TraitHunter

Mapping and extraction of biomedical traits via text embeddings

IEU Programme 3 talk

Yi Liu
3rd September 2024

Today’s talk

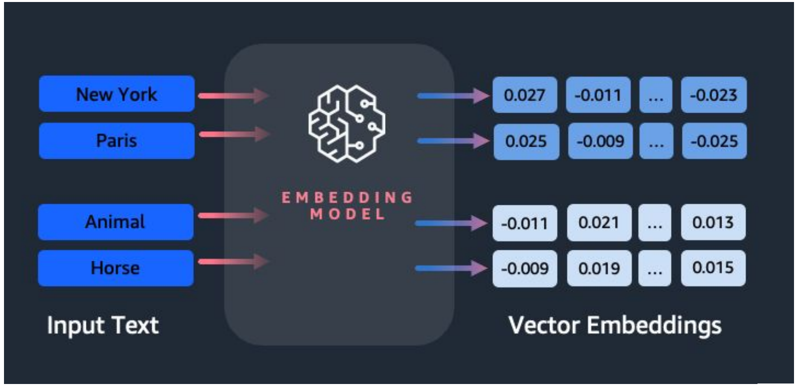
The main aim for today’s talk is to showcase a web service for trait mapping.

This is really quick and dirty hack, however feedbacks on various aspects are deeply appreciated.

Outline

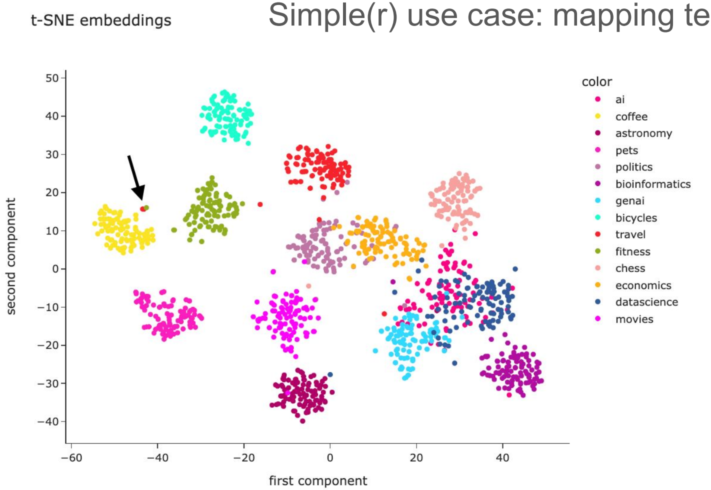
- Concepts
- Background
- TraitHunter
  - Concepts
  - Live demo
- Next steps

