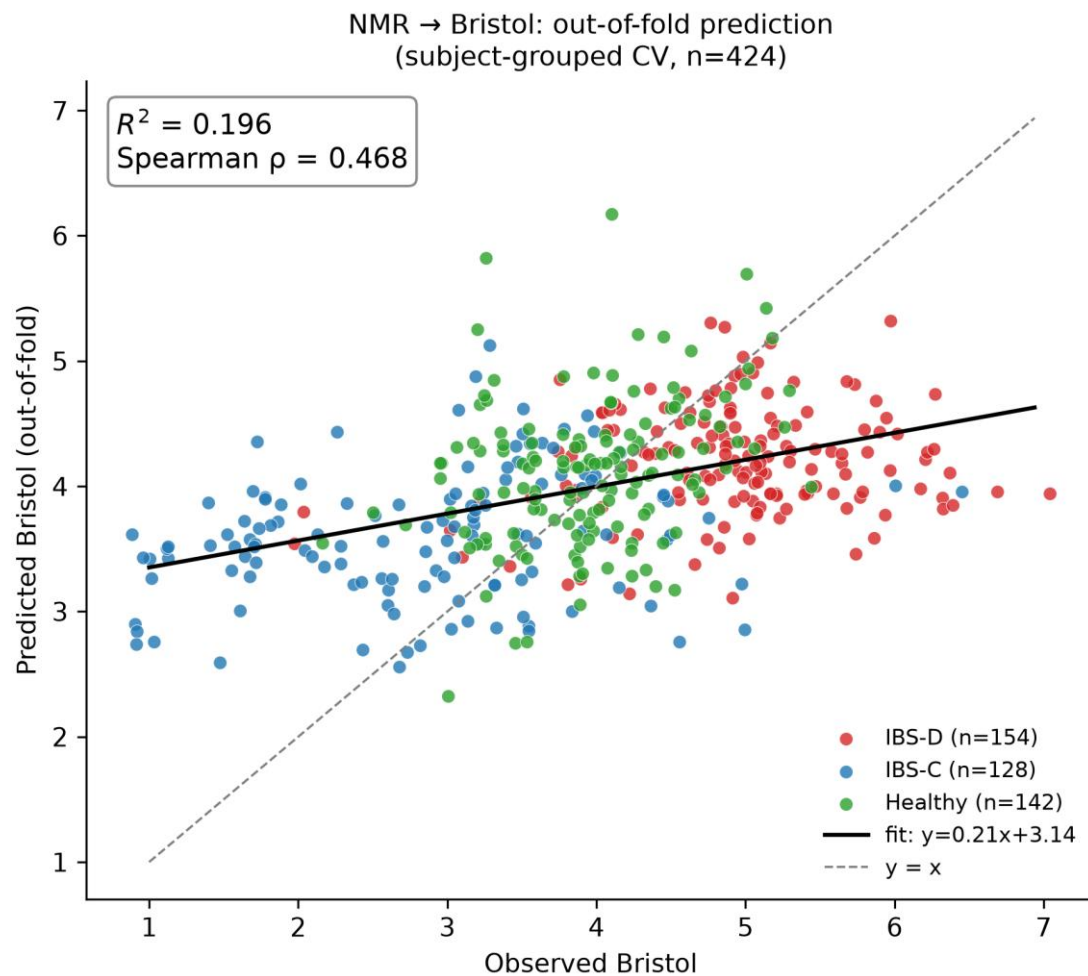


Figure 3. Subject-grouped cross-validated OOF predictions of the NMR metabolome for Bristol. $R^2=0.196$, $\rho=0.468$ (n=424); points colored by subtype (IBS-D red, IBS-C blue, healthy green); black line=linear fit, gray dashed line= $y=x$.

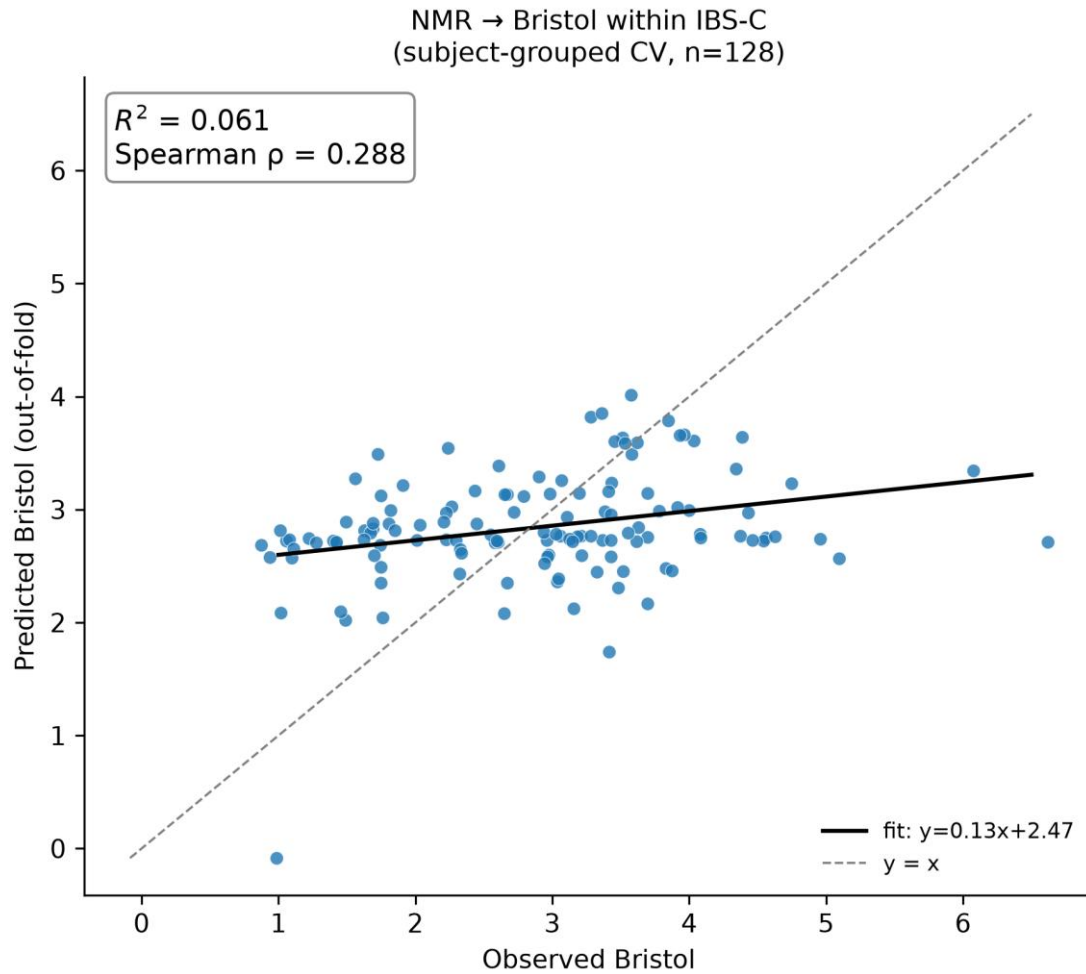


Figure 3b. Within-IBS-C robustness. n=128, $R^2=0.061$, $\rho=0.288$ (5/19 significant).

This asymmetry is only an exploratory observation (exploratory hypothesis-generating): it suggests (rather than proves) that the signal does not all come from subtype labels; under the power constraint of $n \approx 150$, sampling fluctuation cannot be excluded, and validation in larger within-subtype cohorts is needed (§4.7; bootstrap confidence intervals in Figure 10).

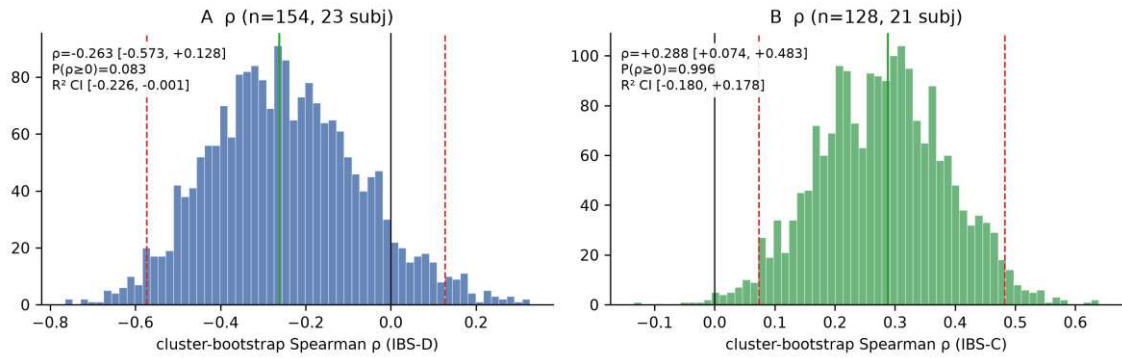


Figure 10. Subject-cluster bootstrap of within-subtype ρ (both subtypes). A: IBS-D (n=154, 23 subjects)— ρ point estimate -0.263 , 95% CI $[-0.573, +0.128]$ crossing zero ($P(\rho \geq 0)=0.083$), R^2 point estimate -0.048 , 95% CI $[-0.226, -0.001]$; B: IBS-C (n=128, 21 subjects)— ρ point estimate $+0.288$, 95% CI $[+0.074, +0.483]$ ($P(\rho \geq 0)=0.996$), R^2 CI $[-0.180, +0.178]$. The signals in both subtypes are only exploratory hypotheses.

