A GRAPH KNOWLEDGE BASE FOR MULTI-OMICS ENTITY MAPPING

Andrew I. Webb[1]; Anna Quaglieri[1]; Mansi Aggarwal[1]; Mark R. Condina[1,2]; Aaron Triantafyllidis[1]; Paula Burton Ngov[1]; Giuseppe Infusini[1]
[1]Mass Dynamics, Melbourne, Victoria 3000, Australia; [2]Clinical & Health Sciences, University of South Australia, Adelaide 5095, Australia

Mass Dynamics

Introduction

The rapid growth of proteomics and other omics data has outpaced traditional analysis methods, making it difficult to compare results across experiments and external databases. Researchers struggle with fragmented datasets, inconsistent identifiers for the same biomolecules, poor reproducibility, and the laborious task of manually cross-referencing proteins or genes between studies.

To address these challenges, we have developed a graph-based entity translation layer that systematically **maps biomolecular entities across diverse experiments**, ensuring consistent identification and facilitating seamless data integration. The **goal** of the platform is to enable **faster, more reproducible multi-omics analyses** through:

- **A graph-based entity translation layer:** consistently map biomolecular entities (proteins, peptides, genes, metabolites) across multiple experiments.
- **Explicit handling of key mapping challenges:**
  - Single-to-multiple mappings (e.g., gene-to-isoforms)
  - Ambiguous peptide-to-protein assignments from inference algorithms
- **Use-case-driven design:** Structured around common analytical queries, enabling rapid retrieval for cross-dataset and longitudinal analyses.

Scope:

- **Current focus:** Mapping entities within and between experimental proteomics datasets.
- **Future expansion:** Incorporation of external biological knowledge bases to enrich data with broader context

Graph Schema & Implementation

We developed a layered graph schema in Neo4j, a graph database platform. In this initial iteration, the schema focuses exclusively on **internal experimental datasets**, ensuring **secure handling of sensitive client data**. The graph is structured to represent experimental entities (proteins, peptides, etc.) and their relationships clearly. Neo4j's schema-less property graph model allows for flexible expansion, supporting future integration with external biological knowledge and additional entity types (gene, metabolites, transcripts, etc.) as development progresses.

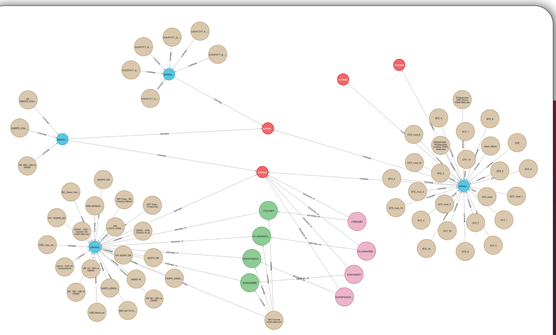


Figure 1. Graph Example Representation of Experimental Data

A visualization demonstrating how experimental proteomics data is modeled within the internal Neo4j graph layer. Protein groups (blue nodes) identified in datasets (brown nodes) are explicitly represented, highlighting direct relationships with individual protein entities (red nodes). Protein group nodes may also connect to peptide nodes (ModifiedSequence shown in green, StrippedSequence in pink), facilitating rapid mapping between proteins and peptides, even if these associations originate from separate data-processing steps.

Entities & Relationships Creation

All **biomolecular entities** are modeled as **nodes** (e.g. *ProteinGroup*, *ModifiedSequence*, etc.) and their **interactions** or mappings are captured as **edges**. For example, a *ModifiedSequence* node connects to the *ProteinGroup* node(s) it was assigned to, explicitly representing cases where one peptide sequence was assigned to multiple proteins.

Step by step:

1. *ProteinGroup* nodes contain Gene information derived directly from processing software annotations.
2. Additional explicit edges (e.g *'is expressed from'*) are introduced in the knowledge layer, linking proteins to dedicated Gene nodes and naturally grouping all isoforms associated with a given gene.
3. Within the internal layer, Dataset nodes are linked to the proteins and peptides identified in them, so we can trace where each observation came from.

This graph design captures **one-to-one** as well as **one-to-many** and **many-to-many relationships** natively, avoiding the pitfalls of protein inference by making each possible connection explicit and traceable.

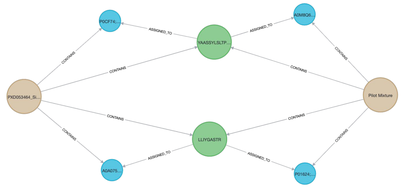


Figure 2. Peptide-to-Protein Connectivity in the Graph Model

An example visualization demonstrating how peptides (central nodes) explicitly connect to their respective protein groups across datasets. The graph structure captures ambiguous peptide-protein relationships clearly, with peptides linked directly to all potential source proteins, thereby enabling precise peptide-to-protein disambiguation and facilitating accurate cross-dataset comparisons.

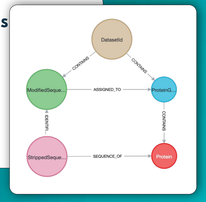
External Knowledge Integration

The platform is purposely designed to integrate curated external knowledge bases into the graph's external layer over time. For instance, UniProt provides protein metadata and known isoforms, Reactome contributes pathway associations, and Gene Ontology (GO) adds functional annotations.

