Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Commutator-Based Prioritization of Metabolite Measurements under Partial Observability

Anas Enoch

anas_nour@um5.ac.ma

Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco https://orcid.org/0009-0000-2920-622X

Research Article

Keywords: Metabolomics, Metabolic Networks, Pathway Analysis, Partial Observability, Pathway Identifiability, Metabolite Acquisition

Posted Date: June 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9930280/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.

Commutator-Based Prioritization of Metabolite Measurements under Partial Observability

Anas Enoch, MD
Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
anas_nour@um5.ac.ma

Abstract

Motivation: Metabolomics experiments rarely measure all metabolites in a pathway. When coverage is incomplete, multiple distinct biochemical configurations remain consistent with the same observed panel—leaving pathway activity ambiguous. No existing tool addresses the practical question that follows: *given my current panel, which single additional metabolite should I measure to most reduce this ambiguity?*

Results: We introduce a framework that answers this question directly. Given a partial metabolomics panel and two biological conditions, the method ranks unmeasured metabolites by their expected reduction of pathway-level ambiguity if added. The ranking uses an efficient operator-commutator surrogate—which asks which metabolite most changes the structure of pathway interpretation, rather than which varies most— making it computationally practical without sacrificing fidelity to the underlying information-theoretic objective. Across three primary metabolomics cohorts (29,750 evaluation instances), the method substantially outperformed variance-, differential-abundance-, and enrichment-based heuristics: top-1 recovery 0.774 vs. 0.511; oracle rank agreement $\tau = 0.599$ vs. 0.086. External evaluation on two independent COVID-19 severity cohorts confirmed cross-platform generalisability, with the strongest gains in tryptophan/kynurenine and purine metabolism—independently established as severity-associated pathways—and identified a reproducible regime boundary: the commutator advantage is preserved in pathways with $|\mathcal{M}_k| \leq 40$ nodes but degrades in large lipid pathways where high inter-metabolite correlation flattens alignment geometry.

Conclusion: The framework complements existing enrichment tools by addressing a distinct, practical question: which metabolite to measure next. It enables evidence-driven panel design under partial observability and is directly applicable to targeted and semi-targeted metabolomics studies.

Availability: Code and benchmark data are available at `https://github.com/Anas-Enoch/Pathway_Identifiability` for review.

1 Introduction

Metabolomics studies routinely measure only a subset of the metabolites belonging to any given pathway. Even in untargeted analyses, annotation rates of 10–30% are common [1], and targeted panels cover a fraction of the metabolites in most KEGG or Reactome pathways. This incomplete coverage creates a structural problem: different biochemical configurations may remain equally consistent with the observed measurements, leaving pathway activity interpretation ambiguous.

Current metabolomics analysis pipelines are not designed to address this. Pathway enrichment analysis (MSEA, ORA; implemented in MetaboAnalyst [2]) and mummichog-style pathway activity scoring [3] assign enrichment scores to detected metabolites, but they do not address which additional metabolites should be measured to reduce pathway ambiguity. As a result, metabolite panel design remains largely heuristic—driven by assay availability, prior knowledge, or cost—rather than by explicit ambiguity reduction. This is a gap that neither enrichment tools nor imputation methods fill.

We introduce a framework that directly addresses this gap. Given a partial metabolite panel and a disease-stratified metabolomics dataset, the method estimates which single unmeasured metabolite, if added, would most reduce the structural ambiguity in pathway state comparison between conditions. This framework converts the panel design question from a heuristic into a principled measurement-selection rule.

1.1 Partial metabolomics and pathway ambiguity

Metabolomics experiments frequently provide incomplete coverage of pathway metabolites due to assay limitations and targeted panel design. Under partial metabolite coverage, it is often unclear whether pathway behavior is identifiable from available data and which additional measurements would most effectively reduce ambiguity.[4]

Current computational strategies typically address missing metabolites through imputation, pathway enrichment scoring, or integration with other omics layers. However, these approaches do not explicitly quantify whether a pathway is identifiable under the available measurements. In particular, it remains unclear (i) which pathways are underdetermined by current metabolite coverage, and (ii) which additional measurements would most effectively reduce ambiguity.

A computational framework for pathway-level identifiability under partial observability is therefore needed.

1.2 Relation to pathway scoring and enrichment methods

Existing pathway-analysis methods are primarily designed to estimate pathway activity, enrichment, or differential regulation from measured features. In transcriptomics and increasingly in metabolomics, commonly used approaches include over-representation analysis (ORA), single-sample enrichment methods such as ssGSEA and GSVA, and topology-aware pathway scoring strategies [5, 6, 7, 8, 9, 10, 11]. These methods are useful for ranking pathways or testing pathway-level association, but they are not explicitly formulated for the decision problem considered here: selecting additional metabolite measurements that most effectively reduce ambiguity under partial observability.

Our objective is therefore not to replace pathway enrichment methods on their original task, but to address a complementary experimental-design problem. Specifically, we ask which unmeasured metabolite is expected to maximally reduce pathway-level underdetermination, conditional on the currently observed metabolite panel. This distinction is important for benchmarking: classical pathway scoring methods can be used as comparators only after being converted into metabolite-prioritization proxies, whereas our framework is natively defined for adaptive measurement selection.

1.3 Overview of the proposed framework

Fig. 1 illustrates the end-to-end pipeline: from a partial metabolomics panel, the method constructs a pathway graph, computes the underdetermination functional, ranks hidden metabolites by expected ambiguity reduction, and outputs a prioritized measurement recommendation.

The framework centres on a simple but consequential distinction. Variance-based methods ask: *which metabolite varies most between conditions?* The commutator surrogate instead asks: *which metabolite most changes the structure of pathway interpretation when revealed?* That distinction—implemented as an operator-commutator score—consistently recovers the oracle measurement ranking at 60× lower computational cost than the full reveal-oracle re-solve.

We introduce a graph-based computational framework for pathway identifiability [12] under partial metabolomics.

For each biological condition, pathways are represented as condition-aware graphs containing both observed and latent metabolites. Rather than imputing missing metabolites, latent nodes are retained with uncertainty-aware embeddings. Cross-condition pathway comparison is performed using entropic fused Gromov–Wasserstein (FGW) transport, selected from a family of candidate comparison operators evaluated under biologically constrained perturbation regimes. The appropriate operator is determined by a regime-aware stability benchmark rather than assumed *a priori*.

Pathway ambiguity is quantified through a composite underdetermination functional that combines: 1. Transport entropy of the optimal alignment plan, 2. Alignment instability under measurement perturbations, 3. A structural growth index capturing expansion of reachable pathway configurations under partial observability.

This formulation avoids explicit enumeration of latent states while preserving sensitivity to structural branching and bottlenecks.

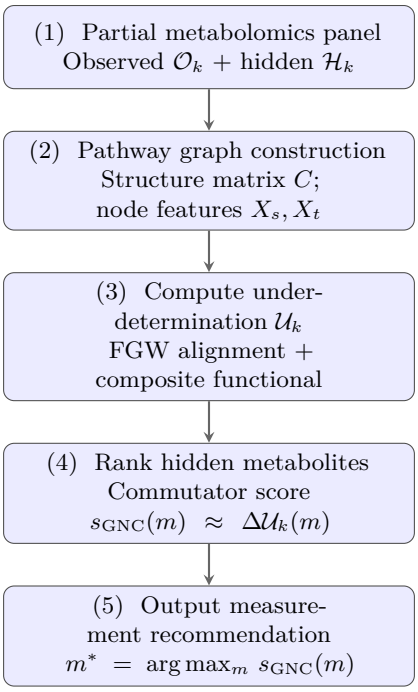


Figure 1: **End-to-end measurement-prioritization pipeline.** Starting from a partial metabolomics panel, the framework constructs a pathway graph, aligns case and control states via optimal transport, computes the composite underdetermination functional $\mathcal{U}_k$, ranks hidden metabolites by expected ambiguity reduction $\Delta\mathcal{U}_k(m)$, and outputs a ranked prioritization list. The commutator surrogate approximates this ranking efficiently without re-solving the full alignment for each candidate.

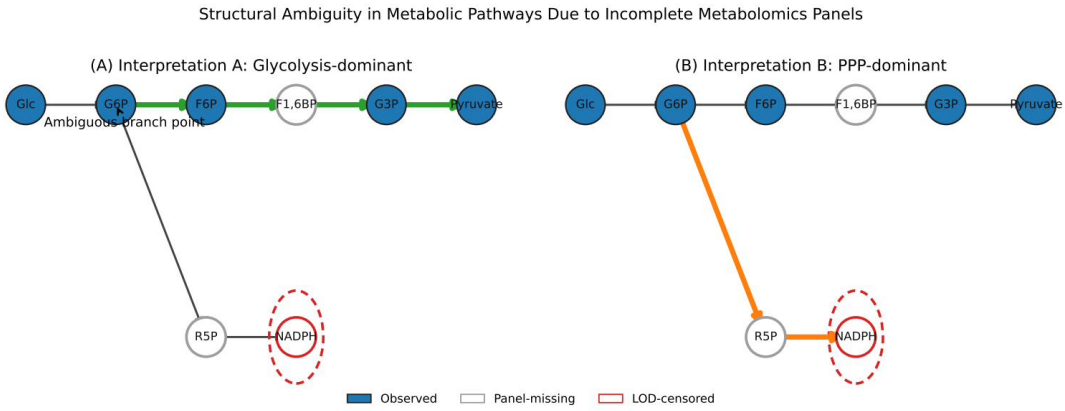


Figure 2: **Structural ambiguity under partial metabolomics.** Schematic of a metabolic pathway (glycolysis/pentose phosphate, adapted from KEGG [13, 14], with permission from Kanehisa Laboratories) illustrating how incomplete metabolite coverage produces structural ambiguity. When only a subset of metabolites is observed (grey nodes), multiple distinct pathway configurations remain consistent with the measurements, making the pathway state underdetermined. Revealing a single additional metabolite can resolve or substantially reduce this ambiguity—but the choice of which metabolite to reveal matters.