3.6 Cross-layer validation: selective weak coupling between functional potential and metabolite pools (Figure 5)

Among 253 associations across 11 functional features $\times$ 23 targets (16 NMR + 6 BA + Bristol, n=423):

Functional feature	Target	ρ	q
bcd (K00248)	Butyrate	+0.265	8×10^{-6}
bcd	Hypoxanthine	+0.209	9×10^{-4}
bcd	Lysine / succinate	+0.189	0.003
bcd	Acetate	+0.182	0.005
bcd	Uracil	+0.183	0.005
bcd	Bristol	+0.175	0.007
adhE (K04072)	DCA-sulfate	-0.231	2×10^{-4}
adhE	Bristol	-0.203	0.001
ko00380 tryptophan metabolism	Bristol	-0.215	7×10^{-4}
ko00280 BCAA degradation	Bristol	-0.181	0.005
Butyrate synthesis gene set (7 KOs)	Lysine	+0.161	0.017
Butyrate synthesis gene set	2-Methylbutyrate	-0.144	0.050

Key observations: (i) the single core gene marker **bcd** (K00248, a genomic abundance marker of butyrate-producing communities, not the rate-limiting enzyme of butyrate synthesis) showed consistent weak-to-moderate coupling with the major fecal SCFA pools and Bristol; (ii) however, **aggregation diluted the signal**—neither the 7-gene sum nor the ko00650 pathway aggregation showed significant associations, possibly reflecting isoenzyme heterogeneity, annotation noise, and the absence of expression regulation; (iii) bile acid-related functional genes (BSH K01442, bai-related ko00121)

showed no significant association with the bile acid pool—a typical case of “**functional potential** $\neq$ **enzyme activity** $\neq$ **metabolite pool**”; (iv) ko00280/ko00380 (BCAA/tryptophan degradation potential) were negatively correlated with Bristol and positively correlated with fecal free AAs, forming a “low potential, high product” combination, suggesting that when transit accelerates, the host absorption window shortens, and the fate of AAs in the lumen is determined by competition between absorption and microbial activity, rather than by degradation gene abundance alone.

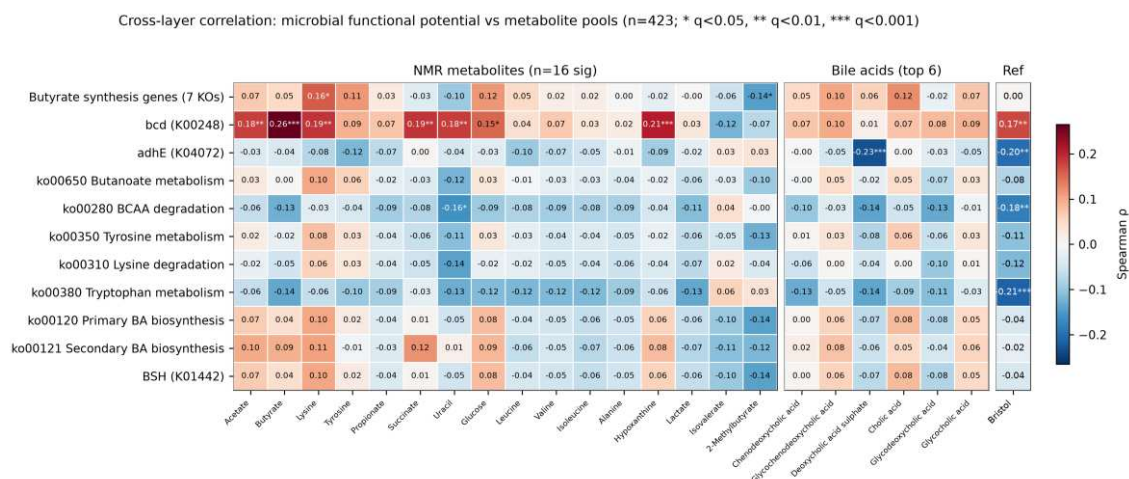


Figure 5. Cross-layer association heatmap: functional potential vs metabolite pools. 11 functional features $\times$ (16 NMR + 6 bile acids + Bristol), n=423; global BH-FDR correction (253 tests). Core findings: bcd $\leftrightarrow$ butyrate ($\rho=0.265$), adhE $\leftrightarrow$ DCA-sulfate (-0.231), ko00380/ko00280 $\leftrightarrow$ Bristol ($-0.215/-0.181$).

3.7 External reference: Microbiome 2023 [3] (Jacobs et al., PRJNA812699)

In the independent cohort (368 samples $\times$ 601 metabolites; IBS-C 89 / IBS-D 84 / mixed 37), the metabolome distinguished IBS from controls with AUC $\approx$ 0.85; however, classification by BH subtype (D/C/M) was weak (balanced acc $\approx$ 0.42–0.52), consistent with the cohort's own report of “limited metabolic differences between subtypes, with elevated bile acids and polyamines in the D subtype,” and also consistent with the positive bile acid–Bristol direction in our cohort. Both cohorts point to the same conclusion: **the metabolic layer is good at capturing “disease vs healthy” and continuous severity axes, rather than discrete subtype labels.**

3.8 Zero-shot cross-cohort domain generalization: Cell 2020 $\rightarrow$ Jacobs 2023 (Figure 6)

Feature alignment: The 16 Bristol-significant markers from Cell 2020 were matched by prespecified names to the 601 LC-MS panel of Jacobs 2023. The Jacobs platform (untargeted LC-MS) does not cover SCFAs; the transferable intersection was **11/16**

(Lysine, Tyrosine, Succinate, Uracil, Glucose, Leucine, Valine, Isoleucine, Alanine, Hypoxanthine, Lactate); the 5 missing items (acetate/butyrate/propionate/isovalerate/2-methylbutyrate) were neutrally imputed at the training mean in z-space.

Zero-shot blind test: The model was trained only on Cell 2020 (subject-grouped CV for α); weights, standardization statistics, and intercept were frozen and then applied directly to 368 Jacobs samples—no training, tuning, or feature selection was performed in the target cohort. BH groups (Constipation 89 / Mixed 37 / Diarrhea 84) were used as a proxy endpoint for intestinal motility (prespecified hypothesis: predicted score $D > M > C$).

Results: The shared 11-feature model (R11) performed well zero-shot—predicted score Diarrhea vs Constipation+Mixed AUC=**0.647** (MWU $p=3.2 \times 10^{-5}$; D vs C sensitivity AUC=0.676), with BH ordinal Spearman $\rho=+0.278$ ($p=4.4 \times 10^{-5}$), Kruskal-Wallis $p=3.1 \times 10^{-4}$; consistent across batches (Batch One AUC=0.679; Batch Two=0.586); no significant confounding by age/BMI/sex ($p>0.05$). Within the training domain, this submodel had OOF $R^2=0.115$, $\rho=0.370$.

Failure-mode attribution (key): The full 16-feature model including 5 imputed SCFAs (F16) collapsed to AUC=0.558 ($\rho=+0.136$, $p=0.049$) on the same task—when the SCFAs with the largest variance and strongest weights in the training domain were unmeasurable in the target domain, the effective transferable variance of the full model was absorbed by them; **signal recovered after shrinking to the measurable intersection**. This pattern complements the conclusions of DANN domain-adversarial experiments: the core constraint on cross-center generalization is not only distribution shift but also **feature transferability**—non-transferable high-weight features systematically poison transfer.

Directional replication: 11/11 shared features had the same correlation direction with BH ordinal in Jacobs as with Bristol in Cell 2020 (one-sided binomial sign test $p=4.9 \times 10^{-4}$).

Random and demographic baselines (new): Under the same Cell→Jacobs zero-shot pipeline, the mean AUC of 200 random 11-feature models was 0.506 (95% quantile 0.434–0.589), and R11's 0.647 was above this null distribution ($p=0.005$). The demographics-only model (age/BMI/sex) had AUC=0.593 (95% CI 0.509–0.675); R11's Δ AUC relative to demographics was +0.053 (95% CI -0.045 – $+0.153$, $p=0.148$), not significant. Therefore, R11's discriminative ability cannot be attributed to random features, but it is also

insufficient to prove independence from demographic variables; the metabolome increment should be read as directional, exploratory evidence.