Concept: Text embeddings

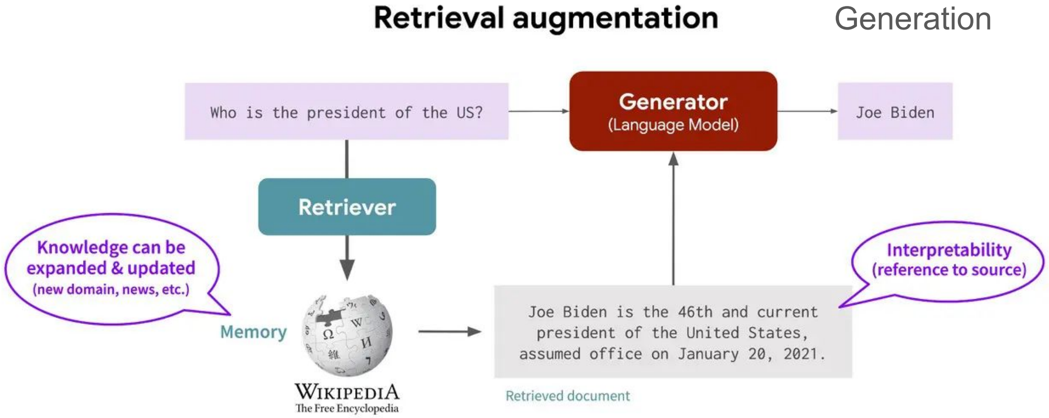


Convert text into their (semantic) vector representations via an encoder model

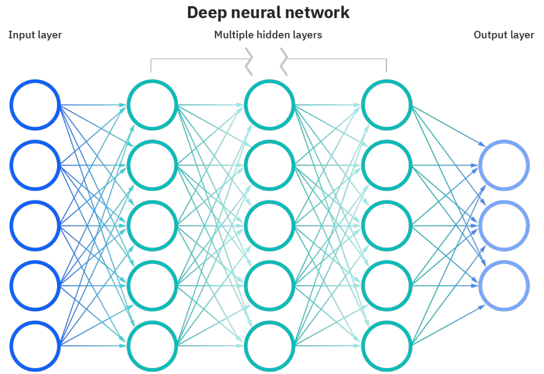
Simple(r) use case: mapping text



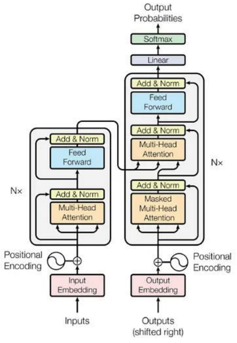
Retrieval Augmented Generation



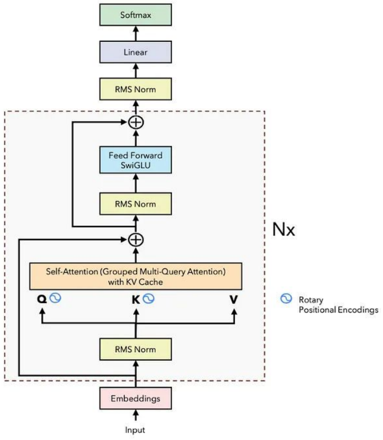
Concept: Large Language Models (LLM)



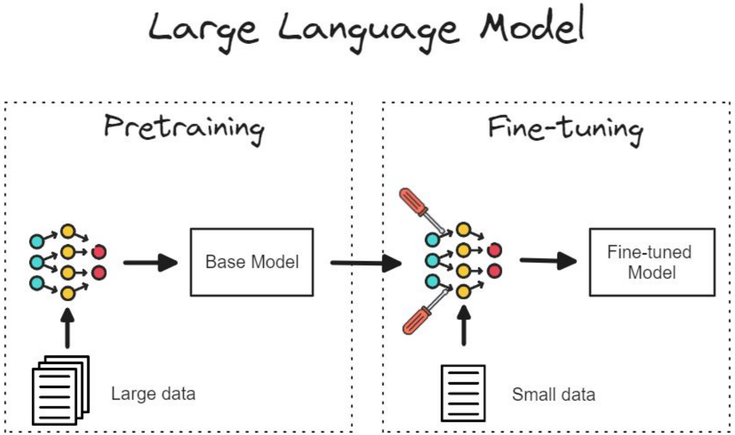
Transformer vs LLaMA

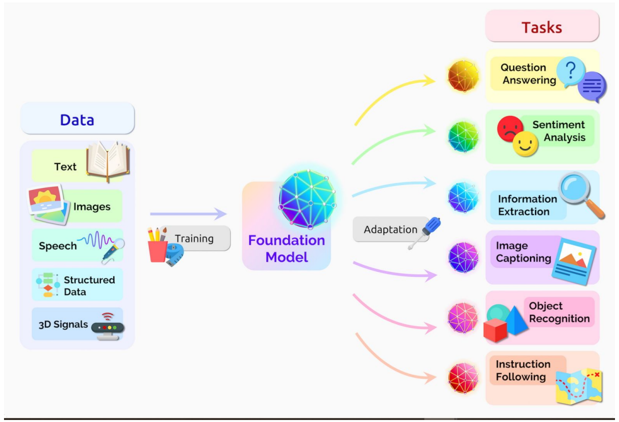


Transformer
("Attention is all you need")


Softmax
Linear
RMS Norm
Feed Forward SwiGLU
RMS Norm
Self-Attention (Grouped Multi-Query Attention) with KV Cache
Q
K
V
RMS Norm
Nx
Rotary Positional Encodings
Embeddings
Input

