Identifying Biologically Compatible Experiments via Semantic Similarity

Núria Fàbrega[1], Ian Simpson[1], Kenneth Baillie[2], Mirella Lapata[1]

1. School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom,
2. Baillie Gifford Pandemic Science Hub, University of Edinburgh, Edinburgh, UK

There are **hundreds of thousands of transcriptomics experiments publicly available** in repositories such as GEO, with substantial potential for the **validation** of results, **meta-analysis**, and **discovery** of new biological insights. Yet these datasets are **rarely reused** after the original publication, as the **metadata** is often **incomplete and unstandardised**, making it **very time-consuming** for researchers to determine which studies are biologically compatible, leaving this **golden source of information mostly untapped**.

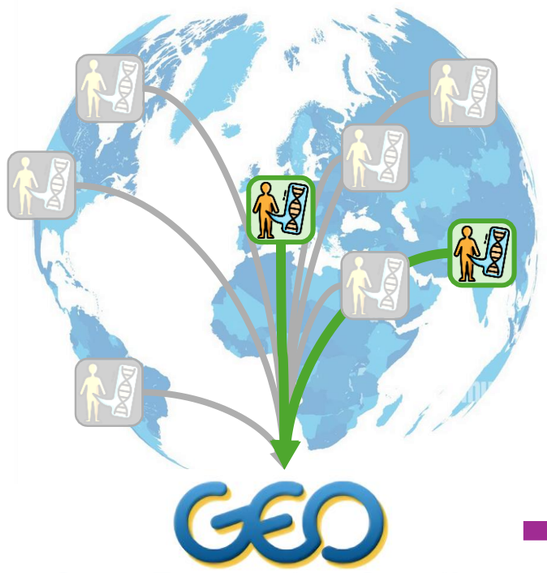
Can we identify biologically compatible experiments by comparing their metadata?

We build a pipeline that enriches experiment metadata and measures semantic similarity to retrieve compatible studies

BIOLOGICALLY COMPATIBLE:
comparable experimental conditions suitable for joint analysis

GOAL

Help scientists efficiently identify compatible experiments



Gene Expression Omnibus
Microarray data

Experiment description

Embed description

Create embeddings
Sentence Transformers

Retrieve nearest-neighbours via cosine similarity

Extract details from PubMed article

Standardise via ontologies

Metadata enrichment

Find **similar experiments**

Characterise **similarities & differences**

Compatible experiments

Compute gene expression similarity

GOAL

Determine the overlap between semantic similarity and biological similarity

Evaluation

Benchmark against curated experiments from Gemma
Assess retrieval quality using expert-curated annotations, ontology mappings, and known experiment relationships.

Compare retrieved experiments with published meta-analysis results
Do our retrieved experiments match those selected by domain experts for meta-analysis?

Measure alignment between semantic and biological similarity
Quantify how well metadata-based similarity reflects biological similarity, and analyse cases where they diverge.

Downstream applications

- Support validation of biological findings
- Select studies suitable for meta-analysis
- Provide a foundation for future multi-omics integration