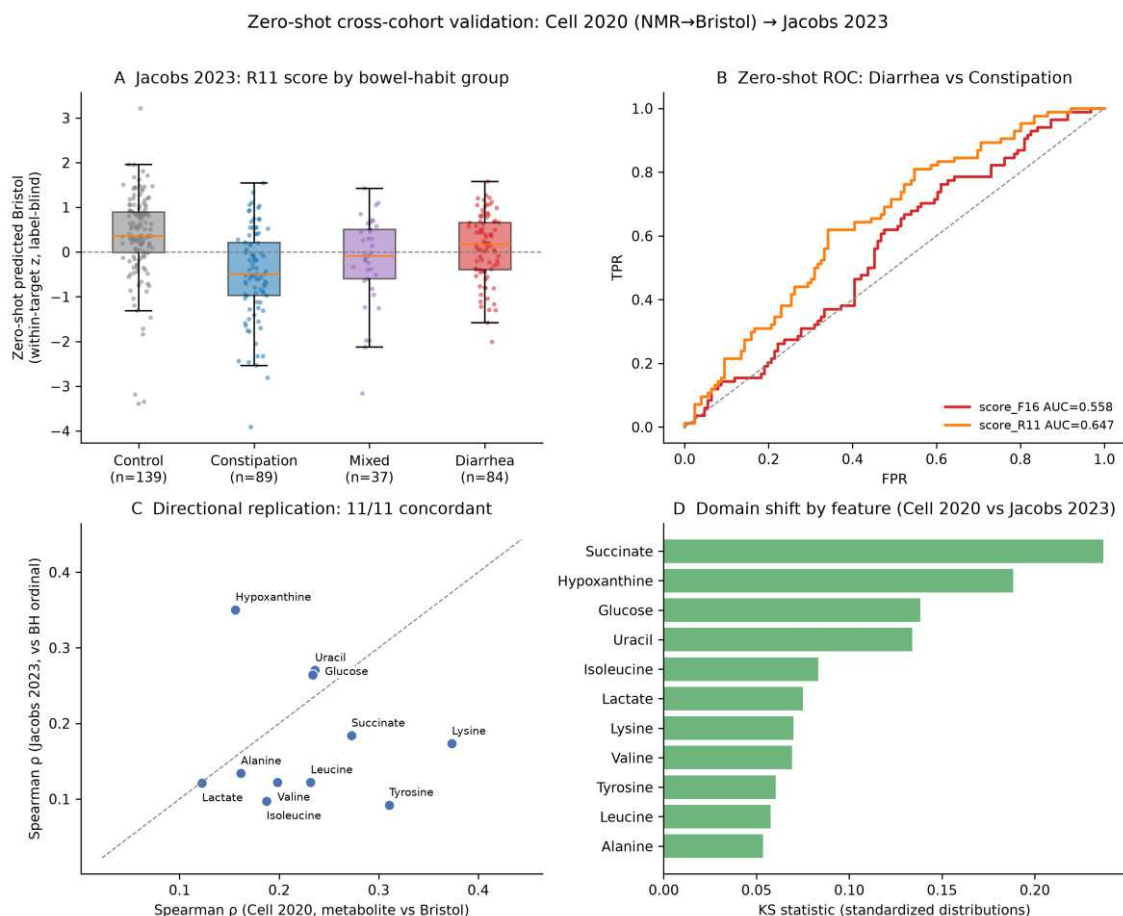


Figure 6. Zero-shot cross-cohort domain generalization (Cell 2020 → Jacobs 2023). A: R11 predicted scores by Jacobs motility group (z-standardized within the target domain, label-blind; healthy controls (Control; BH=Normal; n=139) included as reference, controls not included in ROC)—a clear C<M<D monotonic trend; B: Diarrhea vs Constipation+Mixed zero-shot ROC (R11 AUC=0.647; D vs C sensitivity 0.676; F16 collapsed to 0.558 because of imputed SCFAs); C: directional replication of the 11 shared features (11/11 consistent); D: domain-shift diagnostics (self-standardized KS statistics).

3.9 Source-domain closed-loop stability selection and held-out zero-shot evaluation: the Iron-Core platform-robust core (Figure 7)

Cross-cohort stability selection was upgraded to a **source-domain closed-loop protocol** (label-blind transductive feature selection; criteria in §2.5): inclusion decisions used only ① Cell 2020 grouped bootstrap stability $\pi_{\text{uni_cell}}$; ② the platform-shift KS of target-domain **unlabeled** X; ③ measurability in both groups—target-domain labels had zero contact during selection. The selector (features, scaling parameters, Ridge coefficients) was frozen before zero-shot evaluation on the fully held-out target cohort (2000× paired bootstrap).

Source-domain stability ($\pi_{\text{uni_cell}}$): In Cell 2020, acetate/butyrate/lysine had $\pi_{\text{uni_cell}}$ =1.0, tyrosine 0.990, uracil 0.910, glucose 0.883, and leucine 0.883; succinate 0.957, valine 0.753, 2-methylbutyrate 0.700, isoleucine 0.680, alanine 0.597, hypoxanthine 0.570, and lactate 0.403—those below 0.8 or borderline were excluded by the stability gate.

Platform shift (KS, unlabeled diagnostic): Unmeasurable SCFAs (acetate/butyrate/propionate/isovalerate/2-methylbutyrate) had no KS definition and were treated as non-transferable; among measurable items, succinate (0.237) and hypoxanthine (0.188) exceeded the 0.15 threshold and were excluded; the 5 core features all had KS < 0.15 (0.058–0.138).

Iron-Core determination (three gates): Features simultaneously satisfying ① $\pi_{\text{uni_cell}} \geq 0.8$, ② KS < 0.15, and ③ measurability in both groups totaled **5 items: lysine, tyrosine, uracil, glucose, and leucine**.

Feature	$\pi_{\text{uni_cell}}$	$\pi_{\text{multi_cell}}$	KS	Decision
Lysine	1.000	0.840	0.070	Iron ✓
Tyrosine	0.990	0.990	0.060	Iron ✓
Uracil	0.910	0.950	0.134	Iron ✓
Glucose	0.883	0.963	0.138	Iron ✓
Leucine	0.883	0.913	0.058	Iron ✓
Succinate	0.957	—	0.237	Platform-shift exclusion
Hypoxanthine	0.570	—	0.188	Platform-shift + source-stability exclusion
Valine	0.753	—	0.069	Source-stability exclusion
Isoleucine	0.680	—	0.083	Source-stability exclusion
Lactate	0.403	—	0.075	Source-stability exclusion
Acetate/butyrate/propionate/isovalerate/2-methylbutyrate	—	—	Unmeasurable	Non-jointly-measurable exclusion

Held-out zero-shot evaluation: The frozen 5-item model (within-domain OOF R^2 =0.072; 0.080 under the unified primary protocol, see the permutation paragraph below) was applied directly to Jacobs (D vs C+M): AUC=**0.634** (95% CI 0.559–0.705), BH ordinal ρ =**+0.243**; **not significantly different** from R11 (0.647, CI 0.572–0.716; paired Δ AUC=−0.013, 95% CI −0.044–+0.017, p =0.805), and significantly better than F16 (0.558, CI 0.478–0.630; Δ AUC=+0.075, CI +0.011–+0.140, p =0.013). **Two-level nested sensitivity:** under pre-screening nesting (B=300, gates fixed), the 5-item selection survival rates were 0.997/0.977/0.867/0.857/0.803, and the core was fully reproduced; under fully nested analysis (gates recomputed within resampling, outer full bootstrap B_{out}=100 ×

inner π protocol $B_{in}=100$), the per-feature all-gates-off selection frequencies were lysine 0.96, tyrosine 0.79, leucine 0.36, glucose 0.28, and uracil 0.23—the screening layer (0.94–1.00) and KS gate (amino acids 1.00; uracil 0.78, glucose 0.63) were highly stable, and the bottleneck was estimation noise of the π gate for intermediate values (0.883–0.910, near the 0.80 gate boundary); the conservative composite half-sample variant (outer 50% subject subsampling, inner another 50% of the outer, simulating a “sample size halved” composite conservative lower bound) was overall lower (lysine 0.75, tyrosine 0.58, no feature reached 0.80), consistent with the power loss from halving the sample. The frozen core should be read as a transferable set “conditional on source-domain gate estimation”: conditional stability fully holds, unconditional robustness concentrates on lysine/tyrosine, and the other three await external replication. That is: **source-domain stability shrinkage (16→5) repairs F16 migration poisoning and maintains discrimination not significantly different from R11 with fewer features—the number of platform-robust features is determined by the data.**

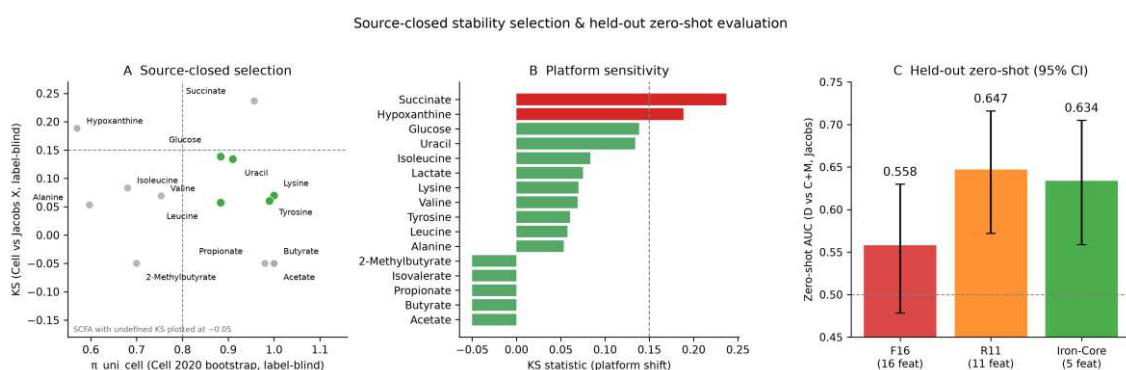


Figure 7. Source-domain closed-loop stability selection and held-out zero-shot evaluation (Iron-Core). A: source-domain stability $\times$ platform-shift scatter ($x=\pi_{uni_cell}$, $y=KS$; neither uses target-domain labels; green points=5 core items; dashed lines=0.8/0.15 thresholds); B: platform-shift diagnostics (self-standardized distribution KS; red=above the 0.15 threshold: Succinate 0.237, Hypoxanthine 0.188); C: held-out zero-shot AUC (D vs C+M; error bars=2000x paired bootstrap 95% CI)—Iron-Core (5 items, 0.634 [0.559, 0.705]) was not significantly different from R11 (11 items, 0.647 [0.572, 0.716]; $\Delta AUC=-0.013$, $p=0.805$) and was significantly better than F16 (16 items, 0.558 [0.478, 0.630]; $\Delta AUC=+0.075$, $p=0.013$).

Permutation significance of the within-domain fit (Figure 12): A subject-level permutation test (1,000 permutations, unified primary protocol) was performed on the within-domain OOF fits: full 19 items observed $R^2=0.196$, $p=0.001$; Iron-Core 5 items within-domain $R^2=0.080$, $p=0.002$ —both within-domain fits were significantly better than the permutation null (null mean ≈ -0.040). Definition note: the Iron-Core value of 0.080 comes from the unified primary protocol (log1p+standardization, nested α selection, same as §3.3); under the selector's original evaluation protocol (no standardization, pooled α selection) it was 0.0720 (previous paragraph), and the difference between the two

definitions arises only from evaluation preprocessing. By design, the primary evidence for Iron-Core remains cross-domain transferability (held-out zero-shot AUC=0.634 and paired bootstrap), and within-domain significance does not replace cross-domain evidence.

