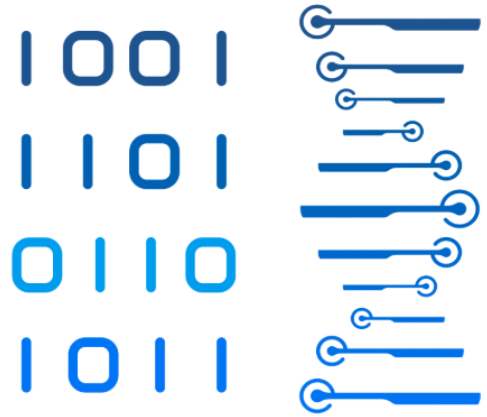
XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2026)

Emerging Challenges and Opportunities for In Silico Drug Discovery

19-21 October, 2026 Virtual



LITBIOME: AN INTEGRATED WEB SERVICE FOR BIOMEDICAL TEXT MINING

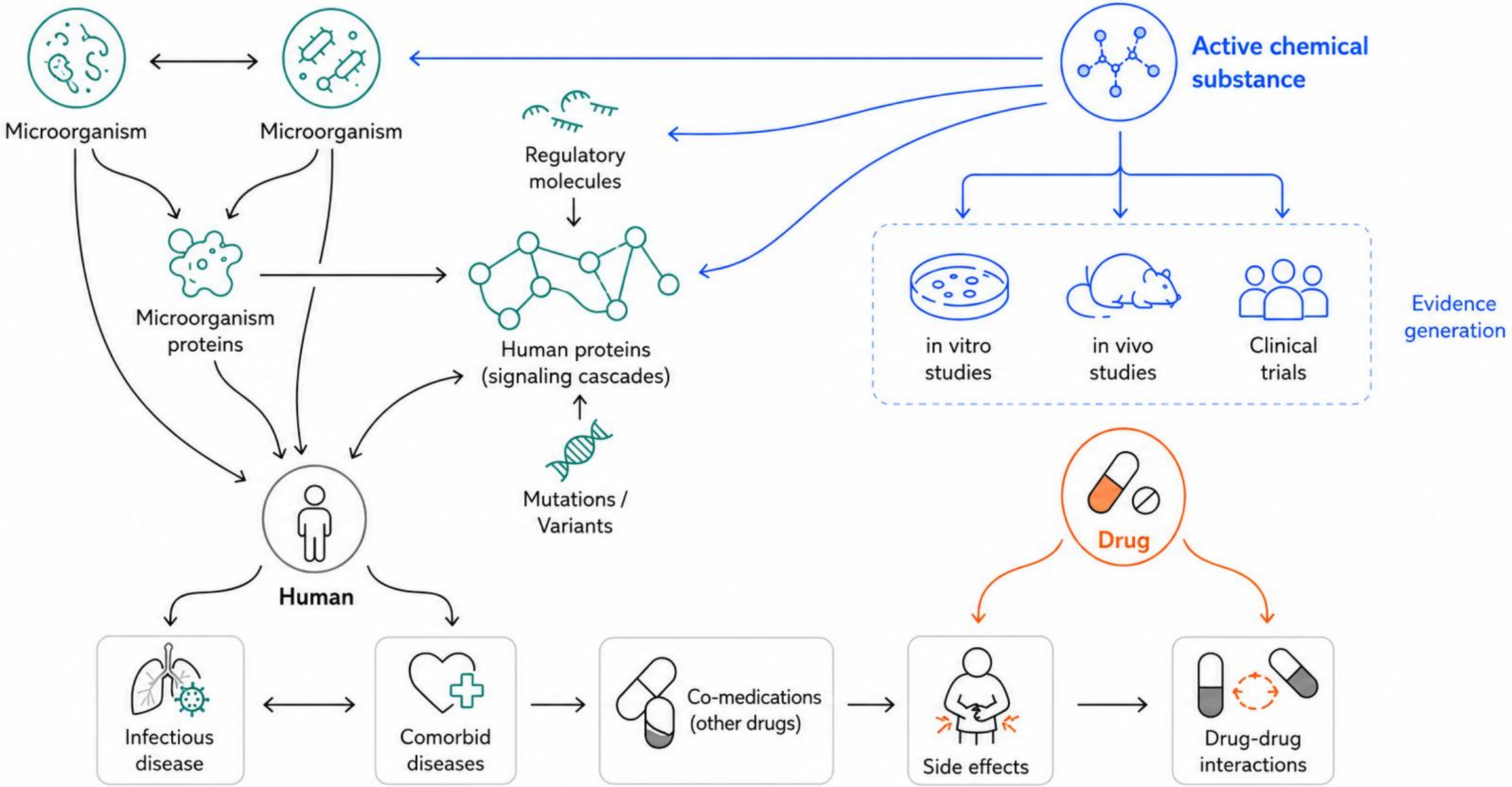
Nadezhda Biziukova

Laboratory of Big Data Analysis for Digital Pharmacology

Institute of Biomedical Chemistry, Moscow, Russia

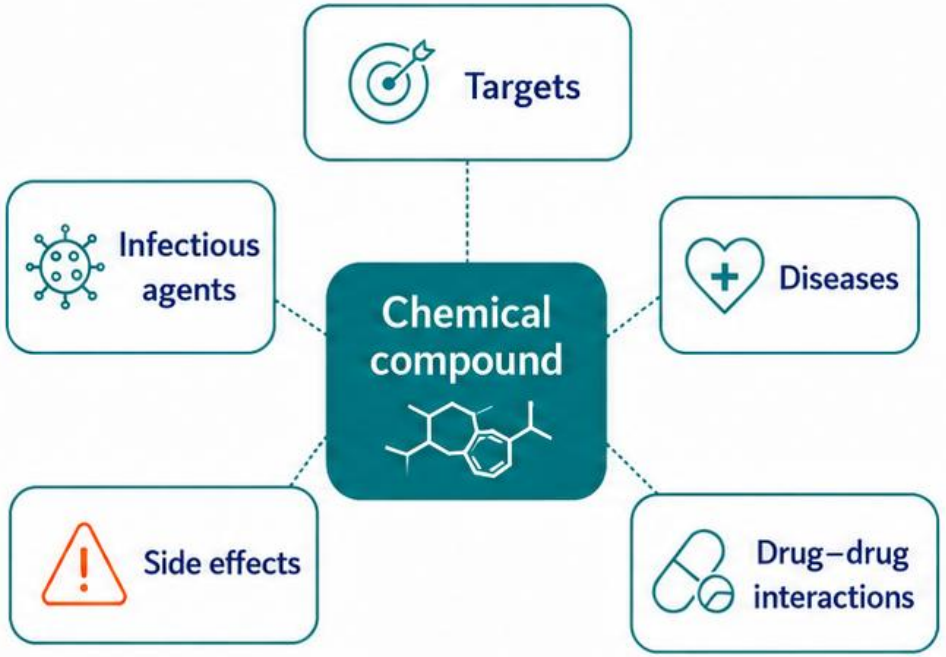