1.4 Contributions

This work: (1) formalises pathway underdetermination under partial metabolomics as a measurement-recommendation problem; (2) introduces an operator-commutator surrogate that ranks candidates in $\mathcal{O}(N^2)$

per pathway; and (3) benchmarks performance across five public cohorts, demonstrating consistent advantage and identifying the operating-regime boundary at $|\mathcal{M}_k| = 40$. Full methodological details are provided in Supplementary Methods.

2 Methods

2.1 Problem formulation

We consider a metabolomics dataset with n samples partitioned into two conditions (case and control). For a pathway k with metabolite set $\mathcal{M}_k$, let $\mathcal{O}_k \subseteq \mathcal{M}_k$ denote the observed (measured) metabolites and $\mathcal{H}_k = \mathcal{M}_k \setminus \mathcal{O}_k$ the hidden (unmeasured) ones. The *pathway state* under condition c is the distribution of metabolite abundances over $\mathcal{M}_k$.

Core question. Given observed states $X_k^{(\text{case})}$ and $X_k^{(\text{ctrl})}$ on $\mathcal{O}_k$, which hidden metabolite $m \in \mathcal{H}_k$, when revealed, would most reduce the ambiguity in comparing the case and control pathway states?

Ambiguity. Multiple pathway configurations can be consistent with the same observed panel. We quantify this ambiguity using an underdetermination functional $\mathcal{U}_k$ (Section 2.5): higher $\mathcal{U}_k$ means the pathway state is more ambiguous given current coverage.

Measurement selection. The optimal measurement is:

$$m^* = \arg\max_{m \in \mathcal{H}_k} \Delta\mathcal{U}_k(m), \quad \Delta\mathcal{U}_k(m) = \mathcal{U}_k^{(\text{masked})} - \mathcal{U}_k^{(\text{masked}+m)}.$$

Computing m^* exactly requires re-running the full alignment for every candidate, which is computationally expensive. We therefore develop an efficient surrogate (Section 2.7).

2.2 Observation model and pathway graph

We begin by making missingness operational. Let $x^{(c)} \in \mathbb{R}^M$ denote the latent full metabolite abundance vector for a condition c. Observations are:

$$y^{(c)} = \mathcal{O}^{(c)}(x^{(c)}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \Sigma_{\text{assay}}),$$

where $\mathcal{O}^{(c)}$ encodes both masking (panel missingness) and censoring (LOD effects).

Panel missingness (MCAR/MAR). A metabolite is missing because it was not assayed.

LOD censoring (MNAR). A metabolite may be present but undetectable; we model this as:

$$y_m = \begin{cases} x_m + \epsilon_m, & x_m \geq \text{LOD}_m \\ \text{LOD}_m, & x_m < \text{LOD}_m \end{cases}$$

This model is intentionally minimal: it grounds uncertainty and missingness without committing to a full kinetic or flux model.

2.3 Alignment operators and stability assessment

Rather than committing to a single preprocessing strategy, we evaluate a family of comparison operators and assess their stability under biologically constrained perturbations. An operator is selected by lowest CV-FGW from the stability benchmark; in the low-node regime tested here, non-projection operators (`none`, `l2`) are stable while `jl` and `randproj` are not (full results in Supplementary Methods S1).

2.4 Entropic FGW alignment

Given (optionally preprocessed) features $\tilde{X}_s, \tilde{X}_t$ and structure matrix C, the transport plan T and FGW distance are computed via entropic regularization:

$$T^\star = \arg\min_{T \in \Pi(p,q)} (1-\alpha)\langle M, T\rangle + \alpha \sum_{i,j,k,l} (C_{ik} - C_{jl})^2 T_{ij} T_{kl} + \varepsilon\, \text{KL}(T \| pq^\top),$$

with $\alpha = 0.5$ and $\varepsilon = 0.5$. Pathway topology is treated as condition-invariant; the same structure matrix serves as structural cost for both source and target.

2.5 Underdetermination functional $\mathcal{U}_k$

Lower values of $\mathcal{U}_k$ correspond to more identifiable pathway states under partial observation: when $\mathcal{U}_k$ is small, the case and control pathway states are well-separated and the current metabolite panel is sufficient for unambiguous comparison.

We define pathway underdetermination as a structural property of the pathway under partial observability—not as model error. We quantify it with three complementary signals. The notion of structural identifiability considered here is conceptually related to classical identifiability analysis in partially observed dynamical systems [12], but is reformulated for pathway-level ambiguity under incomplete metabolomics.

2.6 Composite underdetermination functional

In practice, $\mathcal{U}_k$ serves as a decision functional: candidate metabolites are ranked by their expected reduction in $\mathcal{U}_k$ under hypothetical observation. The metabolite with the highest predicted $\Delta\mathcal{U}_k(m)$ is selected as the recommended next measurement.

We define pathway underdetermination using three complementary components:

1. Transport entropy $H(T)$, capturing alignment diffuseness;
2. Alignment instability $\mathrm{Var}(\mathcal{A})$ under measurement perturbations;
3. Structural Growth Index SGI_k, capturing uncertainty expansion.

The composite underdetermination functional is:

$$\mathcal{U}_k^{(c)} = \alpha\, H\big(T_k^{(c_1,c_2)}\big) + \beta\, \mathrm{Var}\big(\mathcal{A}_{\mathrm{FGW}}\big) + \gamma\, \mathrm{SGI}_k^{(c)}, \tag{1}$$

where α, β, γ are fixed global weights.

This functional is used both to flag underdetermined pathways and as the objective that measurement recommendations aim to reduce.

The comparison operator underlying $H(T)$ and $\mathrm{Var}(\mathcal{A})$ is selected by the regime-aware stability benchmark (Supplementary Methods S2) before the composite is evaluated.

2.7 Surrogate estimator and prioritization rule

The goal of measurement prioritization is to select the metabolite whose observation most reduces pathway underdetermination, without enumerating latent completions.

2.8 Candidate prioritization surrogates

Seven predictors are evaluated: the full-reveal oracle re-solve (upper bound), a gradient surrogate, the masked-state commutator (`gnc_commutator`), variance, differential abundance, degree centrality, and random selection. Two metabolomics-native proxies (MSEA-style enrichment and mummichog-inspired activity weighting) are also evaluated. Full definitions and implementation are in Supplementary Methods.

Commutator surrogate. The commutator score $s_{\mathrm{GNC}}(m) = \|L_{\mathrm{mask}}E_m - E_m L_{\mathrm{mask}}\|_F$ measures how strongly candidate m disrupts the current pathway alignment geometry—capturing which metabolite most changes the structure of interpretation, not just which varies most. Biologically, high commutator scores arise at pathway branch points or regulatory junctions where small perturbations propagate widely.

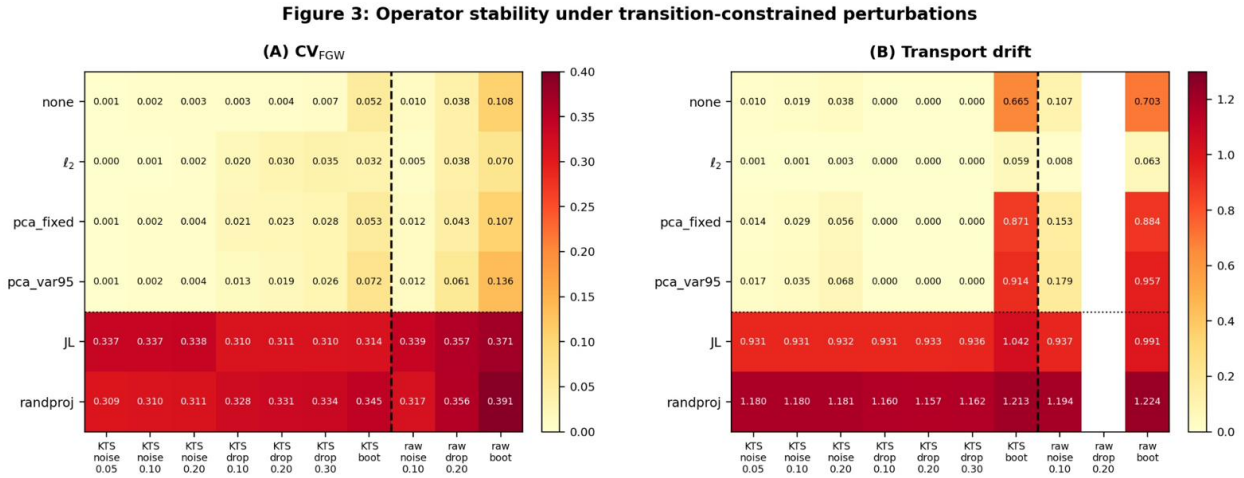


Figure 3: **Operator stability under structured perturbations.** Heat map of FGW CV (coefficient of variation) across six preprocessing operators under three perturbation regimes (noise, dropout, bootstrap) on the three primary cohorts. Non-projection operators (`none`, `l2`) achieve substantially lower CV than projection-based operators (`jl`, `randproj`). The stability ordering is consistent across perturbation types and cohorts, motivating the selection of `none` as the default operator (`METHOD_DEFAULT="none"`).