LLaMA



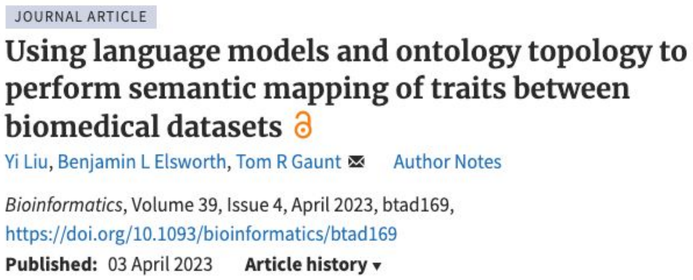


Background: related works

In Liu, et al. (2023) we approached the trait mapping problem with a two stage approach:

- Identify a sub set of related traits with naive text embeddings, from the broad set
- Map the traits using an LLM task model finetuned on ontologies

This approach is put into use in the ASQ platform (Liu & Gaunt, 2024) for querying EpiGraphDB evidence.



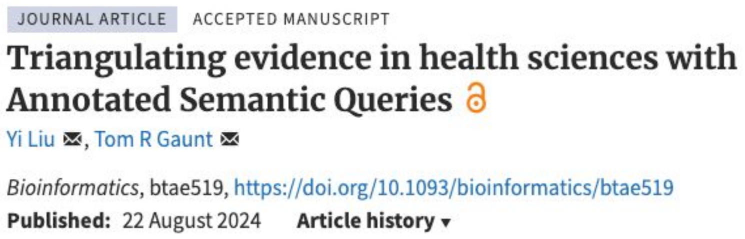
JOURNAL ARTICLE

Using language models and ontology topology to perform semantic mapping of traits between biomedical datasets

Yi Liu, Benjamin L Elsworth, Tom R Gaunt ✉ Author Notes

Bioinformatics, Volume 39, Issue 4, April 2023, btad169, https://doi.org/10.1093/bioinformatics/btad169

Published: 03 April 2023 **Article history ▾**



JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Triangulating evidence in health sciences with Annotated Semantic Queries

Yi Liu ✉, Tom R Gaunt ✉

Bioinformatics, btae519, https://doi.org/10.1093/bioinformatics/btae519

Published: 22 August 2024 **Article history ▾**

Limitations:

- Locked to a specific ontology (EFO)
- Finetuned task model not widely adaptable
- Just using naive embeddings might be preferable with the latest-and-greats of LLMs (in ASQ next gen)

Background: related works in LLM based service

Vectology

- Provides naive embedding and mapping of text with
  - BioSentVec
  - BERT
- Direct precursor to TraitHunter
- Some form of vectology is still internally accessible

epigraphdb-neural

- Pre-computes embedding of EpiGraphDB entities in a elasticsearch vector store
  - ScispaCy
- Powers the search functionalities in EpiGraphDB

Why yet another LLM based recommender service

Models

- LLaMA3 ('nuff said!)

Tooling

- Huggingface and other tools (ollama) are now much more mature
- Elasticsearch only recently supports 4095 dim dense vector embeddings

Research

- We collaborated with a few research teams on the problem of trait mapping for pre-select traits of interest
- So we probably should have a formal service and short paper up so collaborators can cite it
- In preparation for a few research projects

TraitHunter

TraitHunter is

- A publicly accessible IEU web service for
  - **Mapping traits across biomedical dictionaries**
  - Linking traits with biomedical text and literature
- Embedding and extraction of traits
  - Will replace vectology and epigraphdb-neural
  - Semi-publicly accessible with rate limit

- Web app: https://traithunter.epigraphdb.org
- API: https://traithunter-api.epigraphdb.org
- code: https://github.com/MRCIEU/traithunter

TraitHunter: concepts

Dictionary

A vocabulary set of trait entities

Regular examples:

