

Introduction

Scientists working across omics are impeded by siloed datasets and fragmented knowledge bases. Interpreting matched metabolomics, proteomics, phosphoproteomics and transcriptomics data is hindered by fragmented entity references and by the lengthy manual matching of entities from one omics layer to the next.

Public knowledge bases are rich, but they are built for breadth rather than for one task, so lookups are too slow for interactive entity mapping. AI lowers the barrier to using them, but provenance and hallucination concerns limit trust in AI-only answers for scientific decisions.

We built an **automated graph-based entity mapping system**, grounded in experimental data and curated external knowledge, and built for one job: **mapping entities across datasets, experiments and omics layers**. It enables faster, more reproducible longitudinal and multi-omics analysis through:

- A graph-based entity translation layer: proteins, peptides, PTMs, genes, transcripts and metabolites mapped consistently across experiments.
- Explicit handling of the hard cases: gene-to-isoform mappings and ambiguous peptide-to-protein assignments from inference algorithms.
- Use-case-driven design: built around the queries researchers actually run, and provenance-grounded so every connection is traceable.

1

One graph across all omics layer

Graph schema and implementation

The system runs on **Amazon Neptune** and integrates **curated external resources**, including UniProt, Ensembl and Reactome, with **experimental omics datasets** while preserving provenance and access control.

Biological entities with their identifiers and isoforms are modelled as nodes. Peptide-protein, transcript-gene and cross-omics relationships are modelled as edges, so one-to-many and many-to-many cases stay explicit and traceable instead of being collapsed at load time.