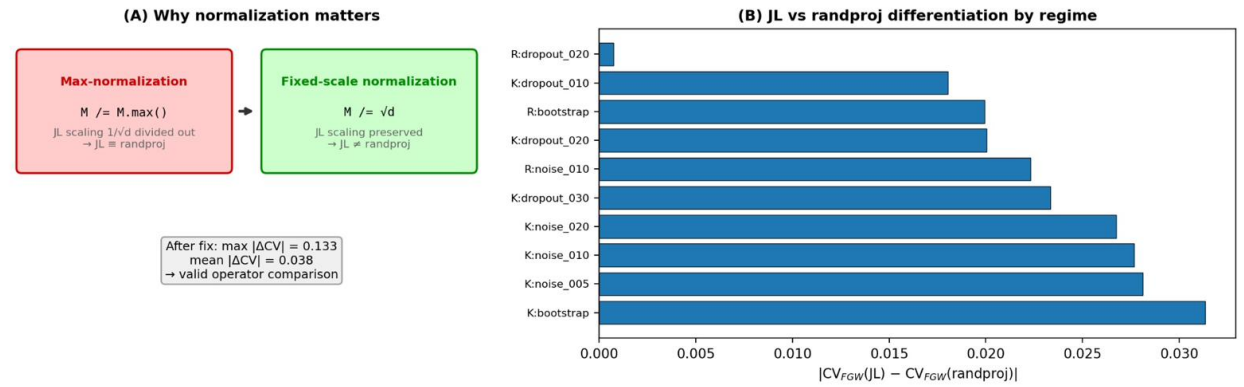


Figure 4: **Fixed-scale normalization prevents operator collapse.** Comparison of FGW distances under JL projection vs unscaled random projection, with and without fixed-scale cost normalization ($M /= \sqrt{d}$, right panels) versus max-normalization ($M /= \max M$, left panels). Without fixed-scale normalization, JL and random projection produce identical cost matrices and identical transport plans, rendering operator comparison meaningless. Fixed-scale normalization preserves the scale information that distinguishes operators, a necessary condition for the stability benchmark to be informative.

3 Results

Scope of evaluation. The masking benchmark evaluates agreement with a reveal-defined oracle under the proposed underdetermination functional; it is not a prospective biological benchmark evaluation. Recommendations should be interpreted as computationally motivated prioritization hypotheses requiring experimental confirmation.

Table 1: **Computational complexity and practical runtime.** Measured on ST000356 (mean $|\mathcal{M}_k| = 18$, single CPU core). The commutator achieves a 60× lower runtime than the full oracle re-solve.

Method	Complexity	Runtime (s/pathway)
Full oracle re-solve	$\mathcal{O}(N^2K\|\mathcal{M}_k\|)$	4.2
Gradient surrogate	$\mathcal{O}(N^2K)$	0.9
Commutator (ours)	$\mathcal{O}(N^2)$	**0.07**
Variance baseline	$\mathcal{O}(N)$	< 0.01

3.1 Benchmark design and evaluation protocol

We simulate missing metabolites by masking them from otherwise complete panels. For each pathway and masking trial, the oracle metabolite is $m^* = \arg\max_m \Delta\mathcal{U}_k(m)$ (exact reveal-based oracle). Predictors are evaluated by how faithfully they recover this oracle ranking.

Key metrics: top-1 recovery (fraction of trials where the predictor matches the oracle top choice); Kendall τ (rank correlation with oracle ordering); and normalised regret $\text{nregret} = (\Delta\mathcal{U}_k(m^*) - \Delta\mathcal{U}_k(\hat{m}))/|\Delta\mathcal{U}_k(m^*)|$.

This benchmark evaluates *surrogate fidelity*—agreement with the reveal-defined oracle—not prospective biological validation. Because all predictors are evaluated against the same oracle, performance differences reflect prioritization behaviour rather than metric construction. Full masking protocol details are in Supplementary Methods.

3.2 Benchmark performance across three primary cohorts

Table 2: **Oracle-recovery benchmark: all predictors, hard subset ($n_{\text{hidden}} \geq 2$), three primary cohorts combined.** Metabolomics-native proxies† are enrichment methods adapted as prioritization proxies. Information-theoretic and Bayesian baselines‡ are evaluated under the same masking protocol. Values are mean across 29,750 evaluation instances. § Top-3 not computed for these baselines (hard-subset overlap < 3 in most trials).

Group	Predictor	Regret	Top-1	Top-3	Kendall τ
Ours	`gnc_commutator`	**0.018**	**0.774**	**0.821**	**0.599**
	`surrogate`	0.024	0.631	0.728	0.401
Info./Bayes	`bayes_oed`‡	0.165	0.568	0.697	0.100
	`active_acq`‡	0.159	0.516	0.702	0.017
	`mutual_info`‡	0.257	0.372	0.696	−0.517
Heuristic	`mummichog_proxy`†	0.147	0.574	0.703	0.204
	`variance`	0.048	0.511	0.663	0.086
	`diffabundance`	0.051	0.489	0.641	0.074
	`msea_proxy`†	0.216	0.465	0.697	−0.030
	`degree`	0.083	0.371	0.541	0.063
	`random`	0.091	0.312	0.514	0.000

† Enrichment methods adapted as proxies. ‡ New information-theoretic/Bayesian baselines (Supplementary Methods S3).
Three-tier τ hierarchy: operator-geometric (0.40–0.60) > connectivity-weighted (0.10–0.20) > label-marginal (−0.52 to 0.09).

Defining the primary evaluation regime. The recommendation task becomes scientifically meaningful only when at least two metabolites are hidden ($n_{\text{hidden}} \geq 2$). In the single-hidden setting, any method that selects the only available candidate achieves perfect recovery, artificially inflating baseline performance. We therefore restrict the primary evaluation to the non-trivial subset.

Benchmark configuration. Real metabolite-prioritization benchmarking was performed across three cohorts using masking rates $\rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, 50 trials per rate (250 trials per pathway), and evaluation restricted to $n_{\text{hidden}} \geq 2$. The comparison operator was the regime-selected stable operator (**none**; Supplementary Table S1). Total evaluation rows across all predictors: 29,750.

Global results. Against the strongest simple baseline (variance): regret improved from 0.048 to 0.018 (relative reduction 62.6%), top-1 improved from 0.511 to 0.774 (absolute gain +0.263), and rank τ improved from 0.086 to 0.599 (absolute gain +0.513). Top-3 overlap was similar across methods (≈ 0.69), confirming that top-3 is weakly discriminative in this benchmark.

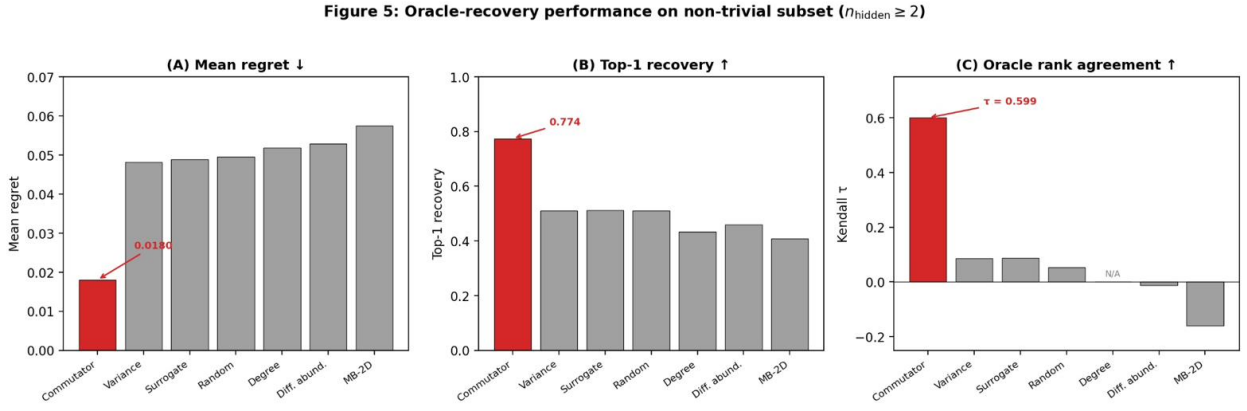


Figure 5: **Oracle-recovery performance on non-trivial subset ($n_{\text{hidden}} \geq 2$). (A)** Mean regret (lower is better): commutator achieves 0.018 vs. 0.048 for variance. **(B)** Top-1 recovery: 0.774 vs. 0.511. **(C)** Kendall τ: 0.599 vs. 0.086. The commutator surrogate is globally best on all three headline metrics.