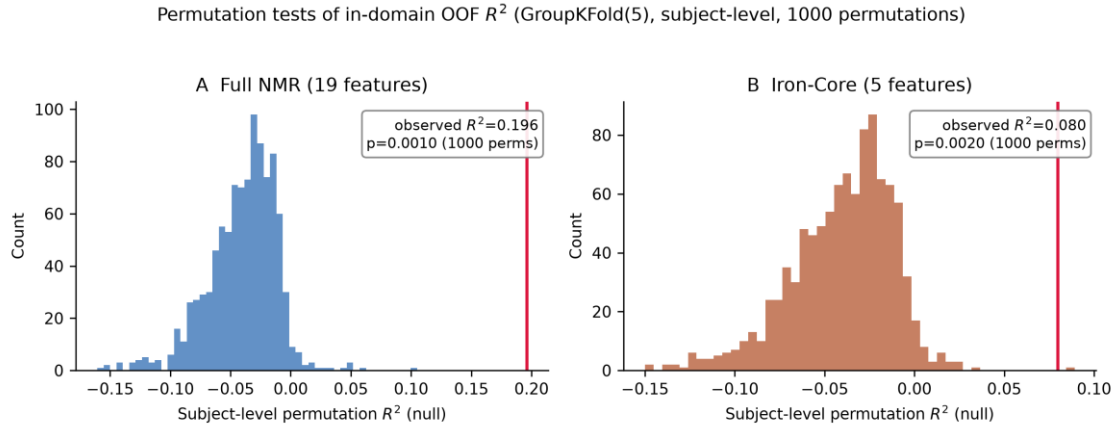


Figure 12. Subject-level permutation test of within-domain R^2 . A: full NMR (19 items, observed $R^2=0.196$, $p=0.001$); B: Iron-Core (5 items, observed $R^2=0.080$, $p=0.002$); null distribution=1,000 subject-level permutations (unified primary protocol, null mean ≈ -0.040), solid red line=observed.

Old protocol audit: The 2-item iron core (Uracil, Glucose) in the early protocol version was derived from a stability determination that included target-domain labels, which has selection optimism bias; it is deprecated and no longer used as a primary result; full audit in Appendix S1.

Threshold robustness (Figure 11): The prespecified gates ($\pi_{\text{uni_cell}} \geq 0.80$, $\text{KS} < 0.15$) lie on a plateau—threshold scanning showed that within $\pi \in [0.80, 0.85] \times \text{KSt} \in [0.15, 0.20]$ the core was always the same 5 items; the core shrank only when $\pi=0.90$ (to 3 items: Lysine/Tyrosine/Uracil) or $\text{KSt} \leq 0.12$ (the baseline row shrank to 3 items, with only amino acids remaining); relaxing to $\pi \leq 0.75$ increased it to 6 items (+Valine)—the qualitative conclusion of core composition was unchanged within ± 1 step (see §4.7 item 8).

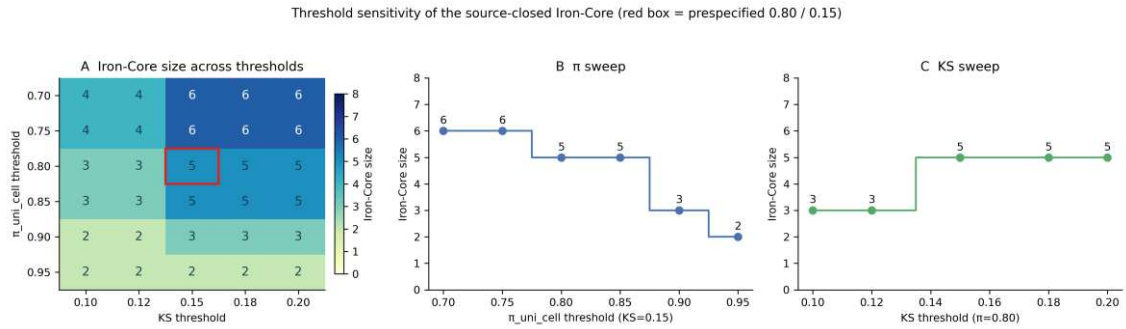


Figure 11. Threshold sweep for stability selection. In the figure, π_{uni_cell} (abbreviated π in the text) is source-domain univariate stability; KSt is the KS threshold. A: Iron-Core size on the grid $\pi \in [0.70, 0.95] \times KSt \in [0.10, 0.20]$ (number in each cell=number of core items, color=size; red box=prespecified 0.80/0.15, located on the plateau; $\pi \in [0.80, 0.85] \times KSt \in [0.15, 0.20]$ always yielded the same 5 items); B: π sweep with KSt=0.15 fixed ($\pi \leq 0.75 \rightarrow 6$ items, +Valine; $\pi = 0.90 \rightarrow 3$ items); C: KS sweep with $\pi = 0.80$ fixed (KSt $\leq 0.12 \rightarrow 3$ items, Uracil/Glucose cut, only amino acids remained).

3.10 Two-sample MR: genus $\rightarrow$ IBS risk (Figure 8)

Among 27 candidate genera, 313 pruned SNPs were obtained; after rsID matching and allele harmonization, 259 entered the IVW primary analysis (4–16 per genus; mean F statistic 20.6–28.9, all >10). Full results are shown in Figure 8; the top 5 nominal signals were:

Genus	nSNP	IVW OR (95% CI)	IVW p	Weighted median p	Egger intercept p	Cochran Q p	Observational ρ (Bristol)
Bifidobacterium	12	1.070 (1.004–1.141)	0.039	0.189	0.868	0.717	−0.215 (q=7×10 ^{−5})
Eubacterium hallii group	12	1.074 (0.996–1.157)	0.062	0.196	0.937	0.894	+0.227 (q=3×10 ^{−5})
Barnesiella	12	1.051 (0.980–1.128)	0.165	0.297	0.787	0.646	−0.154 (q=0.007)
Coproccoccus1	11	1.060 (0.976–1.151)	0.168	0.021	0.647	0.226	+0.072
Roseburia	11	0.930 (0.828–1.044)	0.219	0.130	0.607	0.070	+0.284 (q=4×10 ^{−8})

The remaining 22 genera had IVW $p > 0.23$. **Under the current instrumental-variable statistical power and distance-pruning approximation framework, no genus showed a robust potential causal effect (minimum FDR=0.835).** Full 27-genus results and STROBE-MR item mapping are in Appendix S2.

Key points: (i) The observed number with nominal $p < 0.05$ (1) was indistinguishable from the global null expectation ($27 \times 0.05 = 1.35$), consistent overall with “no robust potential causal effect.” (ii) The nominally strongest Bifidobacterium and Eubacterium hallii group showed no evidence of horizontal pleiotropy (Egger intercept $p > 0.85$), low heterogeneity (Q $p > 0.71$), and stable LOO (maximum shift < 0.03), but neither reached

FDR; Coprococcus1's nominal WM $p=0.021$ was inconsistent with IVW $p=0.168$, suggesting that this signal is not robust under the approximate LD-pruning framework. (iii) Directional controls did not converge: Bifidobacterium was nominally in the risk direction ($OR>1$), whereas its observational Bristol correlation was negative (higher abundance closer to the constipation side); Roseburia was nominally protective ($OR=0.930$) but had the most stable positive observational correlation ($\rho=+0.284$). (iv) Figure 8 shows that the overall estimates were approximately symmetric around the null line—consistent with the picture that “genus main effects on IBS are limited, and observational associations are more likely affected by confounding/reverse causation.” (v) Independence of evidence chains: MR exposure is a genetic proxy at the genus level, whereas Iron-Core is a metabolite panel, with no one-to-one correspondence—this analysis tests the potential causal effect of genus→IBS and does not constitute a test of the “genus→metabolite→IBS” chain; the two evidence lines are independent.

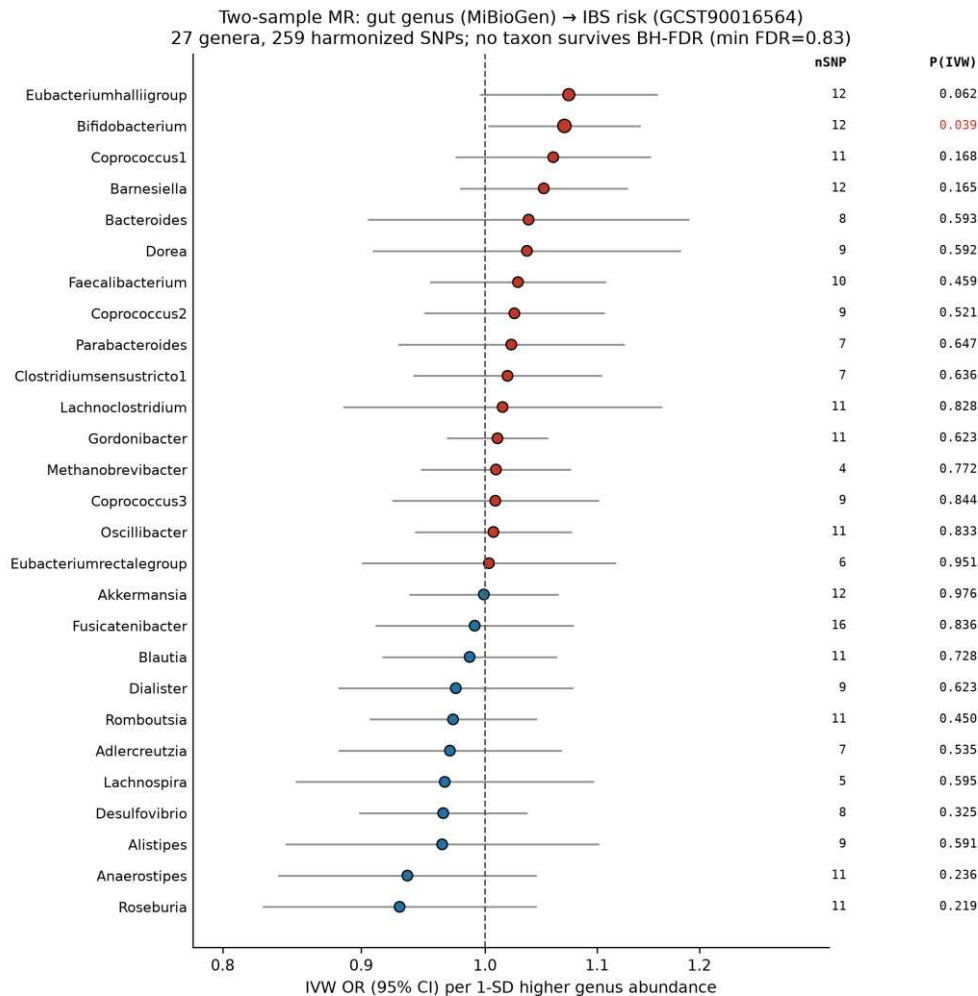


Figure 8. Two-sample MR: IVW effects of 27 candidate genera → IBS risk (forest plot). Each row=genus name, 95% CI, IVW OR (red >1, blue <1), number of instruments (nSNP), and nominal p; dashed line=null line OR=1. No genus passed BH-FDR (minimum FDR=0.835); the strongest nominal signals were Bifidobacterium (OR=1.070, p=0.039) and Eubacterium hallii group (OR=1.074, p=0.062). Exposure: MiBioGen ($p < 10^{-5}$, ± 1 Mb distance pruning); outcome: GCST90016564 (Eijsbouts 2021).