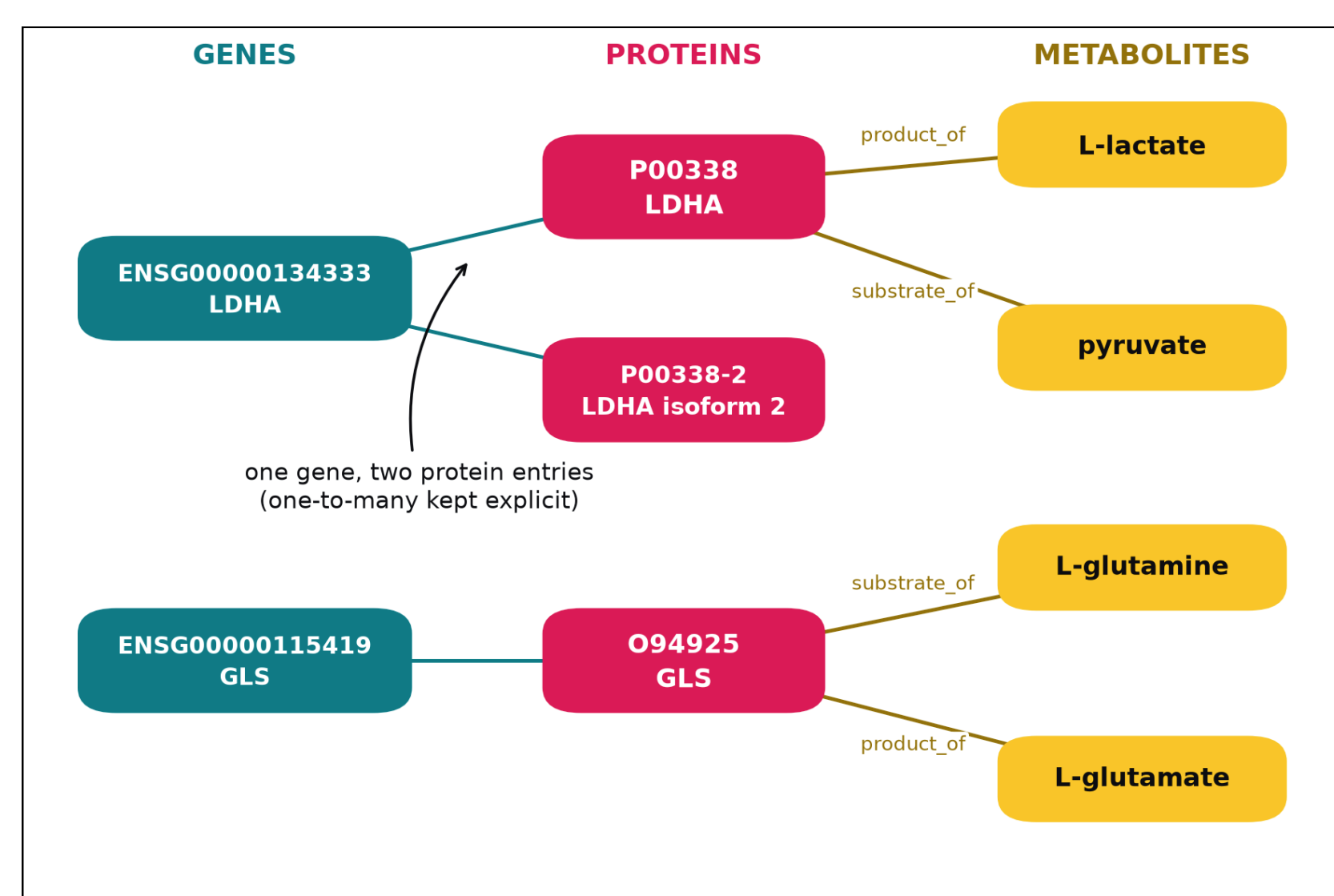


Figure 1. Multi-omics cross-reference graph schema. Example instance graph over the three node types: genes, proteins and metabolites, with the edge types that connect them.

Evaluation dataset

Hu 2024 ccRCC¹: The data showcased here comprises of **matched proteomics, transcriptomics and metabolomics data** from a study of clear cell renal cell carcinoma. Each dataset was separately uploaded and processed using Mass Dynamics.

References

1. Hu J, et al. Multi-omic profiling of clear cell renal cell carcinoma identifies metabolic reprogramming associated with disease progression. *Nat Genet* 56, 442-457 (2024).
2. Argelaguet R, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 21, 111 (2020).

Conflicts of interest

Giuseppe Infusini, Aaron Triantafyllidis, Paula Burton Ngov and Andrew I. Webb are co-founders of Mass Dynamics, a for-profit enterprise delivering software as a service for the processing, analysis and sharing of proteomics data. The remaining authors are employees of Mass Dynamics.

2

Resolve entities across identifiers, isoforms and modalities

Entity mapping runs at upload and on request, connecting each entity to related entities in other molecular layers, and to the same entity in datasets previously acquired, so a group of experiments is linked into a longitudinal knowledge base.

The **mapping modes** cover:

- Identifiers recorded as reported by the raw data processing software, with cross-references to the canonical UniProt and Ensembl entries. We never redo the protein inference step.
- Peptide and modified-sequence assignments kept as they were made at search time, alongside every other protein the same stretch of sequence was assigned to in other searches, so peptides and proteins still match when the inference changes.
- Isoform and protein-group resolution via shared peptide evidence

Planned

The same node and edge model extended to transcript-level entities, metabolites and pathways, with public knowledge bases as additional evidence sources.