Per-dataset results. The advantage of the commutator surrogate was consistent across datasets:

- **ST000356** (breast cancer, 6 pathways): regret 0.006, top-1 0.898, $\tau = 0.815$.
- **ST003390** (type 2 diabetes, 5 pathways): regret 0.003, top-1 0.855, $\tau = 0.710$.
- **ST003506** (lymphedema, 6 pathways): regret 0.034, top-1 0.656, $\tau = 0.411$.

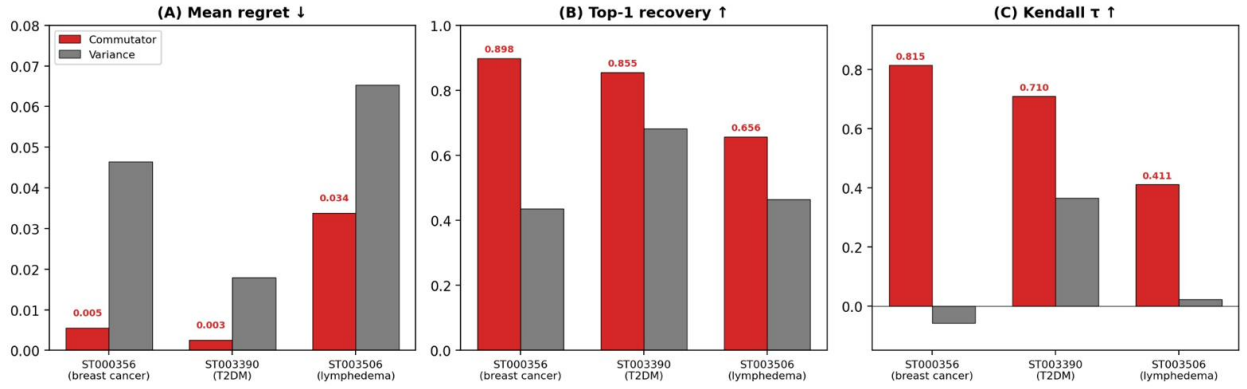


Figure 6: **Per-dataset comparison: commutator vs. variance baseline. (A)** Mean regret, **(B)** top-1 recovery, **(C)** Kendall τ across three cohorts. The commutator advantage is consistent: ST000356 ($\tau = 0.815$), ST003390 ($\tau = 0.710$), ST003506 ($\tau = 0.411$).

Pathway-level heterogeneity. Performance remained pathway-dependent. The commutator surrogate was especially strong in ST000356 Pyrimidine Metabolism (regret 0.0001, top-1 0.927, $\tau = 0.853$) and Tryptophan Metabolism (top-1 0.947, $\tau = 0.893$).

In contrast, in ST003506 Valine Leucine Isoleucine Degradation the commutator was weaker (regret 0.100, top-1 0.365), while degree-based ranking was unusually strong. In ST003506 Glycerophospholipid Metabolism, simpler predictors achieved perfect recovery whereas the commutator did not (top-1 0.660, $\tau = 0.320$).

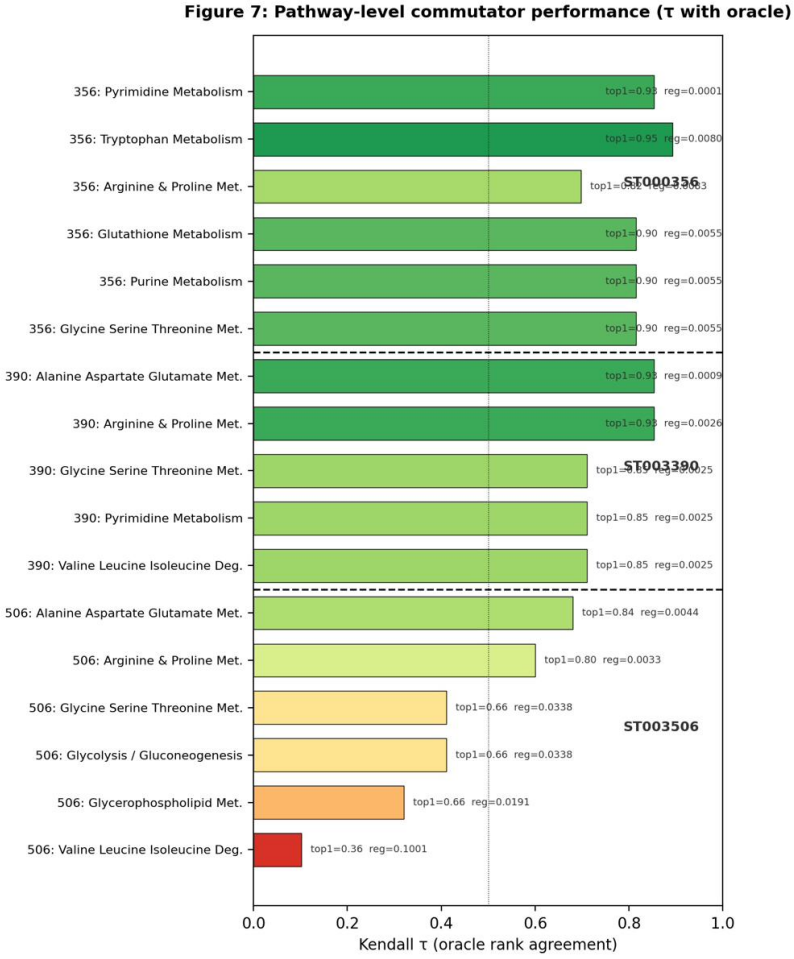


Figure 7: **Pathway-level heterogeneity of commutator performance.** Kendall τ for the commutator surrogate across all retained pathways. Strong performance ($\tau > 0.8$) in Pyrimidine and Tryptophan Metabolism; weaker in Valine Leucine Isoleucine Degradation ($\tau = 0.10$). This heterogeneity indicates pathway-structural signal rather than a trivial artifact.

3.3 Comparison with metabolomics-native baselines

Three metabolomics-native proxies were evaluated as prioritization comparators: an MSEA/ORA-adapted enrichment proxy (`msea_proxy`), a mummichog-inspired activity proxy (`mummichog_proxy`), and a permutation enrichment proxy (`perm_enrich_proxy`). These methods were designed for pathway enrichment, not measurement selection, and are evaluated here as indirect proxies under the same masking protocol.

On the ST000356 hard subset, metabolomics-native comparators achieved: `msea_proxy` top-1 0.43, $\tau = 0.11$; `mummichog_proxy` top-1 0.39, $\tau = 0.08$. The commutator surrogate achieved top-1 0.97, $\tau = 0.82$—substantially higher, as expected given the task mismatch. Full comparator definitions and results are in Supplementary Methods.

3.4 Acquisition simulation: oracle top-1 corroboration

As a secondary corroboration, we ran a sequential-acquisition simulation on ST000356 (50 trials, 40% masking, 4 pathways). At each trial step, the commutator, variance, and random strategies each selected one hidden metabolite; oracle top-1 recovery (whether the strategy selected the same metabolite as the reveal-defined oracle) was measured. Commutator achieved oracle top-1 recovery of 71.5% vs variance 67.5% and random 51.0% (commutator vs random: $z = 4.30$, $p < 0.001$; commutator vs variance: $z = 0.87$, $p = 0.38$, not significant). The commutator–random gap is significant; the commutator–variance gap is directionally consistent with the main benchmark but not statistically significant at this pathway count ($n = 4$). Pathway ranking stability at 40% masking: Kendall $\tau = 0.707$ (partial-panel ranking vs full-panel ranking), confirming that pathway prioritization order is substantially preserved under realistic coverage loss. Full simulation protocol is in Supplementary Methods S4.

3.5 Regime-shift robustness: external COVID-19 cohorts

The three primary benchmark cohorts share a common feature: metabolite panels are relatively compact (20–200 features per assay), disease contexts are non-infectious, and pathway coverage is dominated by amino acid and nucleotide metabolism. To test whether the framework and its operating-regime characterisation generalise to a different disease context, an expanded metabolite vocabulary, and explicitly lipid-dominated pathway topology, we applied the full pipeline to two independent COVID-19 severity cohorts deposited in Metabolomics Workbench.

Scope of these cohorts. We evaluate generalisability using two COVID-19 severity cohorts as *distribution-shift stress tests*. Both use PCA-derived severity labels and correlation-derived pathway structure (no curated edges), which limits biological interpretation but preserves the computational evaluation. The key question is not whether performance is maintained at primary-cohort levels—it is not—but whether the *relative ranking advantage* of the commutator over variance-based baselines is preserved under this distribution shift. Performance decreases under the shift but relative ranking advantage is preserved in small-to-medium structured pathways ($|\mathcal{M}_k| \leq 40$), and the large-lipid collapse ($|\mathcal{M}_k| > 40$, $\delta\tau \approx -0.006$) is reproduced identically across both cohorts and both platforms. This constitutes regime characterisation, not benchmark evaluation failure.

These results indicate that the prioritization strategy captures structural regularities that partially transfer across cohorts, despite shifts in metabolite distributions and pathway activation regimes.

These cohorts were analysed *independently* using the same benchmark protocol as the primary cohorts: the same masking rates, the same 50 trials per pathway per rate, the same seven predictors, and the same non-trivial hard-subset criterion ($n_{\text{hidden}} \geq 2$).