4 Discussion

Three mechanistic axes (SCFA–5-HT, proteolytic fermentation, bile acid–TGR5) share a more parsimonious common upstream: **intestinal transit time**. Faster transit simultaneously produces: (i) more carbohydrate/protein substrate escaping to distal fermentation (SCFA, BCAA, lysine, and tyrosine rise in parallel); (ii) more pronounced bile acid escape (BAD spectrum); (iii) higher mucosal turnover and nucleic acid degradation (uracil); (iv) a shorter water/salt absorption window (higher Bristol, increased glucose delivery). Within IBS-C, positive correlations remained, and all 5 clean core features under the closed-loop protocol point to transit–substrate exposure processes, consistent with the parsimonious model of “transit time as the main effect and each layer's

signal as its mechanistic branch”; by Occam's principle, this paper does not prioritize three independent parallel mechanisms (distinguishing them requires direct measurement of orocecal transit time, planned next).

4.1 SCFA–serotonin–intestinal motility axis

Acetate ($\rho=0.414$), butyrate (0.379), and propionate (0.297) were positively correlated with Bristol (higher=looser). Mechanistically, SCFAs activate **FFAR2 (GPR43)/FFAR3 (GPR41)** [14] on enterochromaffin cells (EC cells), upregulate tryptophan hydroxylase TPH1, and promote 5-HT release (Reigstad et al., FASEB J 2015 [15]; Yano et al., Cell 2015 [16] confirmed that microbiota–SCFA drive colonic 5-HT synthesis); 5-HT amplifies the peristaltic reflex via 5-HT₃/5-HT₄ receptors [17] on intrinsic primary afferent neurons → faster transit → higher fecal water content. In addition, SCFAs are poorly absorbed anions, and their accumulation in the distal colon produces an osmotic water-retaining effect. Together, the two pathways explain the directional consistency of “high SCFA in IBS-D ↔ loose stools.” Succinate (+0.273), as a SUCNR1 (GPR91) [18] agonist, can also promote 5-HT release from EC cells, in the same direction as the main effect. Branched SCFAs (isovalerate, 2-methylbutyrate: fermentation products of BCAAs) were negatively correlated, suggesting a contrasting picture of a different fermentation niche (slow transit, carbohydrate-replete environment). This mechanistic narrative is consistent with the Shapley decomposition in §3.4b: the SCFA group's marginal predictive contribution ($\phi=0.106$) was greater than that of the amino acid group ($\phi=0.045$); removing SCFAs decreased the full-model R^2 by 0.076 ($p=0.002$), whereas removing amino acids decreased it by only 0.016 ($p=0.085$, not significant)—the SCFA axis is not only the strongest in univariate correlation but also the group with the largest multivariate marginal contribution.

4.2 Proteolytic fermentation fingerprint: BCAAs and aromatic amino acids

Leucine/valine/isoleucine (0.19–0.23), lysine (0.374), and tyrosine (0.311) were synchronously positive, consistent with the substrate-delivery model of “fast transit → proximal carbohydrate depletion → enhanced distal proteolytic fermentation.” Microbial decarboxylation products further amplify motility effects: tyrosine → **tyramine** (TAAR1 agonist, promotes 5-HT release); lysine → cadaverine/polyamines (polyamines are associated with IBS-D; Jacobs et al., Microbiome 2023 [3] also reported elevated polyamines and bile acids in the D subtype, cross-validating the direction in our cohort). **2-Methylbutyrate (−0.185) was the only significant negative marker**, complementary to

the positive BCAA profile, suggesting that the “free state vs fermentation consumption” of branched-chain acids reflects opposite transit ecologies. Unlike §4.1, the group-level marginal contribution of proteolytic fermentation-related amino acids (lysine, tyrosine, leucine) did not reach permutation significance ($\varphi_{AA}=0.045$, $p=0.085$, §3.4b); the narrative in this section is based on the directional consistency of univariate correlations (§3.4), which is not equivalent to an independent group-level predictive contribution—the independent increment of the amino acid profile was not confirmed at the current sample size.

4.3 Bile acid–TGR5 secretomotor axis

Primary bile acids (CA, CDCA), glycine-conjugated forms (GCA, GCDCA, GDCA), and DCA-sulfate were consistently positively correlated (ρ 0.23–0.31). Mechanism: bile acids escaping ileal reabsorption reach the colon and, via **TGR5 (GPBAR1)** [19] on EC cells/enteric neurons $\rightarrow$ cAMP $\rightarrow$ 5-HT release + CFTR-dependent Cl^- /water secretion, and accelerate colonic motility—the classic BAD (bile acid diarrhea) pathway [20]. About one-third of IBS-D patients have BAD [20]; this study suggests that the mechanism exists with weak-to-moderate strength across the **full continuous cohort spectrum**, not only in the D subtype.

4.4 “Potential $\neq$ activity”: methodological implications of cross-layer validation

Cross-layer associations show: (i) bcd (a genomic abundance marker of butyrate-producing communities, not the catalytic rate-limiting step) was significantly associated with the butyrate pool ($\rho=0.265$), demonstrating that metagenomic butyrate-producing capacity can partially reflect the metabolite pool at the community level; (ii) however, the signal disappeared after aggregation (7-gene sum, ko00650), indicating that **community-level aggregation systematically dilutes biological signal**; (iii) BSH/bai showed zero association with the bile acid pool, warning that functional gene abundance cannot replace enzyme activity/metabolite measurement. Direct implications for “multimodal fusion”: **the metabolic layer should be a first-tier feature layer, and the metagenomic functional layer has conditional value only at single-gene/single-enzyme resolution**; simple concatenation (NMR+genus) showed no gain when unfiltered (0.171 vs 0.196), and after filtering at the same prevalence was comparable to NMR (0.204 vs 0.196, ΔR^2 not significant); fusion strategies still require layer-wise weighting or causal structure (next step).

4.5 Interpretability of the stable core: predictive magnitude $\neq$ cross-domain transferability (closed-loop protocol)

The closed-loop protocol rewrote an important ranking from the old protocol: the strongest within-domain features such as acetate/butyrate ($\pi_{\text{uni_cell}}=1.0$) were objectively excluded because they were unmeasurable on the target platform; the old protocol had used target-domain labels in the determination, excluding lysine/tyrosine (Jacobs π_{uni} 0.43/0.08)—under the closed-loop protocol, these two entered the core because of extremely strong source-domain stability ($\pi_{\text{uni_cell}}$ 1.00/0.99) and minimal platform shift (KS 0.070/0.060), and were shown in held-out evaluation to be not significantly different from R11. Methodological warning: **out-of-domain evidence for feature selection should come only from unlabeled signals (platform shift KS), not from target-domain labels**—the latter systematically discards features that are robust in the source domain and measurable on the platform (the old protocol's 2-item iron core was a product of this optimism bias; see Appendix S1).

Biologically, the 5-item core consists of two levels with different stability structures. **Slow-variable anchors:** **uracil** comes from nucleic acid degradation (host epithelial shedding and microbial RNA turnover), reflecting luminal cell turnover and transit exposure time; **glucose** reflects carbon-source delivery and proximal absorption efficiency (more available carbon enters the colon during fast transit)—both point to the slow process of “transit time determining substrate delivery and luminal exposure,” with high point-estimate $\pi_{\text{uni_cell}}$ (0.910/0.883), but their gate values had substantial estimation noise in the fully nested analysis (π gate frequency 0.27/0.42; KS gate 0.78/0.63), and cross-cohort replication requires external confirmation. **Chemically quantifiable fast variables:** **lysine and tyrosine** are proteolytic fermentation fingerprints (§4.2), and **leucine** belongs to the BCAA covariance spectrum (0.19–0.23)—these three are small molecules that can be routinely detected on both NMR and LC-MS platforms and had the smallest distribution shift between the two platforms in this cohort (KS 0.058–0.070, §3.9), with the highest stability under fully nested analysis (KS gate frequency 1.00/1.00/1.00; all-gates-off selection frequency lysine 0.96, tyrosine 0.79, leucine 0.36). The two-level nested sensitivity therefore gives a more precise picture: the core's **conditional stability** (fixed prespecified gates) fully holds (survival 0.997–0.803), while **unconditional robustness** concentrates on lysine/tyrosine; cross-platform transferability depends on both source-domain stability and the chemical quantification maturity of the feature.

Hypoxanthine is a valuable process case: in the old protocol it appeared extremely stable because of Jacobs $\pi_{\text{uni}}=0.993$, but its $\text{KS}=0.188$ exceeded the threshold and its source-domain $\pi_{\text{uni_cell}}$ was only 0.570—platform-shift diagnostics (KS) can warn of such non-transferable features **without any contact with target-domain labels**, which is exactly the engineering value of the closed-loop protocol.

Methodological conclusion: the minimal transferable feature set for cross-center applications should be **jointly determined by source-domain stability and unlabeled platform diagnostics**, rather than by effect-size ranking, and target-domain labels should not be allowed to participate in decisions; the F16→R11→Iron-Core comparison shows: F16 collapsed because of non-transferable feature poisoning (0.558), R11 relied on the 11-item intersection to reach 0.647, and Iron-Core reached 0.634 with only 5 items (not significantly different from R11)—dimensionality reduction did not lose discrimination while providing a more compact, cross-platform quantifiable core set; the two-level nested sensitivity gives its boundary: conditional stability fully reproduced, unconditional robustness concentrated on Lysine/Tyrosine (0.96/0.79), and the other three await external replication.