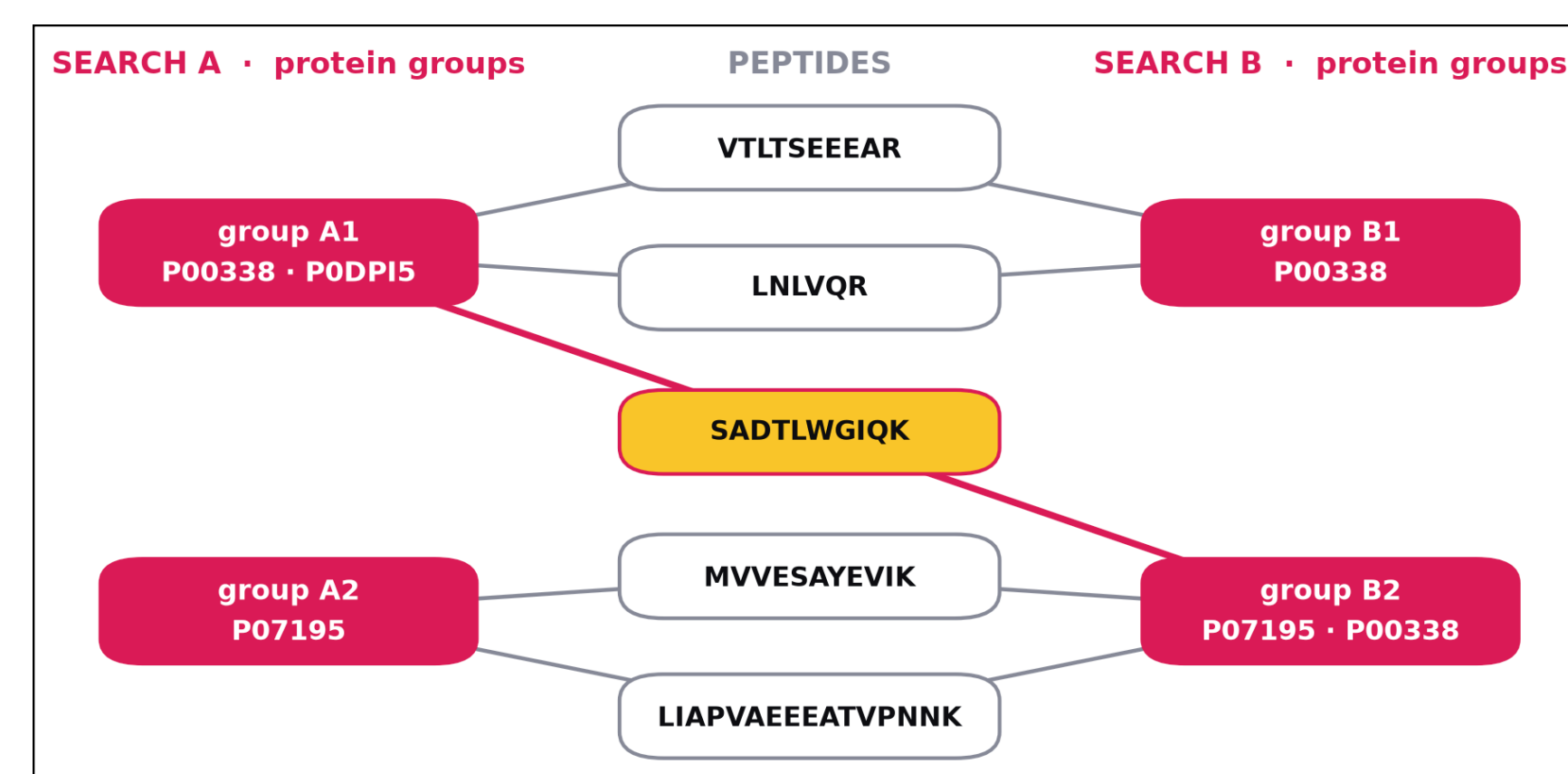


Figure 2. Peptides bridge differently inferred protein groups. SADTLWGIQK is assigned to group A1 in search A and to group B2 in search B. Because the peptide is its own node carrying both assignments, the two groups stay connected and remain matchable across studies.

3

From protein to transcript in one click

Every dataset uploaded to Mass Dynamics is added to the user's own knowledge base, so a researcher can select a gene and the entity mapping service automatically returns the matching protein accessions, together with evidence behind every link.

Differential expression (DE) analysis was performed separately for proteins and genes using the demo dataset. The two DE results are shown side by side, with one transcript highlighted, making agreement and divergence between protein and transcript responses visible at a glance.