- UK Biobank variables
- Trait names of OpenGWAS studies
- EpiGraphDB terms

Ontologies

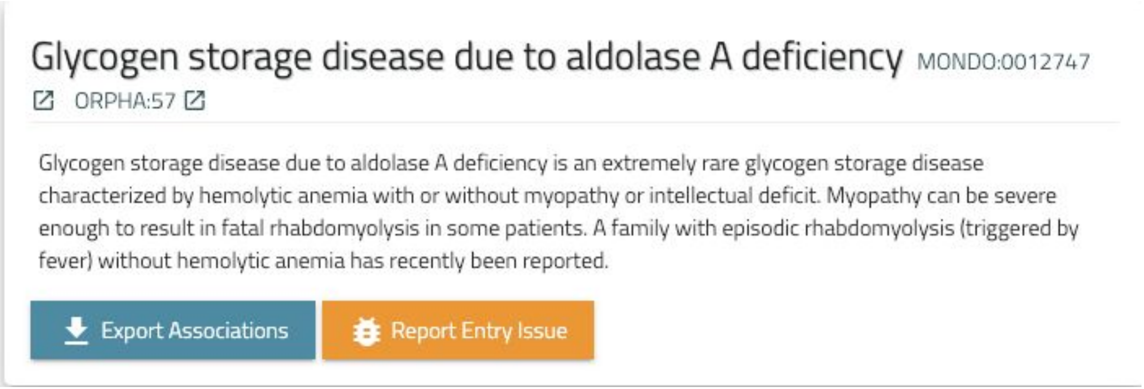
- Hierarchical from general to specific concepts

Trait entity

- ID
- Label
  - Concept singleton
  - Complex trait
- (optional) Description
- (optional) Synonyms
- (optional) Parents, children, and siblings

Full text document

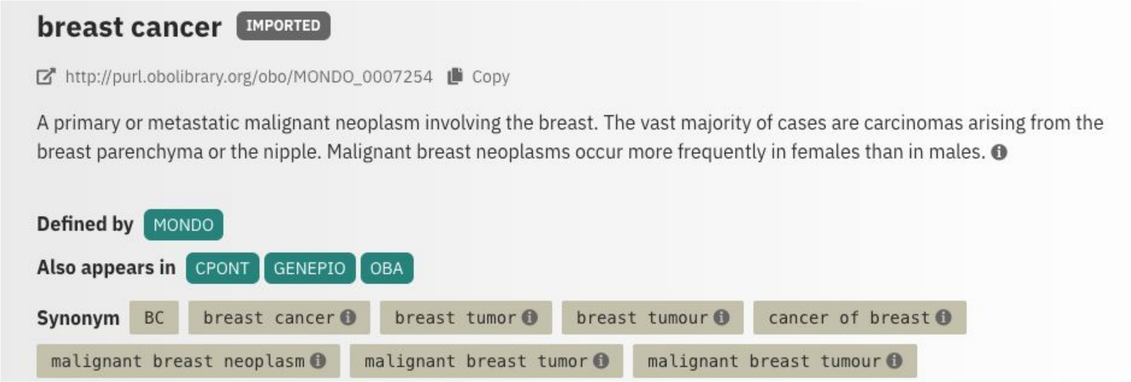
Abstracts, text descriptions, etc.



Glycogen storage disease due to aldolase A deficiency MONDO:0012747

ORPHA:57

Glycogen storage disease due to aldolase A deficiency is an extremely rare glycogen storage disease characterized by hemolytic anemia with or without myopathy or intellectual deficit. Myopathy can be severe enough to result in fatal rhabdomyolysis in some patients. A family with episodic rhabdomyolysis (triggered by fever) without hemolytic anemia has recently been reported.



breast cancer IMPORTED

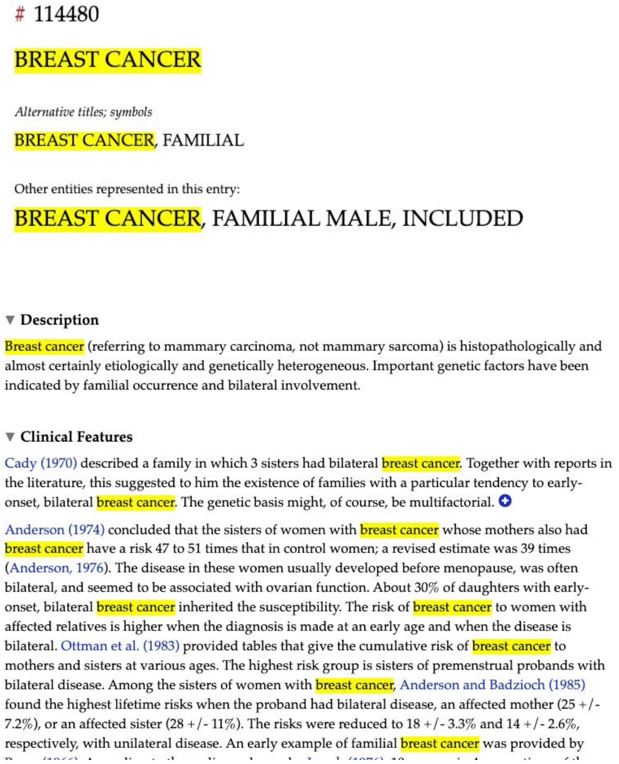
http://purl.obolibrary.org/obo/MONDO_0007254 Copy