4.6 Two-sample MR: potential causal assessment of observational microbial associations

In the two-sample MR of 27 candidate genera × IBS GWAS, no genus passed FDR correction, and nominal signals were weak (strongest $\text{OR}\approx 1.07$)—consistent with the methodological status of the field: single-genus instruments explain little variance (4–16 SNPs per genus; even meeting the weak-instrument lower bounds of $p<10^{-5}$ and $F>10$, explanatory power for the outcome remains limited), IBS phenotypes are highly heterogeneous (case definition, subtype mixing, age/diet modification), and the systematic power of genetic proxies is constrained. The ± 1 Mb distance pruning used here is a conservative approximation without an LD panel: retaining the lowest-p SNP within a window may still leave residual correlation with nearby loci, causing IVW variance to be slightly underestimated; however, the maximum LOO shift <0.05 indicates that the conclusion is not driven by a single point.

Observational–causal directional controls did not converge: *Bifidobacterium*'s nominal risk direction was semantically incompatible with its negative observational Bristol correlation (more toward constipation); *Roseburia*'s MR protective trend ($\text{OR}=0.930$) was

opposite to the stable positive correlation in this cohort ($\rho=+0.284$, $q=4\times 10^{-8}$). Possible explanations include: (i) observational associations are driven by strong confounders such as diet/transit time, whereas MR estimates the net effect of lifelong genetically predicted abundance differences—the two do not target the same counterfactual; (ii) reverse causation: IBS symptoms (and even subclinical motility abnormalities) alter microbiota composition, and MR captures the projection of the net total effect at the genetic level; (iii) different outcome semantics: the cohort phenotype is the continuous Bristol axis, whereas the MR outcome is binary IBS risk. Overall, this study does not support causal conclusions for any genus under the current data; the nominal signals of *Bifidobacterium* and *Eubacterium hallii* group deserve re-examination in larger, subtype-stratified, rigorously LD-pruned frameworks. The methodological boundary must be explicit: this study **does not report MVMR estimates** (unreliable under the current instrument strength)—MVMR estimation of the joint effect of interacting genera requires (i) strong instruments with conditional $F>10$ and (ii) a genetic covariance (LD) matrix between exposures; with 4–16 harmonized SNPs per genus in this cohort, neither can be reliably constructed, and forced execution would fall into weak-instrument bias with uninterpretable estimates. MVMR is therefore left for larger, subtype-stratified microbiome GWAS with complete LD panels (§4.8).

4.7 Limitations

1. **Cross-sectional/observational:** intestinal transit time itself changes substrate delivery and absorption, so reverse causation cannot be excluded; however, signal remained within IBS-C ($\rho=0.288$, 95% CI [$+0.074$, $+0.483$]), supporting that it is not purely a label difference.
2. **Within-subtype sample size and power constraints:** IBS-C 128 samples/21 subjects, IBS-D 154 samples/23 subjects; unsupervised subgrouping (DMM/hierarchical clustering) within a single subtype at $n\approx 150$ is underpowered and prone to pseudo-convergence, so the IBS-D negative signal in §3.5 (no robust effect detected) is only an exploratory hypothesis-generating observation and does not constitute a robustness conclusion; within-subtype subtyping requires replication in larger cohorts.
3. **Bile acid 30.5% missing** was imputed by min/2, possibly underestimating weak signals; tryptophan layer $n=259$.

4. The functional layer is based on relative abundance and KEGG annotation, without metatranscriptome/metaproteome/metabolic flux validation.
5. Bristol is a diary-window aggregate ($\pm 3d/\pm 7d/\text{subject mean}$), with measurement noise.
6. No deep learning/GNN was used: with <500 samples and a disparate feature-to-sample ratio, regularized linear models better fit the bias–variance tradeoff; GNNs require larger samples and external cohort support.
7. Cross-cohort validation covered only the metabolic layer (Jacobs has no SCFA measurements and no matched taxonomic matrix): AUC=0.634 (Iron-Core)/0.647 (R11) are **lower-bound estimates for the transferable subset**; zero-shot validation of the taxonomic layer requires an available abundance matrix in the target cohort (planned for the next stage). In addition, cross-cohort domain shift is a **composite shift** (platform $\times$ feature availability $\times$ sample processing and study protocol); KS can only measure its composite magnitude and cannot decompose it—“platform robustness” should be understood as robustness to composite shift (§2.5).
8. **Stability thresholds are prespecified empirical values** ($\pi_{\text{uni_cell}} \geq 0.8$, $\text{KS} < 0.15$, $B=300$): threshold scanning (Figure 11) shows that $(0.80, 0.15)$ lies on a plateau— $\pi \in [0.80, 0.85] \times \text{KSt} \in [0.15, 0.20]$ always yielded the same 5 items; $\pi \leq 0.75$ increased to 6 items (+Valine), $\pi=0.90$ shrank to 3 items (Glucose/Leucine with $\pi=0.883/0.883$ were cut, leaving Lysine/Tyrosine/Uracil), and $\text{KSt} \leq 0.12$ shrank the baseline row ($\pi=0.80$) to 3 items (Uracil/Glucose with $\text{KS}=0.134/0.138$ were cut, leaving only Lysine/Tyrosine/Leucine; the $\pi \geq 0.90$ row further dropped to 2 items); the qualitative conclusion of core composition was unchanged within ± 1 step. In Jacobs, SCFAs were unmeasurable, so their cross-cohort transferability could not be evaluated (treated as non-transferable); under the closed-loop protocol, Iron-Core and R11 were not significantly different ($\Delta\text{AUC}=-0.013$, CI crossing zero), and the difference from F16 (+0.075) was consistent with the poisoning mechanism of unmeasurable SCFAs. **Two-level nested sensitivity** further limits the conclusion: conditional stability (fixed gates) fully reproduced (0.997–0.803); under fully nested gate recomputation, selection frequencies were graded as Lysine 0.96 > Tyrosine 0.79 >> Leucine/Glucose/Uracil (0.23–0.36), with the bottleneck being estimation noise of the π gate (Monte Carlo SE of selection frequency under $B_{\text{out}}=100 \approx 0.02\text{--}0.05$: Lysine ≈ 0.02 , Tyrosine ≈ 0.04 , the other three 0.04–0.05; the difference between Tyrosine's 0.79 and the 0.80 gate boundary is within error, so its “unconditional robustness” should be read as “near the

gate boundary, direction consistent,” not “stably passing the gate”; Lysine's 0.96 has an SE of about 0.02, significantly above the boundary); the nesting depth is still limited to the 19-item NMR panel and requires re-examination in more independent cohorts.

9. The zero-shot endpoint is a BH motility proxy (not the Bristol gold standard), with label-semantic approximation error.
10. **MR robustness and power:** (i) ± 1 Mb distance pruning is only an approximation without an LD panel; (ii) FDR conclusions for 27 genera are sensitive to instrument thresholds and harmonization procedures; (iii) GCST90016564 is an overall IBS phenotype (without subtype distinction); (iv) the degree of sample overlap between MiBioGen and GCST90016564 is unknown—if overlap exists, estimates would partially inherit observational confounding; (v) There is no correspondence between MR exposure (genus) and Iron-Core (metabolite); they are independent evidence lines, and MR did not test the “genus→metabolite→IBS” mediation chain (§3.10 key point (v)); (vi) GCST90016564 is an **overall IBS** phenotype, and subtype heterogeneity (D/C mechanistic differences) would dilute subtype-specific effects; MR negativity should not be extrapolated to “no potential causal effect at the subtype level.” (vii) This study **does not report MVMR estimates** (unreliable under current instrument strength): conditional $F > 10$ instrument strength and an inter-exposure LD matrix are unavailable with 4–16 harmonized SNPs per genus, joint exposure separation was not tested, and this is a boundary statement for an independent evidence line (§4.6, §4.8).
11. Observational control ρ came from different cohorts/platforms and is only a directional reference, not a magnitude comparison.
12. **NMR internal structure has been decomposed (§3.4b, Figure 13):** two-group Shapley decomposition showed $\varphi_{\text{SCFA}}=0.106$, $\varphi_{\text{AA}}=0.045$ (joint gain 0.151); in subject-level ΔR^2 permutation, removing the SCFA group decreased R^2 by 0.076 ($p=0.002$, significant), and removing the amino acid group decreased it by 0.016 ($p=0.085$, not significant)—“SCFA dominance” holds under current evidence, and the independent increment of the amino acid profile is not confirmed (6 of 7 AA-group items reached univariate significance: alanine, leucine, valine, isoleucine, lysine, tyrosine; only glycine did not, and its within-group contribution was not separately decomposed). Remaining boundary: the decomposition is based on linear Ridge and a fixed three-group partition (5/7/7); within-group feature-level contributions (e.g.,

within BCAA, lysine/tyrosine relative) were not further decomposed, and feature-level Shapley can be performed in larger samples.