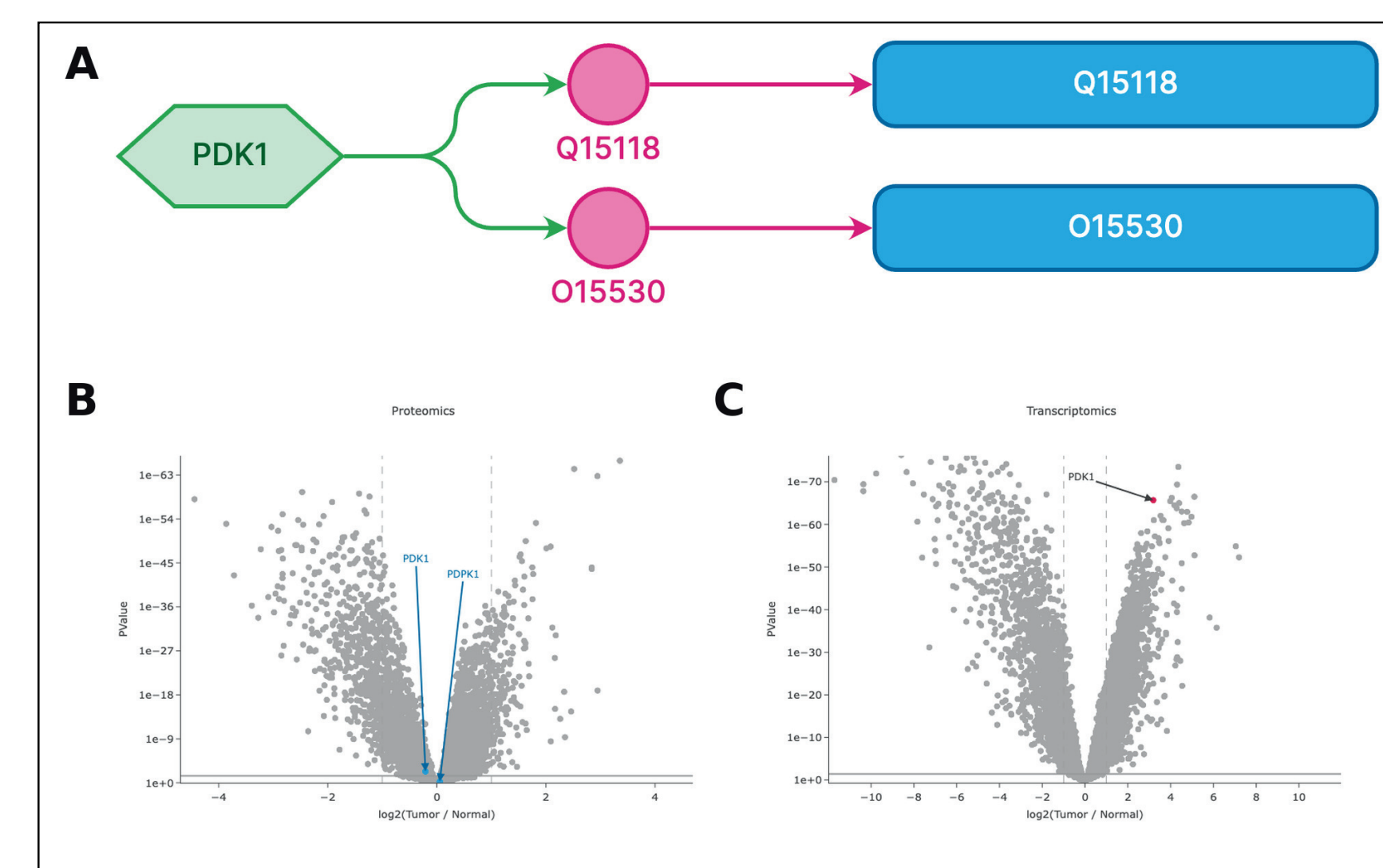


Figure 3. Transcript to proteins in one click.

A) Selecting gene PDK1 on the transcriptomic volcano plot in C) automatically returns the associated protein accessions. The search database annotates the two proteins with different genes, PDK1 for Q15118 and PDK1 for Q15530, while in the reference knowledge base the symbol PDK1 covers both, as the primary symbol of ENSG00000152256 and a legacy alias of PDK1, ENSG00000140992. Gene to accession links are built from UniProt idmapping and Ensembl BioMart, symbols and aliases from HGNC and NCBI Gene, and each edge carries the evidence that produced it. **B)** Proteomics differential expression, tumour versus normal: neither mapped protein moves. **C)** The same gene in the transcriptomic view, strongly up. The discordance becomes visible only once the entities are mapped.

4

Integrate the layers

Multi-Omics Factor Analysis (MOFA)² is an unsupervised generalisation of PCA integrating several omics views measured on the same samples. Each view keeps its own features, and the model infers a small set of latent factors that capture the main axes of variation across all of them. Every factor carries the fraction of variance it explains in each view, which separates variation shared across layers from variation specific to one, and a weight per feature saying which entities drive it. Those top-weighted features arrive as whatever identifier each view was measured in, and entity mapping is what lets them be read together: the factor is interpreted on the graph, with the evidence behind every link.