IBMC Institute of Biomedical Chemistry

nad.smol@gmail.com

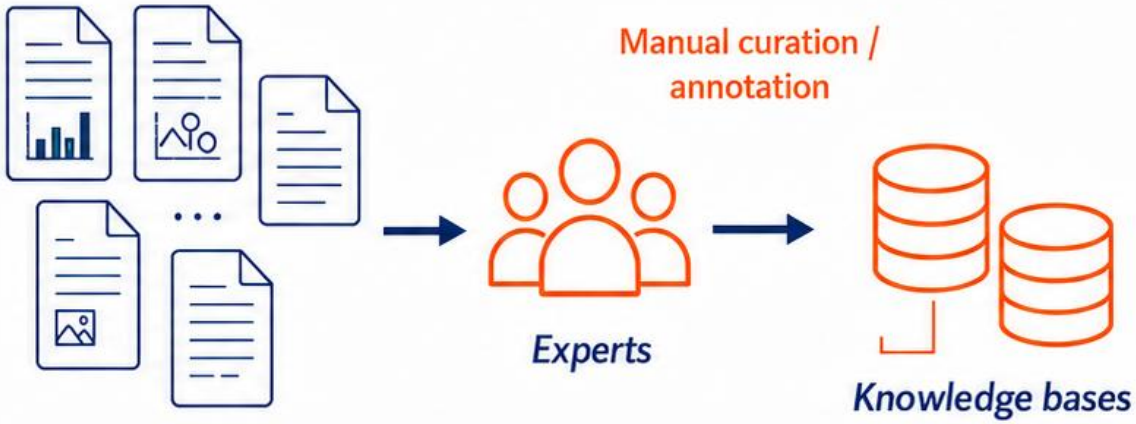

Microorganism
Microorganism
Active chemical substance
Regulatory molecules
Microorganism proteins
Human proteins (signaling cascades)
in vitro studies
in vivo studies
Clinical trials
Evidence generation
Mutations / Variants
Drug
Human
Infectious disease
Comorbid diseases
Co-medications (other drugs)
Side effects
Drug-drug interactions