13. **Iron-Core within-domain fit has been supplemented with a permutation test (§3.9, Figure 12):** under the unified primary protocol, within-domain $R^2=0.080$ (selector original evaluation protocol 0.072), subject-level permutation $p=0.002$ (1,000 permutations)—the within-domain fit is significant but limited in magnitude ($R^2\approx 0.08$); Iron-Core's positioning still relies primarily on cross-domain transferability (held-out zero-shot AUC=0.634 and paired bootstrap), and within-domain significance does not guarantee cross-domain generalization.
14. **Random/demographic baselines for the zero-shot task have been supplemented:** 200 random 11-feature sets AUC=0.506 (95% quantile 0.434–0.589, $p=0.005$); age/BMI/sex model AUC=0.593 (95% CI 0.509–0.675); R11 relative to demographics $\Delta\text{AUC}=+0.053$ (95% CI -0.045 – $+0.153$, $p=0.148$), increment not significant—AUC=0.634/0.647 should still be interpreted as exploratory evidence and cannot claim that the metabolome is independent of demographic variables.
15. **Historical correction of the selection protocol (important):** The early protocol version's Iron-Core (2 items: Uracil, Glucose; full audit in Appendix S1) was derived from a stability determination that included target-domain labels, with selection optimism bias; the current protocol is the source-domain closed-loop protocol (label-blind transductive), with zero contact with target-domain labels in all selection steps, and all cross-cohort numbers are based on this protocol.
16. **Aqueous dilution and transit time cannot be separated:** fecal NMR concentrations were not corrected for water content/dry-wet weight (the available metadata lack this field), so dilution effects and true concentration changes are indistinguishable; transit time affects all concentration-type markers simultaneously, and single-time-point data cannot fully decouple them—future cohorts should add fecal water content and orocecal transit time measurements.
17. **Boundaries of common-subset sensitivity and cross-layer comparison:** common-subset sensitivity assumes missing at random; if missingness of tryptophan/bile acids is phenotype-related, subset results may be affected by selection bias. The common-subset regression $n=255/67$ (limited by Bristol availability) and classification $n=268/75$ (all subjects) were constructed separately and are not the same samples; cross-task comparisons require caution. Cross-layer R^2 /balanced accuracy differences

were not tested with paired bootstrap or DeLong-type tests; cross-layer differences in the text should be read as point-estimate magnitudes, not statistical significance claims; in the classification task, NMR and taxonomic layers were indistinguishable on the common subset and the ordering was not robust, requiring re-examination in fully paired, complete multi-omics subsets. The IBS-SSS task did not undergo common-subset sensitivity.

4.8 Conclusions and outlook

Conclusions: Under strict subject-grouped validation, **the layer with the highest point-estimate OOF R^2 for the Bristol continuous-trait regression task** was the NMR metabolome (SCFA + amino acids, $R^2=0.196$, $\rho=0.468$; **Shapley decomposition:** $\phi_{\text{SCFA}}=0.106$ vs $\phi_{\text{AA}}=0.045$, **remove SCFA $\Delta R^2=0.076/p=0.002$, remove AA $0.016/p=0.085$**), and the metabolite–phenotype coupling was reproducible within IBS-C; **however, in the three-class task, NMR and taxonomic layers were indistinguishable on the common subset (NMR 0.556, genus 0.556, species 0.559, combined 0.565), and the cross-layer ordering in the classification task was not robust; therefore, the single-layer ranking “direct metabolic measurement > taxonomy > functional potential” applies only to the Bristol regression task and should not be extrapolated to subtype classification or combined models (after the same filtering, the combined model was not significantly different from NMR);** functional gene potential was coupled with the metabolite pool only at single-gene (bcd) resolution; held-out zero-shot cross-cohort validation (R11 AUC=0.647, 11/11 direction consistent) and source-domain closed-loop stability selection further conditionally stabilized the platform-robust core to 5 metabolites (Iron-Core: lysine, tyrosine, uracil, glucose, leucine; under fully nested analysis, unconditional robustness concentrated on lysine/tyrosine, the other three conditionally stable; held-out zero-shot AUC=0.634, not significantly different from R11, significantly better than F16). The results support three mechanistic axes in terms of directional consistency, but with different levels of evidence: **SCFA→serotonin→intestinal motility** is supported mainly by within-NMR evidence (Shapley $\phi=0.106$, consistent marker direction spectrum); the **proteolytic fermentation fingerprint** is supported by the directional consistency of BCAAs/aromatic amino acids; and the **bile acid–TGR5 secretomotor** axis is supported only by the directional positive correlation of the bile-acid layer Top6 with loose stools, lacking functional/causal validation (“potential $\neq$ activity”); and it provides the methodological conclusion “predictive magnitude $\neq$ cross-domain

transferability.” External random and demographic baselines further limit the interpretation: R11 was better than random features, but its increment over age/BMI/sex was not significant ($\Delta\text{AUC}=+0.053$, $p=0.148$), and the independent discriminative value of the metabolome still requires larger samples.

Outlook (next steps): (i) On the basis of the completed univariate two-sample MR (§3.10; no FDR-significant effect detected), extend causal inference: species-level/pathway-level exposures with more IBS GWAS (including East Asian cohorts), and mediation MR using metabolite GWAS to test the “genus→metabolite→IBS” chain; **this study does not report MVMR estimates** (unreliable under current instrument strength): separating interacting genera requires instruments with conditional $F>10$ and an inter-exposure genetic covariance (LD) matrix; with only 4–16 harmonized SNPs per genus, conditional strength is likely insufficient, and joint estimation under weak instruments is uninterpretable; MVMR must await larger, subtype-stratified microbiome GWAS and complete LD panels. (ii) Deepen validation of domain adaptation (feature shrinkage/domain adversarial) in more independent cohorts (e.g., PRJEB34103, GMrepo multi-project integrated cohorts); (iii) replace feature concatenation with structured multimodal models and weight by stability strata; (iv) on the basis of the completed source-domain closed-loop protocol, two-level nested sensitivity, threshold scanning, and paired bootstrap, as well as the newly added NMR between-group Shapley decomposition/ ΔR^2 permutation (§3.4b, Figure 13) and within-domain fit permutation test (§3.9, Figure 12), continue to supplement: establishment of demographic/random baselines for the Jacobs zero-shot task, and decomposition of within-group feature-level contributions (e.g., within-BCAA structure).

5 Data and code availability

- **Data:** raw sequencing data at ENA PRJEB37924 [2]; processed multi-omics matrices at Mendeley Data (10.17632/29n2z5r5ph.3) [24]; external validation cohort at PRJNA812699 [3]; taxonomic and functional annotation abundances from GMrepo [11].
- **MR data sources:** MiBioGen full hit table [4] (122,110 rows / 211 taxa) and IBS GWAS summary statistics GCST90016564 [5] (53,400 cases / 433,201 controls; Eijbouts et al., 2021), both public.

- **Analysis code and result tables:** the analysis code for eight-layer feature construction, fair comparison protocol, classification uncertainty, external baselines, cross-cohort validation, and two-sample MR, together with primary result tables and derived data, are provided as Supplementary Code S1 (IBS_Supplementary_Code_S1_v1.0.zip) / Supplementary Data S1 (IBS_Supplementary_Data_S1_v1.0.zip) with this submission. The analysis code is publicly available at <https://github.com/toby27227-dev/IBS-multilayer-comparison>; a permanent Zenodo DOI will be added once archiving is complete. Primary result tables are no longer described as available only from the author on request; supplementary materials S1–S3 are in the appendices.

6 Declarations

Author contributions: Yi Sui was responsible for study design, data analysis, interpretation, and manuscript writing.

Funding: This study received no funding.

Conflicts of interest: The author declares no conflicts of interest.

Ethics: This study is a secondary analysis of publicly available de-identified data (Cell 2020: ENA PRJEB37924; Jacobs 2023: PRJNA812699); no new participant recruitment or sample collection was performed. The original studies obtained the relevant ethics approvals and informed consent, and data use followed the original data licenses (the Cell 2020 processed matrices are CC BY 4.0). Under the authors' institutional and national regulations for secondary analysis of de-identified public data, no additional ethics approval was required. The study was conducted in accordance with the Declaration of Helsinki.

References

1. Enck P, Aziz Q, Barbara G, Farmer AD, Fukudo S, Mayer EA, et al. Irritable bowel syndrome. *Nat Rev Dis Primers*. 2016;2:16014. doi:10.1038/nrdp.2016.14.
2. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell*. 2020;182(6):1460-1473.e17. doi:10.1016/j.cell.2020.08.007.
3. Jacobs JP, Lagishetty V, Hauer MC, Labus JS, Dong TS, Toma R, et al. Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. *Microbiome*. 2023;11:5. doi:10.1186/s40168-022-01450-5.
4. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat Genet*. 2021;53(2):156-165. doi:10.1038/s41588-020-00763-1.
5. Eijsbouts C, Zheng T, Kennedy NA, Bonfiglio F, Anderson CA, Moutsianas L, et al. Genome-wide analysis of 53,400 people with irritable bowel syndrome highlights shared genetic pathways with mood and anxiety disorders. *Nat Genet*. 2021;53(11):1543-1552. doi:10.1038/s41588-021-00950-8.
6. Lewis SJ, Heaton KW. Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol*. 1997;32(9):920-924. doi:10.3109/00365529709011203.
7. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. *Aliment Pharmacol Ther*. 1997;11(2):395-402. doi:10.1046/j.1365-2036.1997.142318000.x.
8. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*. 2012;7(2):e30126. doi:10.1371/journal.pone.0030126.
9. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Knight R, Knights D. SHOGUN: a modular, accurate and scalable framework for microbiome quantification. *Bioinformatics*. 2020;36(13):4088-4090. doi:10.1093/bioinformatics/btaa277.
10. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*. 2021;10:e65088. doi:10.7554/eLife.65088.
11. Dai D, Zhu J, Sun C, Li M, Liu J, Wu S, et al. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res*. 2022;50(D1):D777-D784. doi:10.1093/nar/gkab1019.
12. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the reporting of observational studies in epidemiology using Mendelian randomization: the STROBE-MR statement. *JAMA*. 2021;326(16):1614-1621. doi:10.1001/jama.2021.18236.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.
14. Brown AJ, Goldsworthy SM, Barnes AA, Eilert MM, Tcheang L, Daniels D, et al. The orphan G protein-coupled receptors GPR41 and GPR43 are activated by propionate and other short chain carboxylic acids. *J Biol Chem*. 2003;278(13):11312-11319. doi:10.1074/jbc.M211609200.