A primary or metastatic malignant neoplasm involving the breast. The vast majority of cases are carcinomas arising from the breast parenchyma or the nipple. Malignant breast neoplasms occur more frequently in females than in males.

Defined by MONDO

Also appears in CPONT GENEPIO OBA

Synonym BC breast cancer breast tumor breast tumour cancer of breast malignant breast neoplasm malignant breast tumor malignant breast tumour



# 114480

BREAST CANCER

Alternative titles; symbols

BREAST CANCER, FAMILIAL

Other entities represented in this entry:

BREAST CANCER, FAMILIAL MALE, INCLUDED

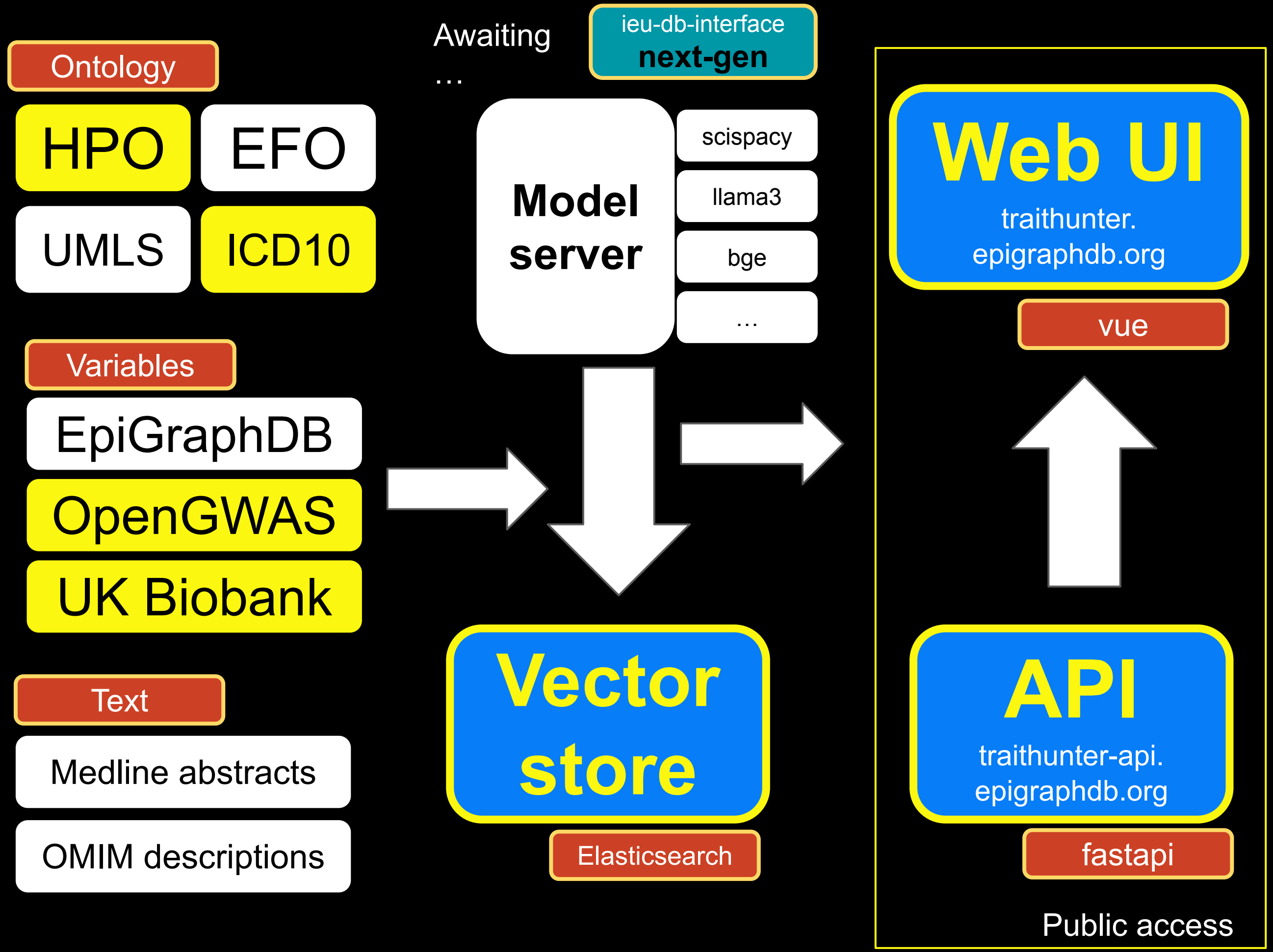
▼ Description

Breast cancer (referring to mammary carcinoma, not mammary sarcoma) is histopathologically and almost certainly etiologically and genetically heterogeneous. Important genetic factors have been indicated by familial occurrence and bilateral involvement.

▼ Clinical Features

Cady (1970) described a family in which 3 sisters had bilateral breast cancer. Together with reports in the literature, this suggested to him the existence of families with a particular tendency to early-onset, bilateral breast cancer. The genetic basis might, of course, be multifactorial.

Anderson (1974) concluded that the sisters of women with breast cancer whose mothers also had breast cancer have a risk 47 to 51 times that in control women; a revised estimate was 39 times (Anderson, 1976). The disease in these women usually developed before menopause, was often bilateral, and seemed to be associated with ovarian function. About 30% of daughters with early-onset, bilateral breast cancer inherited the susceptibility. The risk of breast cancer to women with affected relatives is higher when the diagnosis is made at an early age and when the disease is bilateral. Ottman et al. (1983) provided tables that give the cumulative risk of breast cancer to mothers and sisters at various ages. The highest risk group is sisters of premenstrual probands with bilateral disease. Among the sisters of women with breast cancer, Anderson and Badzioch (1985) found the highest lifetime risks when the proband had bilateral disease, an affected mother (25 +/- 7.2%), or an affected sister (28 +/- 11%). The risks were reduced to 18 +/- 3.3% and 14 +/- 2.6%, respectively, with unilateral disease. An early example of familial breast cancer was provided by