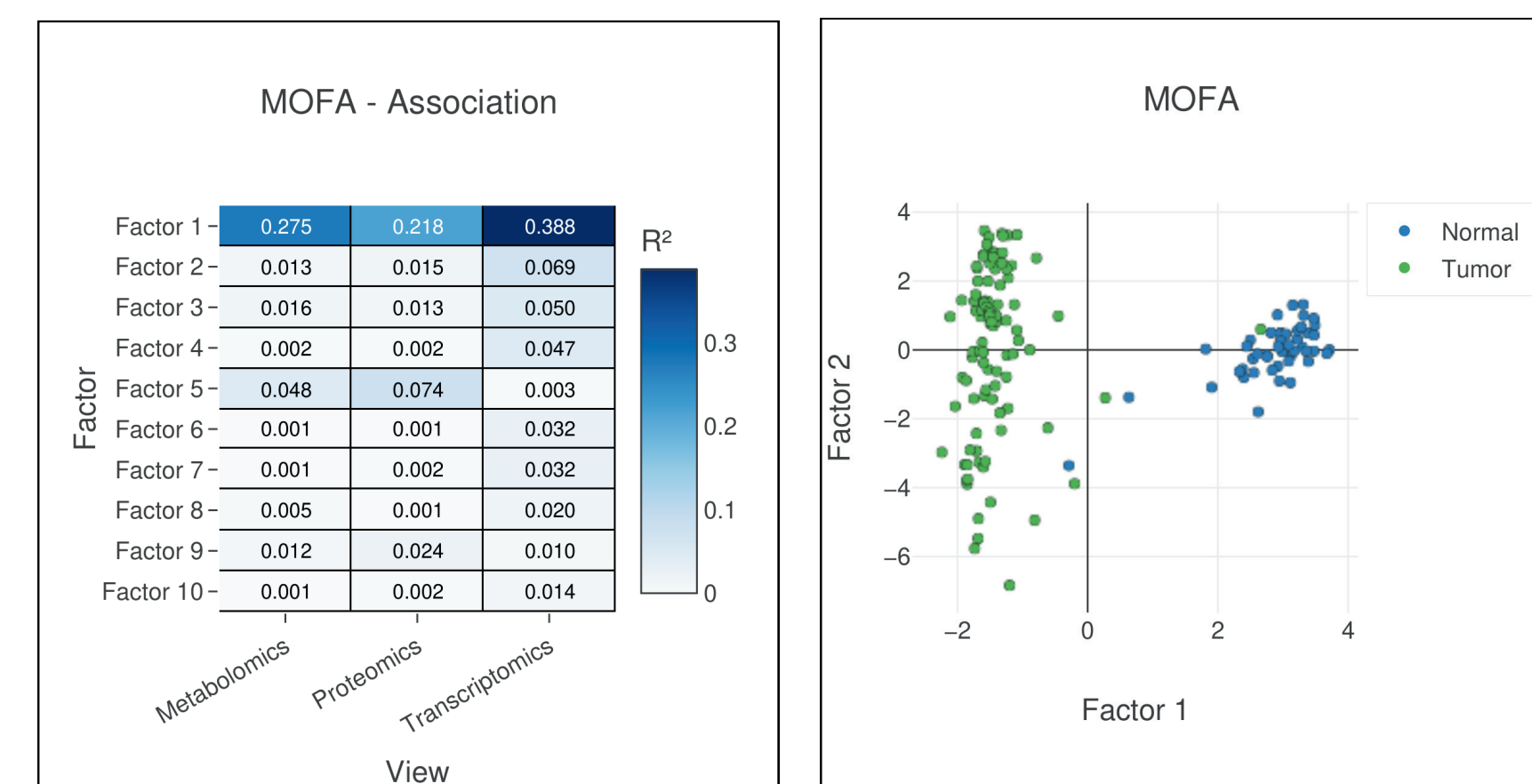


Figure 4. MOFA+ across three views of 150 ccRCC samples. Factor 1 dominates every view and separates the 100 tumours from the 50 adjacent normals, $\eta^2 = 0.92$, $p = 6e-83$. Its top features are the nephron segment markers lost in tumour.