Knowledge needed for safe and effective drug development

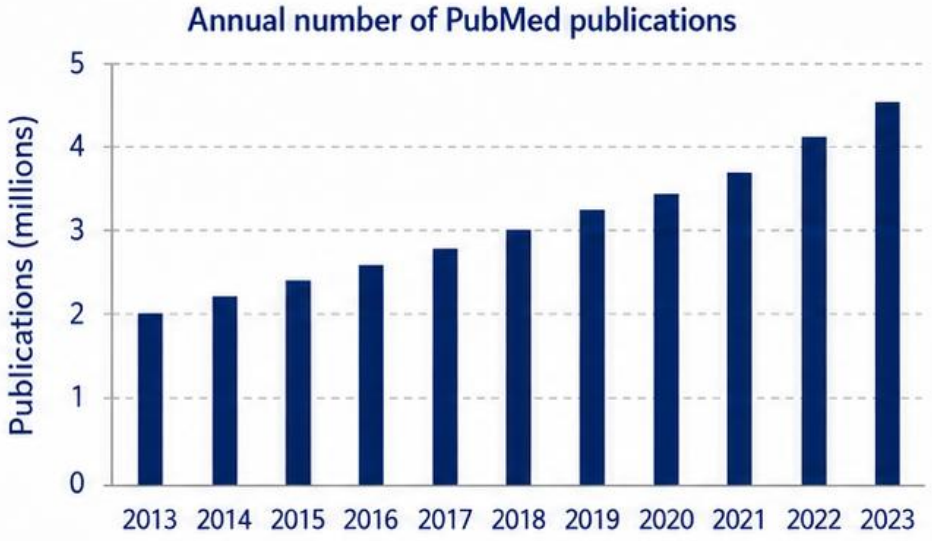
including repositioning opportunities



Main source:
scientific publications



The exponential growth of literature makes manual review unfeasible



As a result, manually curated knowledge bases are often **incomplete and outdated.**

Integrating this diverse knowledge can accelerate drug development and especially support drug repositioning.

1. Entity annotation / semantic search

Example: PubTator 3.0

- AI-powered annotations and relation-aware search across PubMed / PMC
- Strong NER / normalization in a precomputed resource

⚠ Limited custom knowledge-graph construction

2. Rule-based and hybrid relation extraction

Examples: SemRep / SemMedDB

- UMLS-based subject–predicate–object predications
- Broad relation types and factuality assignment

⚠ Limited recall; UMLS-dependent; mostly sentence-level

3. Knowledge assembly / KG construction

Examples: INDRA, BioKGrapher, GePI

- Mechanistic statement assembly and reliability scoring
- Knowledge-graph construction or molecular interaction mining

⚠ Often molecular- or condition-specific

4. LLM-based KG construction

Examples: EDC, AutoBioKG, RELATE, GENET, MDKG

- Context-aware triples and ontology-constrained relations
- Promising for automated KG construction

⚠ Often experimental, domain-specific, or dependent on large / closed LLMs

Goal: to develop an integral text mining algorithm for extracting comprehensive information on the biological activity of chemical compounds from scientific publications.

1 Selection of relevant PubMed publications



- MeSH-based queries
- Domain-specific text classification
- Full-text preprocessing

2 Named entity recognition



- Chemicals
- Proteins / genes
- Diseases
- Species
- Variants and miRNAs (rule-based)

HunFlair 2 + regular expressions

3 Relation extraction



- Trigger-based associations
- Sentence-level scoring
- Corpus-level confidence aggregation