15. Reigstad CS, Salmonson CE, Rainey JF 3rd, Szurszewski JH, Linden DR, Sonnenburg JL, et al. Gut microbes promote colonic serotonin production through an effect of short-chain fatty acids on enterochromaffin cells. *FASEB J.* 2015;29(4):1395-1403. doi:10.1096/fj.14-259598.
16. Yano JM, Yu K, Donaldson GP, Shastri GG, Ann P, Ma L, et al. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell.* 2015;161(2):264-276. doi:10.1016/j.cell.2015.02.047.
17. Gershon MD, Tack J. The serotonin signaling system: from basic understanding to drug development for functional GI disorders. *Gastroenterology.* 2007;132(1):397-414. doi:10.1053/j.gastro.2006.11.002.
18. He W, Miao FJ, Lin DC, Schwandner RT, Wang Z, Gao J, et al. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature.* 2004;429(6988):188-193. doi:10.1038/nature02488.
19. Kawamata Y, Fujii R, Hosoya M, Harada M, Yoshida H, Miwa M, et al. A G protein-coupled receptor responsive to bile acids. *J Biol Chem.* 2003;278(11):9435-9440. doi:10.1074/jbc.M209706200.
20. Walters JR, Pattni SS. Managing bile acid diarrhoea. *Therap Adv Gastroenterol.* 2010;3(6):349-357. doi:10.1177/1756283X10377126.
21. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol.* 2015;44(2):512-525. doi:10.1093/ije/dyv080.
22. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol.* 2016;40(4):304-314. doi:10.1002/gepi.21965.
23. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol.* 2013;37(7):658-665. doi:10.1002/gepi.21758.
24. Mars R. Mars et al, IBS multi-omics [Data set]. Mendeley Data. 2020. doi:10.17632/29n2z5r5ph.3.

Appendix S1 Audit of the old protocol (not a primary result)

The early Iron-Core protocol (2 items: Uracil, Glucose) used target-domain labels in the stability determination and achieved zero-shot AUC=0.668, $\rho=+0.314$; this number has selection optimism bias and has been deprecated. The current protocol is the source-domain closed-loop protocol (label-blind transductive), with zero contact with target-domain labels during selection; all cross-cohort numbers in the main text are based on the current protocol. The complete determination table and code for the old protocol are retained in historical files under IBS/jacobs_validation/ for audit only.

Appendix S2 Key STROBE-MR items and complete results for 27 genera

STROBE-MR item	Corresponding content in this study
Exposure data source	MiBioGen microbiome GWAS [4] (Kurilshikov et al., 2021; 16–24 cohorts; 211 taxa)
Outcome data source	IBS GWAS GCST90016564 [5] (Eijssbouts et al., 2021; 486,601 European samples, 53,400 cases/433,201 controls)
Instrument selection	$p < 10^{-5}$; F statistic 100% > 10 ; ± 1 Mb distance pruning to approximate LD removal (retain the lowest-p SNP within the window)
Harmonization	rsID matching 305/313; strand flipping and complementary pair handling; palindromic SNPs conservatively removed; 259/313 entered estimation
Primary analysis	Random-effects IVW (delta-method SE)
Sensitivity	Weighted median (bootstrap 2000), MR-Egger intercept, Cochran Q, leave-one-out
Multiple correction	BH-FDR of IVW p values for 27 genera
Software	Python 3.13.13; NumPy 2.4.6; SciPy 1.17.1; pandas 3.0.3; scikit-learn 1.8.0
Limitations	Distance pruning is not full LD; overall IBS phenotype; sample overlap unknown; MVMR not reported

Table S2. Complete two-sample MR results for 27 candidate genera (sorted by IVW p ascending; OR>1 indicates the risk direction).

Genus	nSNP	IVW OR (95% CI)	IVW p	BH-FDR	Cochran Q p	Egger intercept p	LOO max shift
Bifidobacterium	12	1.070 (1.004–1.141)	0.039	0.835	0.717	0.868	0.024
Eubacterium hallii group	12	1.074 (0.996–1.157)	0.062	0.835	0.894	0.937	0.020
Barnesiella	12	1.051 (0.980–1.128)	0.165	0.911	0.646	0.787	0.021
Coprococcus1	11	1.060 (0.976–1.151)	0.168	0.911	0.226	0.647	0.026
Roseburia	11	0.930 (0.828–1.044)	0.219	0.911	0.070	0.607	0.033
Anaerostipes	11	0.936 (0.839–1.044)	0.236	0.911	0.089	0.848	0.035
Desulfovibrio	8	0.965 (0.899–1.036)	0.325	0.911	0.585	0.940	0.019
Romboutsia	11	0.973 (0.907–1.044)	0.450	0.911	0.712	0.603	0.021
Faecalibacterium	10	1.028 (0.955–1.108)	0.459	0.911	0.582	0.014	0.020
Coprococcus2	9	1.025 (0.950–1.107)	0.521	0.911	0.449	0.870	0.020
Adlercreutzia	7	0.971 (0.883–1.067)	0.535	0.911	0.135	0.917	0.033
Alistipes	9	0.964 (0.844–1.101)	0.591	0.911	0.070	0.097	0.038
Dorea	9	1.036 (0.910–1.181)	0.592	0.911	0.030	0.090	0.044
Bacteroides	8	1.038 (0.906–1.189)	0.593	0.911	0.080	0.437	0.043
Lachnospira	5	0.966 (0.852–1.096)	0.595	0.911	0.344	0.785	0.043
Gordonibacter	11	1.011 (0.969–1.054)	0.623	0.911	0.351	0.335	0.013
Dialister	9	0.975 (0.883–1.077)	0.623	0.911	0.128	0.771	0.029
Clostridium sensu stricto 1	7	1.019 (0.941–1.104)	0.636	0.911	0.493	0.333	0.028
Parabacteroides	7	1.023 (0.929–1.125)	0.647	0.911	0.754	0.177	0.025
Blautia	11	0.987 (0.917–1.062)	0.728	0.911	0.916	0.861	0.027
Methanobrevibacter	4	1.009 (0.948–1.075)	0.772	0.911	0.616	0.988	0.022
Lachnoclostridium	11	1.015 (0.887–1.162)	0.828	0.911	0.009	0.709	0.034
Oscillibacter	11	1.007 (0.943–1.076)	0.833	0.911	0.264	0.554	0.020
Fusicatenibacter	16	0.991 (0.912–1.078)	0.836	0.911	0.122	0.111	0.017
Coprococcus3	9	1.009 (0.925–1.101)	0.844	0.911	0.764	0.364	0.021
Eubacterium rectale group	6	1.003 (0.901–1.117)	0.951	0.976	0.919	0.547	0.020
Akkermansia	12	0.999 (0.938–1.064)	0.976	0.976	0.659	0.133	0.016

Note: Minimum BH-FDR=0.835; no genus passed FDR 0.05; only Bifidobacterium had nominal p<0.05, indistinguishable from the global null expectation of 1.35.

Appendix S3 Classification uncertainty and OOF confusion matrices

Table S3. Eight-layer three-class OOF performance and uncertainty (subject-cluster 2000× bootstrap 95% CI; subject-level label permutation).

Layer	n	Features	balanced acc (fold mean±SD)	95% CI	macro-F1	Permutation p (permutations)
NMR	445	19	0.558±0.085	[0.477, 0.638]	0.553	0.005 (200)
Bile acids	447	12	0.482±0.034	[0.419, 0.539]	0.470	0.005 (200)
Tryptophan	274	12	0.486±0.109	[0.401, 0.564]	0.477	0.005 (200)
Genus	457	69	0.506±0.077	[0.421, 0.595]	0.508	0.005 (200)
Species	457	145	0.520±0.134	[0.434, 0.611]	0.519	0.005 (200)
NMR+Genus	440	163	0.503±0.099	[0.424, 0.581]	0.501	0.010 (100)
KEGG	459	327	0.436±0.088	[0.347, 0.520]	0.428	0.030 (100)
KO	459	5310	0.384±0.087	[0.313, 0.465]	0.384	0.095 (20)

Table S3b. IBS-SSS binary classification (genus layer, 70 features, n=250/51 subjects): balanced acc=0.500 (fold mean 0.500±0.033), 95% CI [0.420, 0.578], macro-F1=0.500, AUC=0.552, permutation p=0.458 (200 permutations)—indistinguishable from chance.

Confusion matrices (rows=true, columns=predicted; class order H/IBS-C/IBS-D):

NMR (n=445)

True\Predicted	H	IBS-C	IBS-D
H	66	42	36
IBS-C	33	86	14
IBS-D	40	33	95

Bile acids (n=447)

True\Predicted	H	IBS-C	IBS-D
H	52	55	39
IBS-C	31	86	16
IBS-D	40	54	74

Tryptophan (n=274)

True\Predicted	H	IBS-C	IBS-D
H	28	36	26
IBS-C	22	43	17
IBS-D	17	22	63

Genus (n=457)

True\Predicted	H	IBS-C	IBS-D
H	71	37	43
IBS-C	40	65	32
IBS-D	41	30	98

Species (n=457)

True\Predicted	H	IBS-C	IBS-D
H	65	45	41
IBS-C	30	76	31
IBS-D	38	33	98

NMR+Genus (n=440)

True\Predicted	H	IBS-C	IBS-D
H	64	36	43
IBS-C	34	75	20
IBS-D	56	31	81

KEGG (n=459)

True\Predicted	H	IBS-C	IBS-D
H	34	49	67
IBS-C	42	68	30
IBS-D	42	26	101

KO (n=459)

True\Predicted	H	IBS-C	IBS-D
H	38	51	61
IBS-C	48	54	38
IBS-D	49	32	88

Note: Confusion matrices are pooled results of 5-fold OOF predictions; CIs and permutation p values are based on pooled OOF predictions, and fold means are the average of 5 folds; H=Healthy. The permutation p resolution floor is $1/(B+1)$; values of 0.005/0.010 are at that floor, and the KO value of 0.095 (20 permutations) is exploratory only. Full-sample cross-layer differences in the classification task should be read as point estimates, not strong ranking claims (see §3.3, §4.7-17).

Supplementary Files

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

- [IBSSupplementaryCodeS1v1.0.zip](#)
- [IBSSupplementaryDataS1v1.0.zip](#)