Ontology

HPO

EFO

UMLS

ICD10

Variables

EpiGraphDB

OpenGWAS

UK Biobank

Text

Medline abstracts

OMIM descriptions

Awaiting …

ieu-db-interface
next-gen

Model server

scispacy

llama3

bge

…

Vector store

Elasticsearch

Web UI

traithunter.
epigraphdb.org

vue

API

traithunter-api.
epigraphdb.org

fastapi

Public access

Diagram of TraitHunter web service

Explore curated traits

Trait mapping

- via trait labels
- via labels + descriptions
- incorporating synonyms

based on **LLM embeddings** and across dictionaries

Extraction and identification of traits **from biomedical literature**

data mining

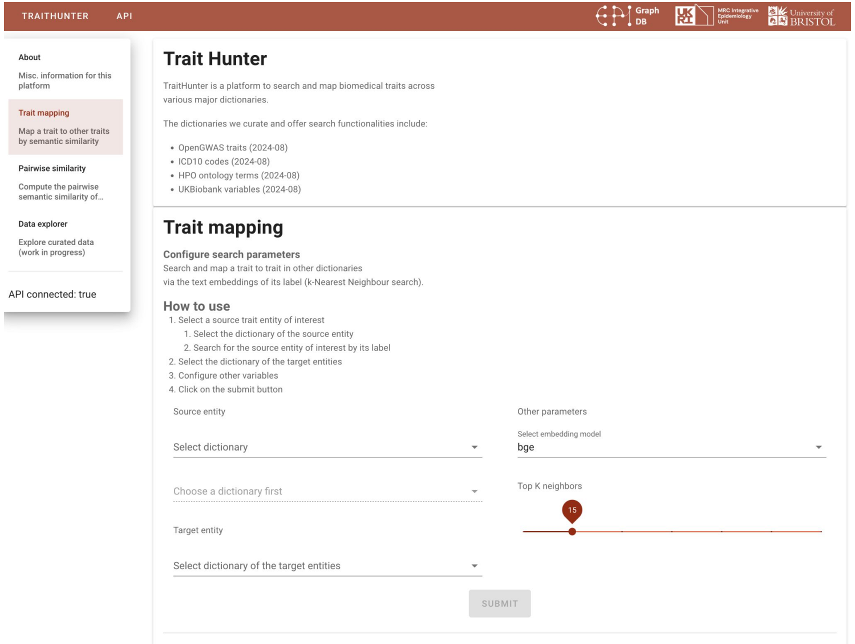
EpiGraphDB

Trait Hunter

Mapping and extraction of biomedical traits

LLM
embeddings

TraitHunter: Demo


TRAITHUNTER
API
Graph DB
MRC Integrative Epidemiology Unit
University of BRISTOL
About
Misc. information for this platform
Trait mapping
Map a trait to other traits by semantic similarity
Pairwise similarity
Compute the pairwise semantic similarity of...
Data explorer
Explore curated data (work in progress)
SUBMIT

API connected: true

Trait Hunter

TraitHunter is a platform to search and map biomedical traits across various major dictionaries.

The dictionaries we curate and offer search functionalities include:

- OpenGWAS traits (2024-08)
- ICD10 codes (2024-08)
- HPO ontology terms (2024-08)
- UKBiobank variables (2024-08)

Trait mapping

Configure search parameters

Search and map a trait to trait in other dictionaries
via the text embeddings of its label (k-Nearest Neighbour search).

How to use

1. Select a source trait entity of interest
   1. Select the dictionary of the source entity
   2. Search for the source entity of interest by its label
2. Select the dictionary of the target entities
3. Configure other variables
4. Click on the submit button

Source entity

Select dictionary

Choose a dictionary first

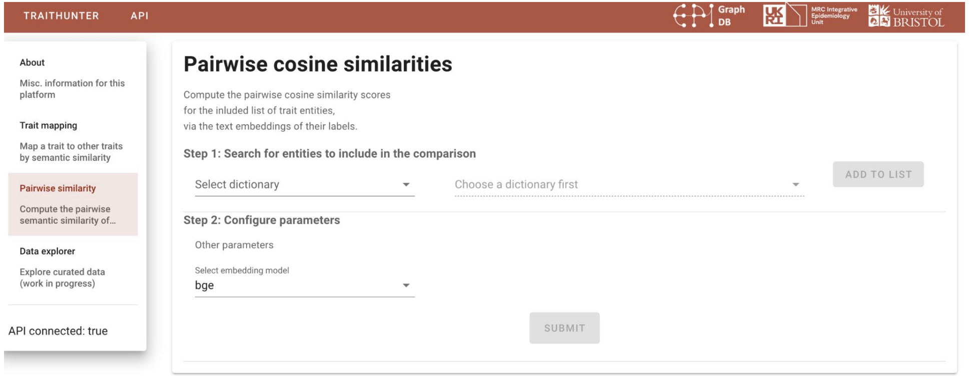
Target entity

Select dictionary of the target entities

Other parameters

Select embedding model: bge

Top K neighbors: 15


TRAITHUNTER
API
Graph DB
MRC Integrative Epidemiology Unit
University of BRISTOL
About
Misc. information for this platform
Trait mapping
Map a trait to other traits by semantic similarity
Pairwise similarity
Compute the pairwise semantic similarity of...
Data explorer
Explore curated data (work in progress)
ADD TO LIST
SUBMIT

API connected: true

Pairwise cosine similarities

Compute the pairwise cosine similarity scores
for the inluded list of trait entities,
via the text embeddings of their labels.