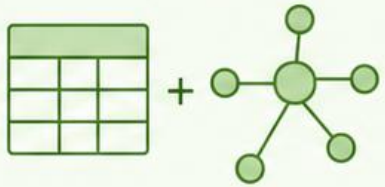
association confidence

4 Normalization & aggregation



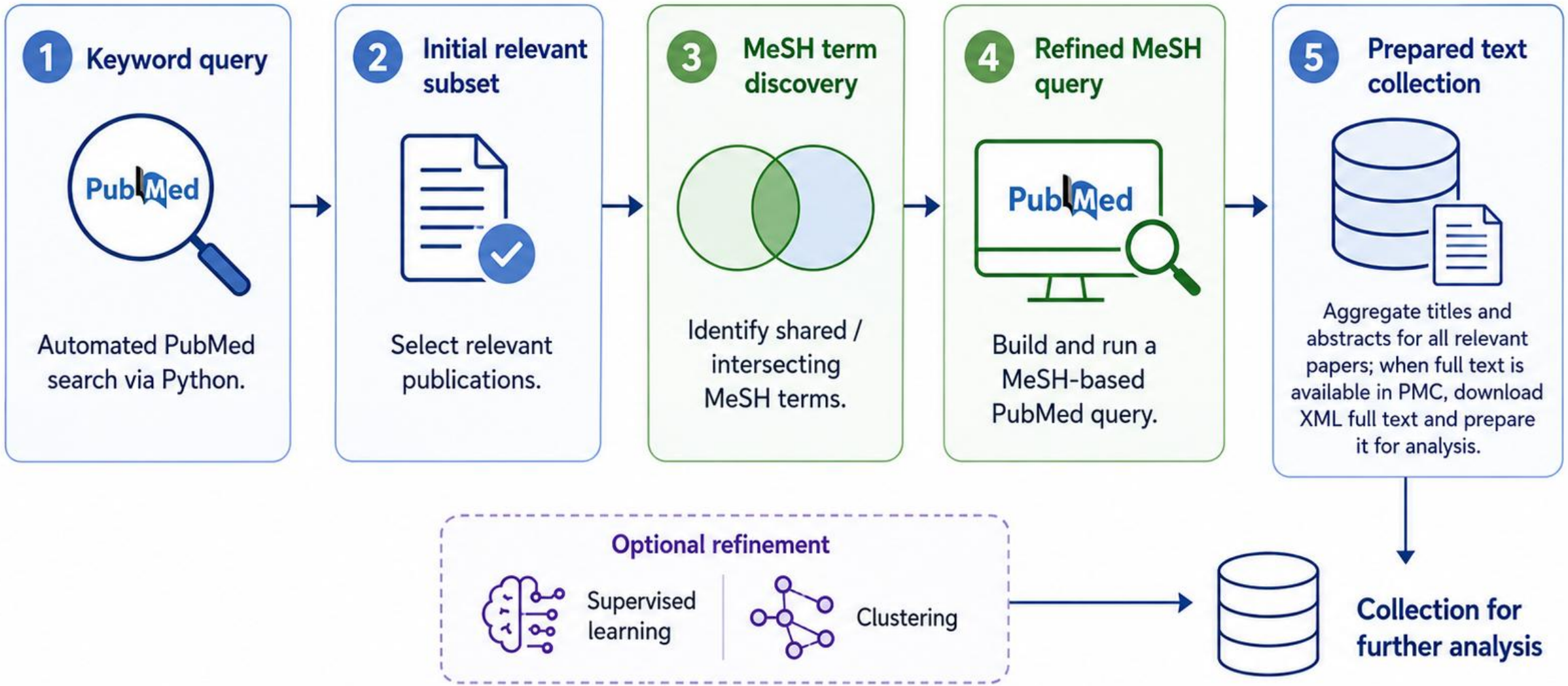
- Exact and fuzzy matching
- Database linking by entity type
- BERT-based similarity
- Abbreviation resolution