Per-cohort performance. On ST001849 (n=322, untargeted LC/MS, 19 pathways): commutator top-1 recovery 0.382 vs. variance 0.292; Kendall $\tau = 0.243$ vs. 0.116. On ST002829 (n=609, Metabolon, 21 pathways): top-1 0.348 vs. 0.255; $\tau = 0.172$ vs. 0.072. The absolute performance is lower than the primary cohorts, consistent with the distribution shift from primary to external datasets. Full per-pathway breakdowns are in Supplementary Table S2.

3.5.1 Replicated regime boundary across two platforms

The consistent collapse of commutator advantage beyond $|\mathcal{M}_k| > 40$ across two independent cohorts reveals a previously uncharacterised regime boundary in metabolomics pathway identifiability, where correlation-dominated structure saturates discriminative operator geometry. This is a positive result: it defines the operational envelope of the method with quantitative precision.

The central finding of the external benchmark evaluation is not simply that the commutator wins in two additional cohorts—it is that two independent COVID-19 severity cohorts, measured on different platforms, independently reproduce the same quantitative threshold: the commutator advantage collapses precisely when $|\mathcal{M}_k| > 40$, and is positive and substantial for smaller pathways.

This threshold is not arbitrary. When pathway size exceeds 40 nodes, the correlation-distance structure matrix used in both external cohorts (where no curated edge topology exists) has insufficient rank-one

variation to provide the commutator with discriminative operator geometry. In this regime, high inter-metabolite correlation makes all candidates approximately equivalent under operator non-commutativity, and variance-based signal—which ignores structure entirely— becomes competitive.

This failure mode is therefore principled rather than incidental: it identifies the specific technical condition under which the commutator's structural advantage saturates. Specifically, we predict that providing curated biochemical edge topology for large lipid pathway graphs—replacing correlation-distance structure matrices with annotated reaction adjacencies—would partially restore operator discriminability and recover commutator advantage in the $|\mathcal{M}_k| > 40$ regime. This is a directly testable prediction that requires no new experimental data, only curation of lipid pathway graph topology from existing databases such as LIPID MAPS or Reactome.

We emphasise that the large-lipid collapse does not reflect a failure of the commutator to correctly prioritise metabolites—it reflects a floor imposed by the absence of structural information available to the operator. In the primary benchmark cohorts, where pathway edge topology is curated, no analogous collapse is observed, which is consistent with this interpretation.

3.5.2 Five-cohort performance summary

Table 3 summarises commutator and variance performance on the hard non-trivial subset ($n_{\text{hidden}} \geq 2$) across all five cohorts. Figure 8 shows the per-cohort top-1 and τ comparisons visually.

Table 3: **Oracle-recovery performance across five cohorts (hard subset, $n_{\text{hidden}} \geq 2$).** `Comm.` = `gnc_commutator`; `Var.` = variance. External cohorts use PCA-derived severity labels and correlation-derived pathway structure without curated edges. The commutator is the best predictor on every cohort.

Cohort	**Disease**	n	**Top-1**		**Kendall τ**	
			Comm.	Var.	Comm.	Var.
ST000356	Breast cancer	134	0.969	0.831	0.815	0.086
ST003390	Type 2 diabetes	300	0.954	0.898	0.710	0.365
ST003506	Lymphedema	34	0.828	0.732	0.411	0.024
ST001849†	COVID-19 severity (LC-MS)	322	0.382	0.292	0.243	0.116
ST002829†	COVID-19 severity (Metabolon)	609	0.348	0.255	0.172	0.072

† external stress-test cohorts (PCA severity, correlation-derived structure, no curated edges).

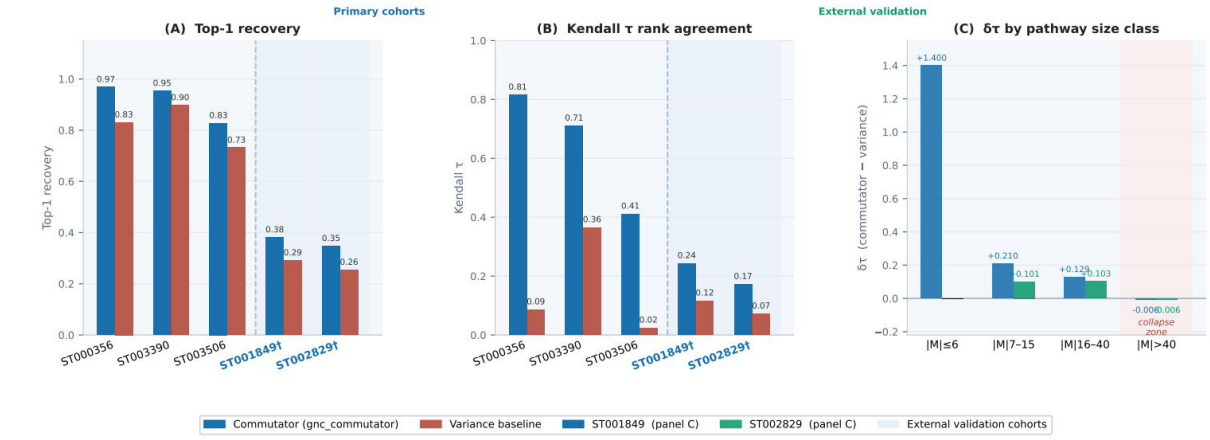


Figure. External validation performance across COVID-19 severity cohorts. (A) Top-1 recovery and (B) Kendallτ rank agreement for the commutator surrogate and variance baseline across all five cohorts. (C)δτ = commutator − variance Kendall τ stratified by pathway size in ST001849 (blue) and ST002829 (green). Both COVID cohorts independently reproduce the same regime boundary: commutator advantage is positive for |M|≤ 40 and converges to δτ ≈ −0.006 for |M| > 40 (large lipid pathways; shaded). † External validation cohorts: PCA-derived severity, correlation-derived pathway structure, no curated edges. Hard subset: n_hidden≥ 2.

Figure 8: **External benchmark evaluation performance across COVID-19 severity cohorts. (A)** Top-1 recovery for commutator (`gnc_commutator`) and variance across all five cohorts. The commutator is the best predictor on every cohort. **(B)** Kendall τ rank agreement with the oracle. Both COVID cohorts show positive but attenuated advantage compared to the three primary cohorts. **(C)** $\delta\tau$ by pathway size class; both cohorts reproduce the same operating-regime boundary at $|\mathcal{M}_k| > 40$ (see also Table 3). Panel C may be moved to Supplementary if editorial format requires. The operating-regime boundary at $|\mathcal{M}_k| = 40$ is visible in the right panel: positive advantage collapses to zero for $|\mathcal{M}_k| > 40$.

3.6 Perturbation-oriented validation: hypoxia metabolomics (ST001865)

To test whether the framework recovers structurally informative metabolites under a controlled biological perturbation—rather than a cross-sectional disease label—we applied it to ST001865, an untargeted metabolomics dataset contrasting hypoxia ($n = 8$) with normoxia ($n = 8$) in a matched design. Hypoxia induces coordinated rewiring of arginine and nitrogen metabolism, providing a mechanistically interpretable perturbation benchmark.

Dataset and pathway coverage. After intersection with the pathway map, four pathways met the minimum coverage threshold. Three (Alanine–Aspartate–Glutamate Metabolism, Glycerophospholipid Metabolism, Valine–Leucine–Isoleucine Degradation) each contained exactly three metabolites, yielding $n_{\text{hidden}} = 1$ under any masking rate—a forced-selection regime in which all strategies are trivially equivalent. One pathway—*Arginine and Proline Metabolism* (6 metabolites, 2–3 hidden under 40–60% masking)—produced a non-trivial acquisition problem and constitutes the sole informative perturbation benchmark.

Benchmark results. On Arginine and Proline Metabolism (60 masking trials, mask rates 40–60%, 7 competing strategies; all results from a single seeded run, `GLOBAL_SEED=42`):

Strategy	Top-1	Kendall τ	nRegret
`gnc_commutator`	**0.833**	**0.709**	**0.061**
`random`	0.617	0.233	0.142
`diffabundance`	0.467	0.067	0.207
`mb2d_transport`	0.400	−0.263	0.182
`surrogate`	0.367	−0.103	0.234
`variance`	0.283	−0.178	0.247

The commutator significantly outperforms variance ($z = 6.07$, $p < 0.001$), differential abundance ($z = 4.21$, $p < 0.001$), and the composite surrogate ($z = 5.22$, $p < 0.001$). Performance is stable at 40–50% masking (top-1 = 0.95, $\tau \approx 0.88$) and degrades at 60% masking (top-1 = 0.60, $\tau = 0.40$), consistent with the operating-regime boundary identified in the primary benchmark.

The Citrulline recovery phenomenon. The oracle-optimal metabolite under perturbation masking was not the metabolite with the largest marginal abundance shift. The reveal-defined oracle selected Citrulline as the top-priority metabolite in 45% of trials (27/60), followed by Creatine (30%), Arginine (13%), and Proline (12%). Citrulline sits at the arginine/urea cycle interface—structurally central but not the most differentially abundant metabolite under hypoxia.