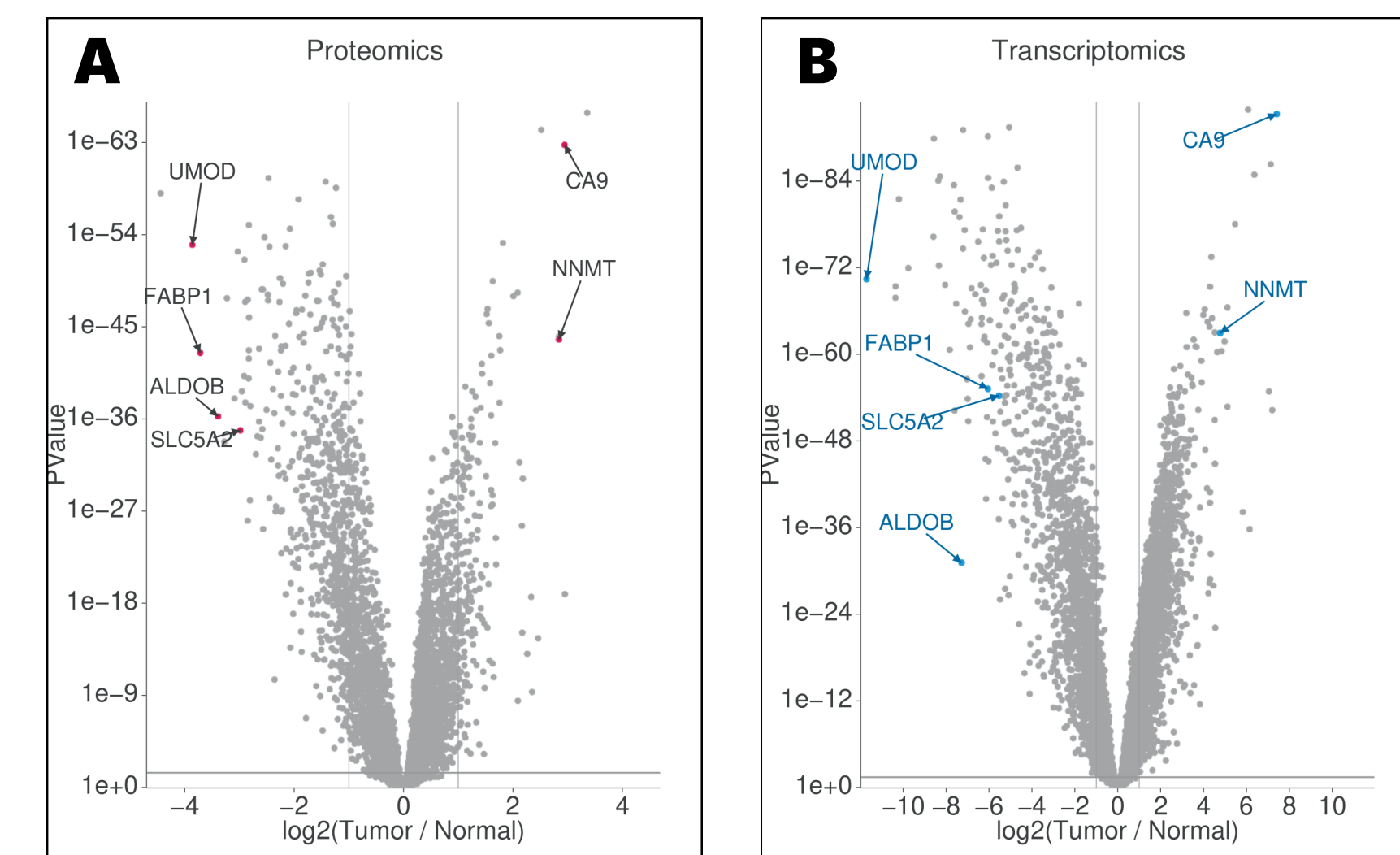


Figure 5. Mapping runs both ways.

A) Six proteins selected on the tumour vs normal proteomics volcano plot.

B) The genes mapped to them, on the transcript volcano of the same data. The click works in reverse too, and it has to: 65 of the 5,994 proteomics features in this study carry a label that is not a UniProt accession at all, mitochondrial legacy names such as ND3 and CYTB, withdrawn symbols such as G6PC, and fusion readthroughs such as SEPT5-GP1BB. The graph resolves those; a string match does not.