5 Visualization

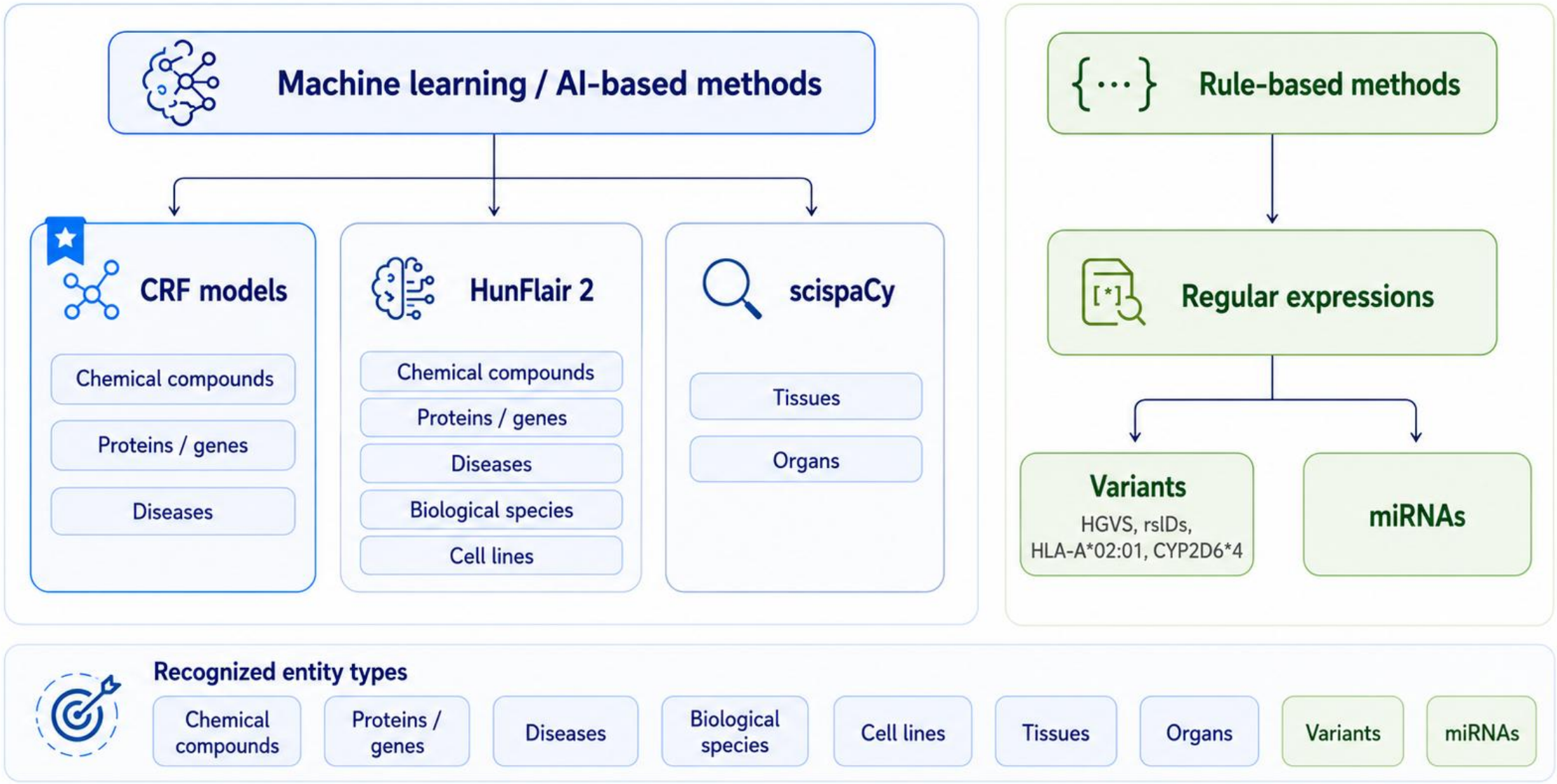


- Tabular view
- Network view
  - Nodes: mention, DB ID, preferred name, entity type
  - Edges: trigger, relation type, PMID, confidence