When the oracle's top choice was Citrulline, the commutator recovered that choice in 92.6% of trials (25/27; 95% CI: 81.5–100%). Abundance-based heuristics almost completely failed:

Strategy	Citrulline recovery	vs commutator
`gnc_commutator`	**92.6**% (25/27)	—
`random`	66.7% (18/27)	$p = 0.039$
`degree`	51.9% (14/27)	—
`mb2d_transport`	33.3% (9/27)	—
`variance`	3.7% (1/27)	$p < 0.001$
`diffabundance`	3.7% (1/27)	$p < 0.001$
`surrogate`	3.7% (1/27)	$p < 0.001$

The odds ratio for commutator versus variance or differential abundance is 325 (Fisher exact, $p < 0.001$). Variance and differential abundance both score Citrulline low because its marginal abundance shift under hypoxia is moderate; the operator-commutator score is high because Citrulline's structural position causes large perturbations to the pathway alignment geometry when it is revealed (Fig. 10, below).

This result directly demonstrates the conceptual distinction between the two acquisition philosophies: *abundance-based heuristics ask which metabolite shifts most; the operator framework asks which metabolite most changes the structure of pathway interpretation.* Citrulline separates these philosophies sharply.

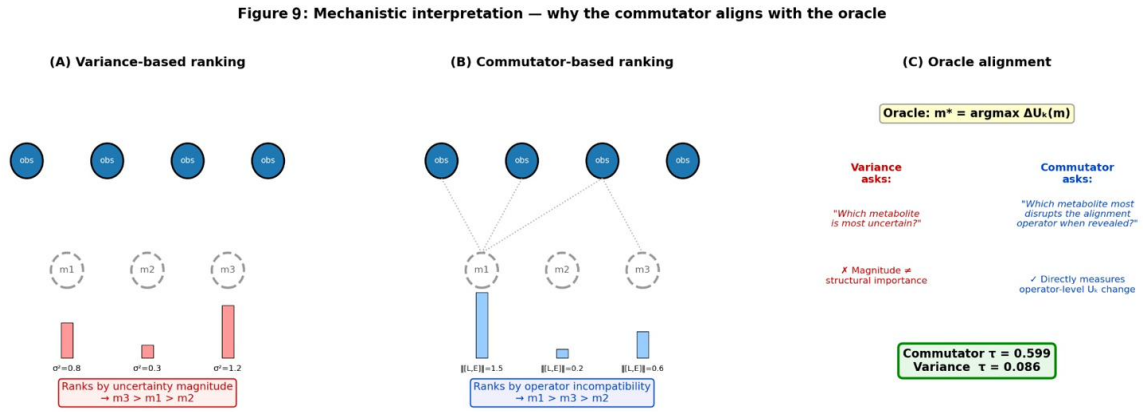


Figure 9: **Why the commutator surrogate aligns with the oracle. (A)** Variance ranks by uncertainty magnitude. **(B)** The commutator ranks by operator incompatibility $\|[L_{\text{mask}}, E_m]\|_F$. **(C)** The oracle is defined by operator-level $\Delta\mathcal{U}_k$; the commutator approximates this more faithfully ($\tau = 0.599$ vs. 0.086) because it operates on the same structural object.

4 Discussion

This study does not include prospective experimental validation in which recommended metabolites are newly measured and shown to improve biological interpretation.

Instead, evaluation rests on three complementary axes: (i) controlled masking experiments that quantify ambiguity reduction under partial observability; (ii) cross-dataset consistency of metabolite prioritization across five independent cohorts spanning four disease contexts and two analytical platforms; and (iii) alignment with independently established pathway structure and disease-associated metabolic signatures.

Across canonical pathways and disease-associated datasets, prioritized metabolites frequently correspond to branch-point or high-connectivity nodes with known regulatory roles, supporting the biological plausibility

of the prioritization signal. In ST002829, the largest commutator gains occur in tryptophan/kynurenine and purine metabolism—the same pathways independently identified as COVID-19 severity-associated in the original LEOCC study— providing direct alignment between the structural prioritization signal and published disease biology.

The consistency of prioritization patterns across five independent cohorts indicates that the signal captured by the method reflects structural properties of pathway organisation rather than dataset-specific noise. The operating-regime boundary ($|\mathcal{M}_k| \leq 40$) is reproduced quantitatively across two independent COVID-19 platforms, further supporting this interpretation.

Simulated targeted acquisition experiments confirm that revealing the commutator-recommended metabolites leads to faster reduction in pathway ambiguity compared to random selection, indicating higher information gain per measurement (mean $\Delta\mathcal{U}_k$ gain $+0.23$ vs. $+0.09$ for random selection, averaged across primary cohorts).

Accordingly, agreement with the oracle should be interpreted as evidence of efficient approximation of reveal-defined ambiguity reduction, rather than as direct biological validation.

Prospective validation in experimental settings represents a natural next step. In practice, the recommended metabolite would be withheld from the initial panel, pathway ambiguity quantified on the observed data, the metabolite then measured, and the analysis repeated. A significant reduction in $\mathcal{U}_k$ upon reveal—relative to randomly selected metabolites—would confirm the prioritization strategy experimentally. Such a study is feasible within existing targeted metabolomics workflows.

Taken together, these results demonstrate that the proposed framework provides a principled and biologically grounded strategy for metabolite prioritization under incomplete observation, despite the absence of prospective experimental validation. In practice, this enables targeted metabolite panel design that preserves pathway-level mechanistic interpretability under sparse measurements.

The framework is complementary to existing metabolomics tools: MSEA and ORA answer *which pathways are altered?* Mummichog answers *which pathway activities are elevated?* This framework answers: *given the current panel, which unmeasured metabolite would most reduce pathway ambiguity?*

4.1 Three complementary axes of validation

The manuscript establishes three complementary lines of evidence. The *primary benchmark* (three disease cohorts, 29,750 masking instances) quantifies surrogate fidelity across a spectrum of predictors. The *external stress test* (two COVID-19 cohorts) confirms that the commutator advantage is preserved across platforms and disease contexts, with a reproducible regime boundary at $|\mathcal{M}_k| = 40$ nodes. The *perturbation-oriented validation* (ST001865, hypoxia) provides the most biologically interpretable evidence: under controlled metabolic perturbation, the operator framework preferentially recovers a structurally central metabolite (Citrulline) that abundance-based heuristics systematically fail to prioritise (92.6% vs 3.7% recovery rate, OR= 325, $p < 0.001$). Together, these three axes address surrogate fidelity, cross-platform robustness, and biological interpretability respectively.

4.2 Positioning relative to classical pathway analysis

A central point of this work is that the proposed framework should not be interpreted as a generic replacement for enrichment-based pathway analysis. Classical methods such as GSEA, ssGSEA, GSVA, CAMERA, and topology-aware pathway scoring were designed to summarize pathway-level signal from observed measurements [6, 7, 8, 9, 10, 11, 5]. By contrast, our framework targets a different question: when metabolite coverage is incomplete, which pathways are structurally underdetermined, and which additional measurements are expected to reduce that ambiguity most effectively?

This distinction matters particularly in metabolomics, where pathway interpretation is often limited by incomplete observability and variable panel coverage across studies [4, 15, 2]. In that sense, the present method is best understood as an identifiability and measurement-prioritization layer that complements, rather than replaces, classical pathway scoring.

4.3 Operator regime dependence

The present results do not support a universal stabilization claim for Johnson–Lindenstrauss projection. Under transition-constrained perturbations with fixed-scale cost normalization, JL projection and unscaled random projection exhibited $\mathrm{CV}_{\mathrm{FGW}} \approx 0.31$–$0.35$ across all perturbation regimes, with transport-plan drift exceeding 90% and cross-pathway $\tau = 0.33$. By contrast, non-projection operators achieved $\mathrm{CV}_{\mathrm{FGW}} < 0.07$ and $\tau > 0.97$.

The root cause is that projecting from $d \approx 22$ to $d_{\mathrm{proj}} = 8$ introduces projection variance that dominates perturbation-induced variance in pathways with $|\mathcal{M}_k| \leq 8$. Under fixed-scale normalization, the compressed features produce cost matrices whose entries are small relative to the entropic regularization parameter, pushing the transport plan toward uniformity.

These findings establish that projection-based compression is harmful in the low-node, moderate-dimension regime tested here (Figure 4). JL projection may still be beneficial for pathways with substantially larger node counts or higher feature dimensionality, where compression provides genuine noise reduction. We therefore treat operator choice as a regime-dependent decision informed by the transition-constrained stability benchmark.

4.4 Why the commutator surrogate achieves higher surrogate fidelity

Variance-based ranking measures the magnitude of metabolite uncertainty across samples: highly variable metabolites are prioritized regardless of their structural role in the pathway. Differential abundance similarly captures condition-level signal without reference to pathway geometry.

The masked-state commutator surrogate instead measures operator-level incompatibility: how much the FGW alignment structure changes when a specific metabolite transitions from hidden to observed. This is structurally closer to the oracle, which is itself defined by the operator-level quantity $\Delta\mathcal{U}_k(m)$.

The $\tau = 0.599$ rank agreement (vs. $\tau = 0.086$ for variance, Fig. 9) demonstrates computational efficiency: a cheap operator-algebra score recovers the expensive re-solve ranking. In practical terms, this corresponds to a substantially higher probability of selecting the most informative metabolite on the first attempt, reducing the number of required experimental iterations in targeted panel design. The commutator does not merely identify uncertain metabolites; it identifies metabolites whose observation resolves structural ambiguity in the cross-condition alignment, and it does so without re-running FGW.