We employed [1]BioCypher, an open-source framework for building biomedical knowledge graphs, to import these resources using standardized ontologies. **External nodes** (e.g. a protein entry from UniProt or a pathway from Reactome) are **linked to internal experimental nodes via shared identifiers** (such as matching UniProt accession numbers). This unification enriches the experimental data with broader biological context without exposing sensitive data, allowing the graph to answer complex queries that span both experimental findings and public knowledge.

Legend: Graph Schema Nodes And Relationship

- **DatasetId** (Brown Node): Represents individual experimental datasets.
- **ProteinGroup** (Blue Node): Groups of proteins identified within a dataset.
- **ModifiedSequence** (Green Node): Peptide sequences identified in experiments, including post-translational modifications.
- **StrippedSequence** (Pink Node): Peptide sequences without modification details, linked directly to proteins.
- **Protein** (Red Node): individual protein entities explicitly connected to Protein Groups or Stripped Sequences.



Query And Visualization Approach

We have integrated the **graph-based entity** layer directly into analytical **workflows within the Mass Dynamics application**. Proteins selected by users from interactive plots are dynamically highlighted in other plots and tables, enabling rapid visualization and data exploration across multiple datasets.

The underlying graph explicitly maps protein groups and peptides between datasets, allowing intuitive **exploration of relationships** and **cross-dataset matches**. Additionally, we have implemented a **flexible search** feature supporting exact and fuzzy matching, enabling **targeted retrieval of proteins**, peptides, or genes from selected datasets.

This workflow-centric approach **seamlessly integrates into researchers' existing analytical processes**, facilitating intuitive data exploration without requiring specialized knowledge of graph databases.

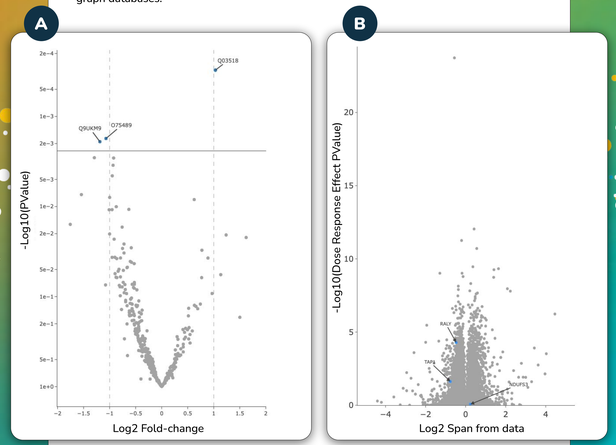


Figure 3. Integrated Exploration of Differentially Expressed Protein.

This figure demonstrates how the graph database facilitates seamless connectivity between different analytical visualizations, specifically linking proteins identified from volcano plot analyses (**A**) to their corresponding quantitative data in dose-response assays (**B**) from independent experiments. By explicitly mapping these cross-dataset protein relationships (**C**), the platform enables researchers to rapidly validate exploratory findings with targeted analyses, enhancing the robustness of biological insights and streamlining the research workflow.

Platform Capabilities & Early Validation

Stable Protein Identifier Mapping:
The platform consistently maps protein identifiers from diverse accession formats to stable canonical references, significantly enhancing reproducibility across experiments

Accurate Isoform Resolution:
Protein isoforms and variants are accurately connected across datasets via gene associations and shared peptide evidence, clarifying previously ambiguous mappings

Clear Peptide–Protein Disambiguation:
Ambiguous peptide identifications are explicitly modeled, linking each peptide to all possible source proteins and enabling precise resolution through additional context

Enhanced Cross-Dataset Discovery:
Immediate visibility of shared proteins across datasets facilitates rapid hypothesis generation and identification of biologically relevant connections that would otherwise remain hidden

Intuitive User Interface & Interactive Visualization:
Researchers can seamlessly search and visually explore protein, peptide, and gene relationships across datasets, without requiring specialized knowledge of graph databases

Improved Proteomics Analysis and Reproducibility:
By consolidating experimental data, resolving entity ambiguities, and simplifying data integration, the platform substantially accelerates analysis workflows and strengthens reproducibility

Impact & Future Directions

Our platform introduces a new, use-case-driven approach to proteomics data integration, employing a graph-based model tailored specifically to researchers' analytical queries. In its current form, it allows intuitive exploration of relationships across experimental datasets.

Next:

- Integrate external biological knowledge to enrich experimental data and context.
- Develop advanced, automated entity mapping for seamless integration of diverse multi-omics data types (proteomics, transcriptomics, metabolomics).
- Enhance visualization and user-interface features, directly aligning platform capabilities with practical research workflows.
- Establish a more interconnected scientific environment, facilitating rapid hypothesis generation, improved reproducibility, and deeper integrative insights across disciplines.

READY TO TRY MASS DYNAMICS?

Keen to try Mass Dynamics using your own data? Simply scan the QR code to book a custom demo.

References: 1: **Lobentanzer, S., Aloy, P., Baumbach, J., et al.** (2023). *Democratizing knowledge representation with BioCypher*. Nature Biotechnology, 41(8), 1056–1059. https://doi.org/10.1038/s41587-023-01848-y

Conflicts of interest: The authors Aaron Triantafyllidis, Paula Burton Ngov, Giuseppe Infusini and Andrew I. Webb declare that they are founders of Mass Dynamics, a for-profit enterprise, delivering software as a service in the processing, analysis and sharing of proteomics data. Mark R. Condina, Brendan Spinks, Mansi Aggarwal and Anna Quaglieri are employees of Mass Dynamics.