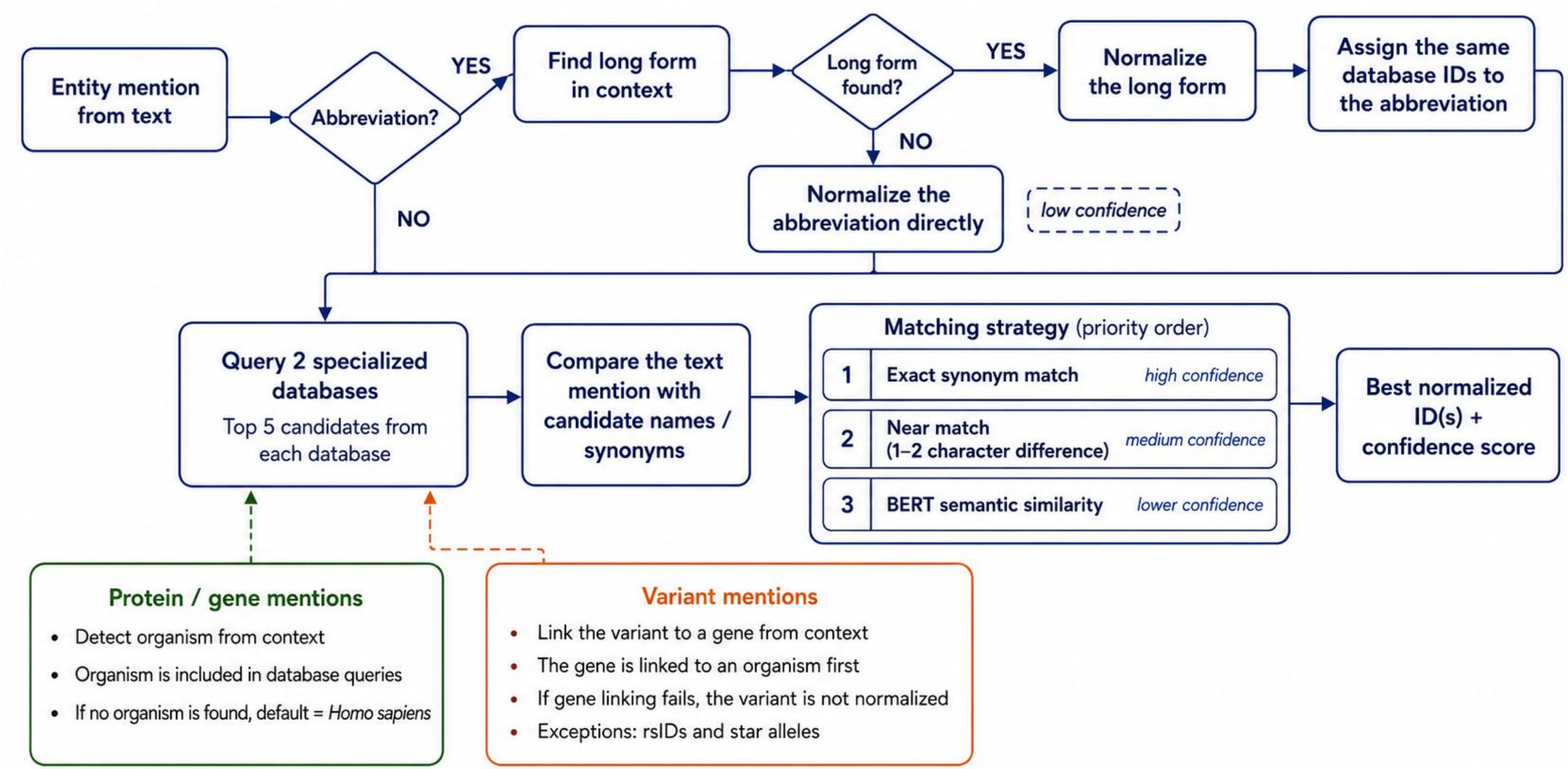
Text collection for further analysis



Named entity recognition



Named entity normalization



Relation extraction

However, increased CD8 cell count may be a moderate risk factor for stroke.



Furthermore, we document chikungunya virus (CV) as a possible triggering agent for DRESS.



- Treatment ~ regimen, administration, receive
- Prevention ~ prophylaxis
- Due to ~ lead to, trigger, etiology, cause (in case of Species-Disease associations)

- 77 groups
- 233 common concepts
- 726 common terms

Associated objects	Group (nature)	Common meaning (concept)	Common term	Pattern phrases
Chemical-Disease	Side effects	Leads to	Induce	Induce, induces, induced, …
			Lead to	Lead, leads, leading, lead to, …
		Increases risk of	Risk factor	Risk, risks, risk factor, risk factors
	Application in therapy	Used in therapy of	Administer	Administer, administered, administration
			Regimen	Regimen, regimens
Chemical-Protein/Gene	Effects on protein/gene	Positive regulation of	Induce	Induce, induces, induced, inducing, induction
		Negative regulation of	Inhibit	Inhibit, inhibits, inhibited, inhibiting, …

Relation extraction

1. Multiple candidate associations in one sentence

... the interaction of the HMG-CoA reductase inhibitor rosuvastatin with the renal anemia drug vadadustat increases the blood concentration of rosuvastatin in vitro.