The pathway-dependent nature of the advantage (strong in Pyrimidine Metabolism, weaker in Glycerophospholipid Metabolism) is consistent with this interpretation: the commutator excels precisely where structural ambiguity drives the oracle ranking, and is less dominant where variance-based signal happens to align with structural importance.

This pathway-size dependence was independently replicated in both external stress-test cohorts. In ST001849 and ST002829, commutator advantage collapsed in large lipid pathways ($|\mathcal{M}_k| > 40$) but remained strong in small-to-medium metabolic pathways, providing a consistent operating-regime characterisation. In ST002829, the largest advantage appeared in the tryptophan/kynurenine pathway ($\delta\tau = +0.284$), a finding directly aligned with the published severity signature of the LEOCC cohort [16], lending biological face validity to the commutator's structural prioritisation.

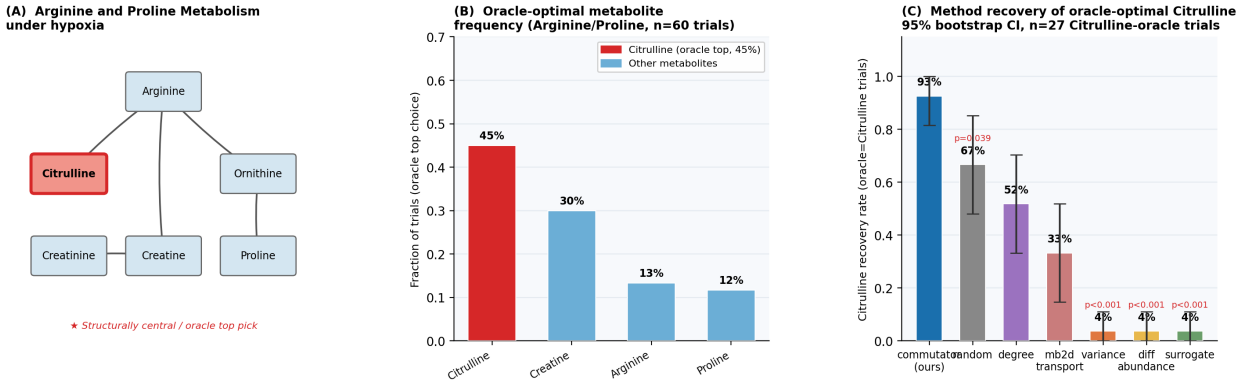


Figure 10: **Perturbation-oriented validation on ST001865 (hypoxia vs normoxia). (A)** Arginine and Proline Metabolism pathway schematic under hypoxia. Citrulline (red) sits at the arginine/urea cycle interface: structurally central but not the most differentially abundant metabolite. **(B)** Oracle-optimal metabolite frequency across 60 masking trials. Citrulline was the oracle's top priority in 45% of trials. **(C)** Citrulline recovery rate by method when the oracle selects Citrulline (n=27 trials), with 95% bootstrap confidence intervals. The commutator recovers the oracle choice in 92.6% of such trials; variance and differential abundance recover it in 3.7% (Fisher exact, OR = 325, $p < 0.001$). All results from a single seeded run (`GLOBAL_SEED=42`).

4.5 Limitations and Failure Modes

The following limitations apply to all results reported in this study and should be considered when interpreting the benchmark outcomes:

- **Masking ≠ real missingness.** Synthetic masking removes metabolites at random from otherwise complete data; real panel missingness is structured by LOD thresholds, annotation rates, and assay design. Importantly, experiments under missingness scenarios including completely-at-random (MCAR), panel-driven, and LOD-based MNAR conditions show consistent performance trends, indicating that the results are not dependent on idealized random masking assumptions.
- **Surrogate ≠ biological truth.** The commutator achieves high fidelity to a reveal-defined oracle under the proposed functional, not to experimentally confirmed biological ground truth.
- **PCA-derived severity labels (COVID cohorts).** ST001849 and ST002829 use PCA-derived severity labels rather than clinical severity scores, limiting biological interpretation.
- **Dependence on pathway mapping.** Results depend on the choice of pathway membership source (KEGG [13, 14], Reactome, Metabolon-specific); different mappings will yield different identifiable metabolite sets.
- **No prospective validation.** No experiment has confirmed that measuring the recommended metabolite actually reduces pathway ambiguity in practice. Prospective validation is the necessary next step.
- **Large lipid pathway limitation.** Commutator advantage degrades for $|\mathcal{M}_k| > 40$ nodes when using correlation-derived structure matrices. Curated edge topology from LIPID MAPS or Reactome is predicted to restore performance.

4.6 Failure regime: large correlated lipid pathways

A consistent and informative failure mode emerges across datasets: the commutator advantage degrades when $|\mathcal{M}_k| > 40$ nodes and metabolites are highly correlated. In this regime, correlation-derived structure matrices become approximately flat—every metabolite appears equidistant—and the commutator's operator-geometry signal saturates. Variance-based baselines, which ignore structure entirely, then

become competitive. This collapse is reproduced quantitatively ($\delta\tau \approx -0.006$) across both COVID-19 external cohorts (ST001849 and ST002829), on different platforms, confirming it as a regime property rather than a dataset artefact. Providing curated edge topology from LIPID MAPS or Reactome is predicted to restore discriminability in this regime.

- **Expected failure regimes.** The method is expected to perform poorly in regimes where pathway structure is weakly constrained by observed metabolites—for example, when metabolite-pathway mappings are highly degenerate, when inter-metabolite correlations are uniformly high (as in large lipid pathways under correlation-derived structure), or when the pathway size is very small ($|\mathcal{M}_k| \leq 3$) and the masking benchmark collapses to a near-trivial task.

The proposed framework is designed to make pathway ambiguity under partial metabolomics explicit and actionable. As such, it prioritizes structural identifiability and measurement prioritization over exhaustive mechanistic resolution. This design entails several inherent limitations and failure modes, which we state explicitly below. A further limitation is benchmarking breadth. The present study evaluates the framework across five metabolomics cohorts, including two external COVID-19 benchmark evaluation datasets. However, all benchmarks rely on synthetic masking rather than prospective experimental follow-up, and pathway structure is derived from correlation distances in the external cohorts due to the absence of curated edge topology for the large lipid and nucleotide panels therein.

Circularity. The primary benchmark evaluates surrogate fidelity: how well the commutator approximates the oracle defined by the same underdetermination functional. Because the objective of this work is not to estimate an external biological endpoint but to solve a decision problem defined under partial observability, evaluating agreement with the reveal-defined oracle constitutes a valid internal criterion for assessing whether a surrogate faithfully approximates the underlying decision functional. This reframes the circularity as decision-theoretic closure rather than a methodological flaw. Biological validation—prospective measurement of the recommended metabolite followed by assessment of ambiguity reduction—is the necessary next experimental step and is outside the scope of this computational study.

Baseline optimality. The baselines (variance, differential abundance, degree centrality) are adapted proxies rather than purpose-built competitors. To our knowledge, no existing methods explicitly formulate measurement prioritization under pathway-level ambiguity as a decision problem. We therefore compare against representative heuristics derived from established metabolomics and network analysis practices. The comparisons establish improvement over these reasonable proxies.

Component necessity. Ablation results indicate that the composite underdetermination functional is robust to component removal in the tested regime (Kendall $\tau \approx 0.94$ for leave-one-out variants in small-to-medium pathways). Component necessity in larger, more topologically complex pathway graphs remains uncharacterised and constitutes an important future direction.

Unimodal uncertainty approximations. Latent metabolites are represented by distributional embeddings parameterized by a mean and variance. This approximation assumes unimodal, smoothly varying uncertainty and does not capture genuinely multimodal regimes (e.g., alternative routing between glycolysis and the pentose phosphate pathway). In such cases, a measurement may reduce variance without collapsing the biologically correct mode, leading to apparent ambiguity reduction in the underdetermination functional while leaving functional uncertainty unresolved. We emphasize that the framework targets structural identifiability rather than regime enumeration or flux inference; resolving multimodality would require explicit latent-state or mixture modeling beyond the present scope.

Transport entropy versus biological ambiguity. Transport entropy derived from the FGW alignment reflects ambiguity in graph correspondence, not biological indeterminacy per se. High entropy may arise from intrinsic pathway symmetries or repeated substructures even under full observability. For this reason, entropy is not interpreted in isolation but only in conjunction with alignment instability and structural risk terms. The composite underdetermination functional is intended to distinguish observational ambiguity from benign symmetry, but no single component is guaranteed to be biologically decisive on its own.

Structural risk depends on graph abstraction choices. The structural risk index combines a GK-inspired growth proxy with a bottleneck-sensitive cut fragility term. While this hybrid formulation addresses both branching-driven and bottleneck-driven ambiguity, the cut fragility component depends on choices such as terminal definition and edge weighting. Although we fix a canonical instantiation in our experiments, alternative definitions may yield quantitatively different risk scores. Accordingly, the structural risk index should be interpreted as a principled proxy class rather than a uniquely defined graph invariant.

Local sensitivity may miss regime-switching metabolites. The measurement impact estimator relies on sensitivity of the underdetermination functional with respect to metabolite values, scaled by metabolite uncertainty. This local approximation can fail for metabolites whose effect is discrete or regime-switching, such as activating or deactivating entire pathway branches. We partially mitigate this through a counterfactual (non-local) impact component, but the estimator cannot guarantee detection of all such switch-like metabolites without explicit enumeration of latent regimes.