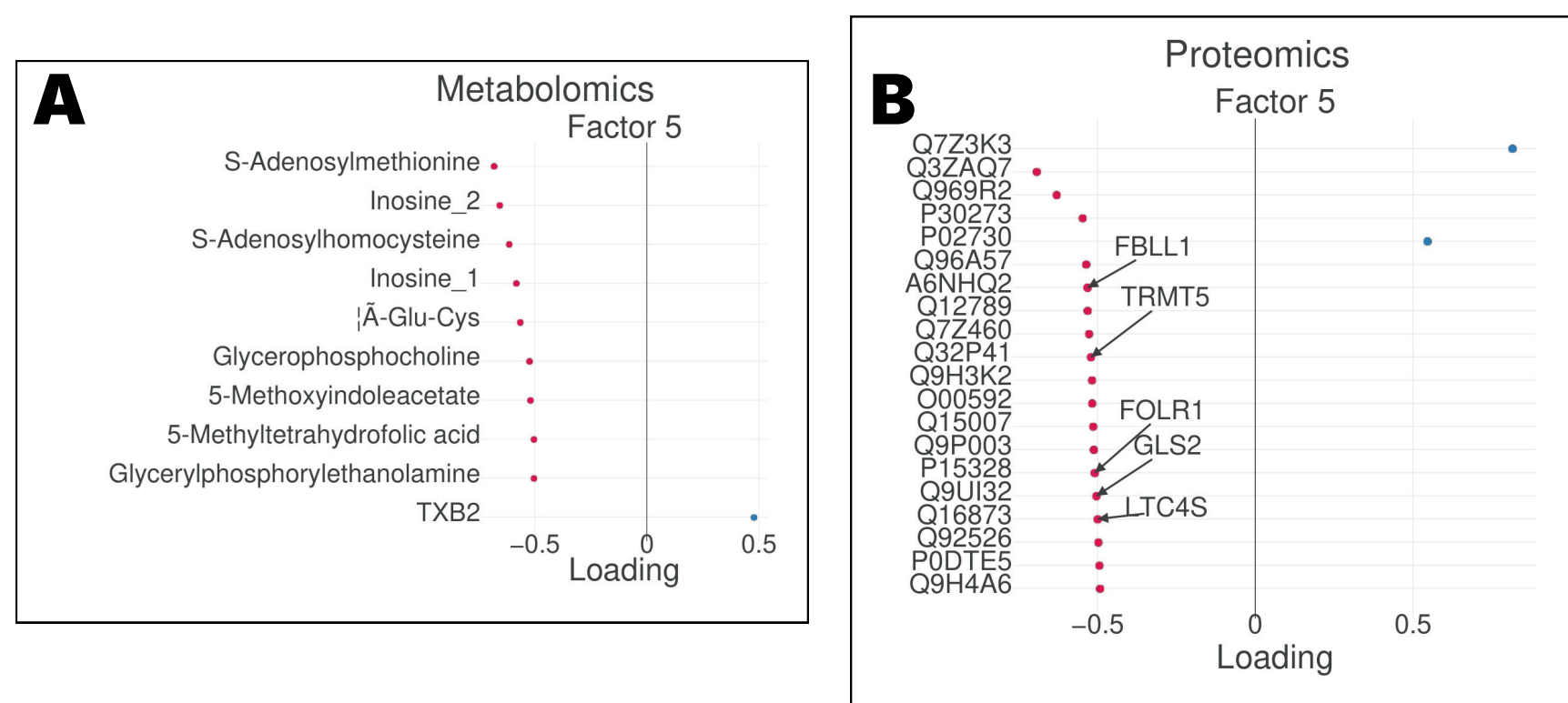


Figure 6. A one-carbon axis resolved only across layers. Factor 5 explains 4.8% of metabolite and 7.4% of protein variance against 0.3% of transcript variance.

A) S-adenosylmethionine (SAM) and S-adenosylhomocysteine (SAH) are among the largest metabolite loadings, at ranks 1 and 3. SAM is the universal methyl donor; SAH is its demethylated product and a competitive inhibitor of the methyltransferases that generate it, so the SAM/SAH ratio constrains methylation capacity. 5-methyltetrahydrofolate (methionine regeneration) and γ -Glu-Cys (committed intermediate of glutathione synthesis) load on the same factor.

B) Five highlighted proteins act on these metabolites. FBLL1 and TRMT5 are SAM-dependent methyltransferases; TRMT5 methylates G37 of mitochondrial tRNA, consuming SAM and releasing SAH per turnover. FOLR1 mediates 5-methyltetrahydrofolate uptake; GLS2 and LTC4S flank γ -Glu-Cys in glutathione synthesis and turnover. Metabolite to protein mapping is in beta.

Conclusions

A purpose-built entity mapping graph is a trusted foundation for multi-omics systems biology: trace provenance, explore connected molecular evidence across independent studies, and run integrative methods such as MOFA on data that would otherwise need manual reconciliation.

Keen to try this with your own data?

Simply scan the QR code to get started.