Step 1: Search for entities to include in the comparison

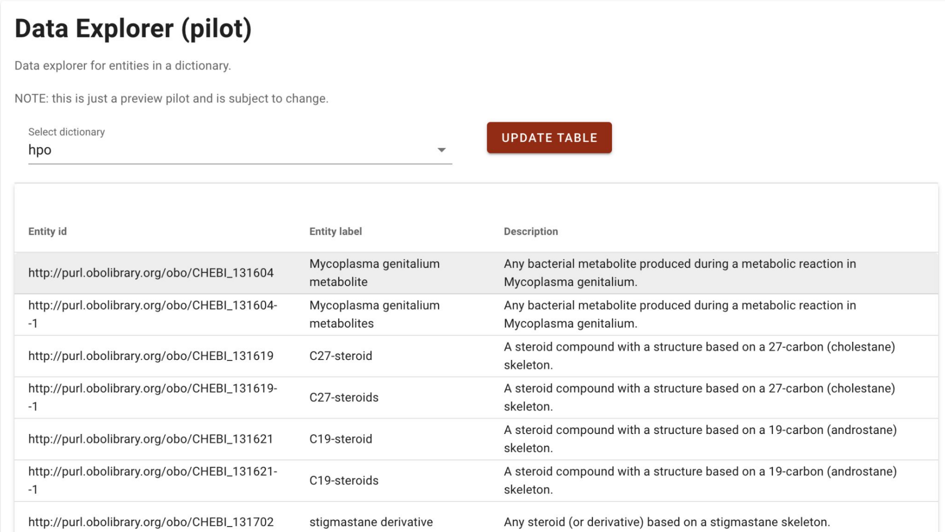
Select dictionary

Choose a dictionary first

Step 2: Configure parameters

Other parameters

Select embedding model: bge



Data Explorer (pilot)

Data explorer for entities in a dictionary.

NOTE: this is just a preview pilot and is subject to change.

Select dictionary: hpo

Entity id	Entity label	Description
http://purl.obolibrary.org/obo/CHEBI_131604	Mycoplasma genitalium metabolite	Any bacterial metabolite produced during a metabolic reaction in Mycoplasma genitalium.
http://purl.obolibrary.org/obo/CHEBI_131604-1	Mycoplasma genitalium metabolites	Any bacterial metabolite produced during a metabolic reaction in Mycoplasma genitalium.
http://purl.obolibrary.org/obo/CHEBI_131619	C27-steroid	A steroid compound with a structure based on a 27-carbon (cholestane) skeleton.
http://purl.obolibrary.org/obo/CHEBI_131619-1	C27-steroids	A steroid compound with a structure based on a 27-carbon (cholestane) skeleton.
http://purl.obolibrary.org/obo/CHEBI_131621	C19-steroid	A steroid compound with a structure based on a 19-carbon (androstane) skeleton.
http://purl.obolibrary.org/obo/CHEBI_131621-1	C19-steroids	A steroid compound with a structure based on a 19-carbon (androstane) skeleton.
http://purl.obolibrary.org/obo/CHEBI_131702	stigmastane derivative	Any steroid (or derivative) based on a stigmastane skeleton.

Next steps, in infra development

Assessment of embedding quality with benchmarks

- How well does an embedding model objectively perform for the task of mapping biomedical traits

Curation of full text into the vector store

- OMIM descriptions
- Medline abstracts

Named entity recognition

- Identification and extraction of a trait entity from full text docs.

Molecular symbols

- Special treatment for gene names, etc.

Next steps, in research

The TraitHunter paper

- Describing the service and methods

Your input is needed on what this platform can help with your research.

Clustering analysis

Investigation on the different clusters of biomedical relationships for novel insights

- Semantics
- Genetic associations

RAG

RAG on knowledge graphs to investigate about biomedical assertions (i.e. ASQ-next-gen).

Acknowledgements

IEU

- Tom Gaunt
- Zhaozhen Xu
- Gibran Hemani
- *(alumni* Benjamin Elsworth*)*

University of Bristol

- Naomi Cornish
- Andrew Mumford

Million Veteran Program

- Brian Ferolito
- Daniel Golden
- Alexandre Pereira