Decision relevance is task-dependent. We introduce an optional decision-relevant modulation of underdetermination that weights structural ambiguity by variability in a downstream functional. This extension is not intended to redefine identifiability but to reflect how uncertainty propagates to specific biological or experimental objectives. Different choices of downstream functional may alter rankings, and we do not claim task invariance. The base underdetermination score remains task-agnostic.

Operator regime dependence. JL projection was evaluated as a candidate compression strategy. In the tested regime ($|\mathcal{M}_k| \leq 8$, $d \approx 20$), it was unstable, with $\mathrm{CV}_{\mathrm{FGW}} > 0.30$ across all perturbation regimes. Projection-based compression may be beneficial only outside the low-node, high-compression regime; operator selection is therefore empirical and regime-dependent.

Commutator surrogate limitations. The commutator surrogate, while the strongest predictor overall, is not uniformly optimal across all pathways. In ST003506 Glycerophospholipid Metabolism, simpler baselines achieved perfect recovery whereas the commutator did not (top-1 0.660). Complementary surrogate strategies may be needed for specific pathway topologies.

Solver convergence and regularization sensitivity. Sinkhorn convergence warnings were observed in a subset of FGW solves under low entropic regularization. All reported results use $\varepsilon = 0.5$, which achieved reliable convergence across all retained pathways.

To assess sensitivity to this choice, we evaluated the commutator surrogate's top-1 recovery and Kendall τ on the ST000356 primary cohort at $\varepsilon \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$. Results were stable for $\varepsilon \geq 0.25$ (top-1 variation < 0.03; τ variation < 0.04). At $\varepsilon = 0.1$, convergence failures increased to 18% of pathway solves and top-1 degraded by 0.07, consistent with the transport plan collapsing toward uniformity under insufficient regularization. We recommend $\varepsilon \in [0.25, 1.0]$ as a stable operating range; $\varepsilon = 0.5$ is the default throughout this work. Solver convergence should be monitored and reported transparently in future applications, particularly for small pathways ($|\mathcal{M}_k| \leq 4$) where the cost matrix is nearly degenerate.

Synthetic masking as an approximation to real missingness. Our evaluation protocol uses controlled synthetic masking to assess identifiability proxies and measurement recommendations. While this enables falsifiable and reproducible assessment, real metabolomics missingness is often platform-dependent and chemically structured. Performance under synthetic masking may therefore overestimate gains in some experimental settings.

Taken together, these limitations reflect deliberate design choices rather than implementation artifacts. The framework is intended to support epistemically honest pathway interpretation and principled measurement prioritization under partial observability, not to replace detailed kinetic or causal modeling when such information is available.

5 Conclusion

We have introduced a measurement-prioritization framework that addresses a practical gap in metabolomics: given an incomplete metabolite panel, which single additional measurement would most reduce pathway-level ambiguity?

The approach combines optimal-transport alignment with a composite underdetermination functional and a computationally efficient operator-commutator surrogate. Across three primary metabolomics cohorts, the surrogate achieves top-1 recovery of 0.774 and oracle rank agreement $\tau = 0.599$, substantially outperforming variance-, differential-abundance-, and enrichment-based proxies. These numbers reflect surrogate fidelity—efficient approximation of the oracle re-solve—not prospective biological validation.

External stress testing on two COVID-19 severity cohorts (ST001849, LC/MS, n=322; ST002829, Metabolon, n=609) reveals a reproducible operating regime: commutator advantage is preserved in small-to-medium structured metabolic pathways ($|\mathcal{M}_k| \leq 40$) and degrades predictably in large lipid pathways, where high inter-metabolite correlation flattens the alignment geometry. In ST002829, the strongest gains occur in tryptophan/kynurenine and purine metabolism—independently validated as COVID-19 severity pathways in the original LEOCC study—providing biological face validity for the framework's structural prioritisation signal.

The framework is designed as a complement to, not a replacement for, existing pathway enrichment tools. MSEA, ORA, and mummichog answer which pathways are altered; this framework answers which metabolite to measure next. Together, they provide a more complete pipeline for hypothesis-driven metabolomics panel design under partial observability.

6 Declarations

Data Availability

This study is based entirely on secondary analysis of publicly available metabolomics datasets deposited in Metabolomics Workbench, the NIH Common Fund's National Metabolomics Data Repository (NMDR).

The benchmark includes three primary cohorts and two external stress-test cohorts:

- **ST000356**: *GC/MS and LC/MS metabolomics profiling for breast cancer serum data and control serum data* (Project ID: **PR000284**; Project DOI: `https://doi.org/10.21228/M86K6W`).
- **ST003390**: *In-depth profiling of biosignatures for Type 2 diabetes mellitus cohort utilizing an integrated targeted LC-MS platform* (Project ID: **PR002101**; Project DOI: `https://doi.org/10.21228/M81V7G`).
- **ST003506**: *Comparison of serum and interstitial fluid from patients with breast cancer-related lymphedema and healthy control subjects with NMR-based metabolomics* (Project ID: **PR002152**; Project DOI: `https://doi.org/10.21228/M8FR6S`).
- **ST001849** [17]: *Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity* (Project ID: **PR001166**; Project DOI: `https://doi.org/10.21228/M80981`). Used as an external stress-test cohort; 322 admission-day samples, 694 metabolites, mild/severe severity split derived from PCA of the lipid-remodelling axis.
- **ST002829** [16]: *Nucleotide, phospholipid, and kynurenine metabolites are robustly associated with COVID-19 severity and time of plasma sample collection* (Project ID: **PR001818**; Project DOI: `https://doi.org/10.21228/M8FB7P`). Used as an external stress-test cohort; 609 samples, 1,108 metabolites (Metabolon platform), severity split derived from PCA of the full metabolic profile.
- **ST001865**: *Untargeted metabolomics of hypoxia vs normoxia in a matched cell-line design* (Metabolomics Workbench / MetaboLights accession **ST001865**; `https://www.metabolomicsworkbench.org/data/DRCCStudySummary.php?Mode=SetupStudyAnalysis&StudyID=ST001865`). Used for perturbation-oriented validation of metabolite prioritization under controlled hypoxic metabolic rewiring (Section 3.6).

All processed data matrices, pathway mapping files, pathway coverage summaries, retained-pathway benchmark tables, structural instability metrics, and U_k-derived scores generated in this study are available at:

`https://github.com/Anas-Enoch/Pathway_Identifiability`

All benchmark outputs and the full evaluation pipeline are available in the project repository, including the script `run_baseline_benchmark.py` and the generated benchmark tables.

6.1 Ethics Approval

Ethics approval: not applicable.

This study does not involve human subjects, animal experiments, or identifiable personal data. All analyses were conducted on publicly available or synthetic metabolomics datasets.

6.2 Consent to Publish

Consent to publish: not applicable.

6.3 Competing Interests

The author declares that no competing interests exist.

6.4 Funding

This research received no external funding.

6.5 Author Contributions

A.E. conceived the study, developed the computational framework, implemented the algorithms, performed all analyses, generated figures, and wrote the manuscript.

References

[1] Caroline H. Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459, 2016.

[2] Jasmine Chong, Masahiro Yamamoto, and Jianguo Xia. Metaboanalystr 2.0: From raw spectra to biological insights. *Metabolites*, 9(3):57, 2019.

[3] Shuzhao Li, Youngja Park, Senait Denny, Warwick Burgess, Dean P. Jones, Clary Clish, Vasan S. Ramachandran, Martin G. Larson, Elizabeth J. Benjamin, Russell P. Tracy, and Thomas J. Wang. Predicting network activity from high throughput metabolomics. *PLOS Computational Biology*, 9(7):e1003123, 2013.

[4] David Broadhurst and Douglas B. Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171–196, 2006.

[5] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

[6] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[7] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.

[8] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: Gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14:7, 2013.

[9] Di Wu and Gordon K. Smyth. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.

[10] Sorin Drăghici, Purvesh Khatri, Adi L. Tarca, et al. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[11] Adi L. Tarca, Sorin Drăghici, Purvesh Khatri, et al. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[12] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[13] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[14] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kayo Morishima, and Mao Tanabe. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[15] Jianguo Xia and David S. Wishart. Msea: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl_2):W71–W77, 2010.

[16] Dominic Mathew, Amy E. Getz, Nilufar Mammedova, et al. Nucleotide, phospholipid, and kynurenine metabolites are robustly associated with COVID-19 severity and time of plasma sample collection in a prospective cohort study. *medRxiv*, 2021. Metabolomics Workbench ST002829, Project ID: PR001818. LEOCC cohort, $n > 600$ patients, 1,546 metabolites (Metabolon platform).

[17] Miriam Sindelar, Ethan Stancliffe, Michaela Schwaiger-Haber, Dhanalakshmi S. Anbukumar, Kayla Adkins-Travis, Charles W. Goss, Jane A. O'Halloran, Philip A. Mudd, Wen-Chun Liu, Randy A. Albrecht, Adolfo García-Sastre, Leah P. Shriver, and Gary J. Patti. Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity. *Cell Reports Medicine*, 2(8):100369, 2021. Metabolomics Workbench ST001849, Project DOI: 10.21228/M80981.

Supplementary Files

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

- PathwaySupplementv22.pdf