Vadadustat - inhibitor - HMG-CoA reductase ✖

rosuvastatin - inhibitor - HMG-CoA reductase ✓

Shorter distance – higher probability

2. Sentence-level ranking

$$S_{sent} = \frac{w_p}{1+d}$$

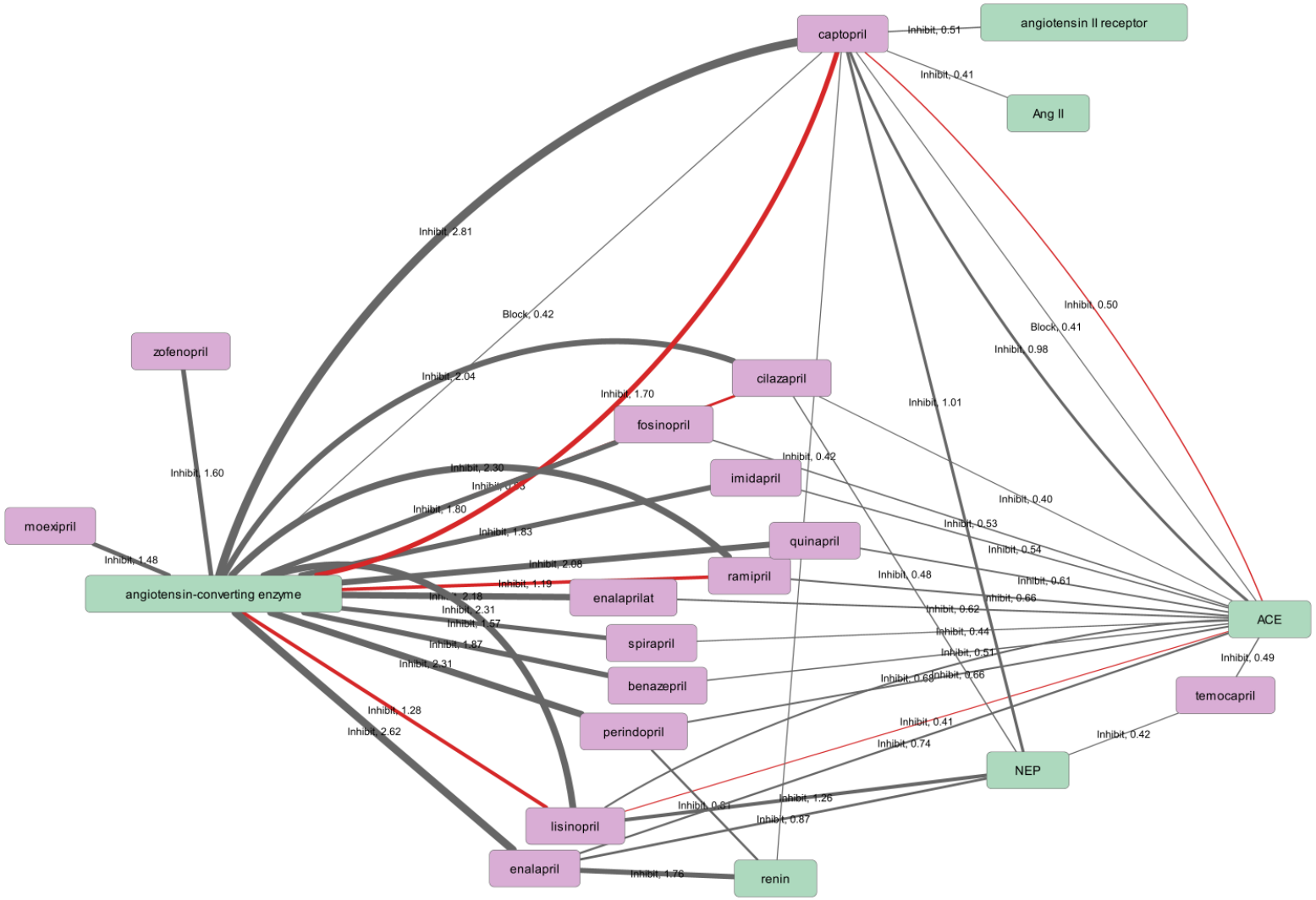
- w_p - trigger strength (3 – strong, 2 – medium, 1 – low)
- d – distance between pattern phrase and entities (in tokens)

3. Collection-level ranking

Confidence score

- Publication type
  - *Clinical trial*
  - *In vivo*
  - *In vitro*
- Association mention frequency in the collection
- Normalization confidence
- Sentence-level ranking

Application



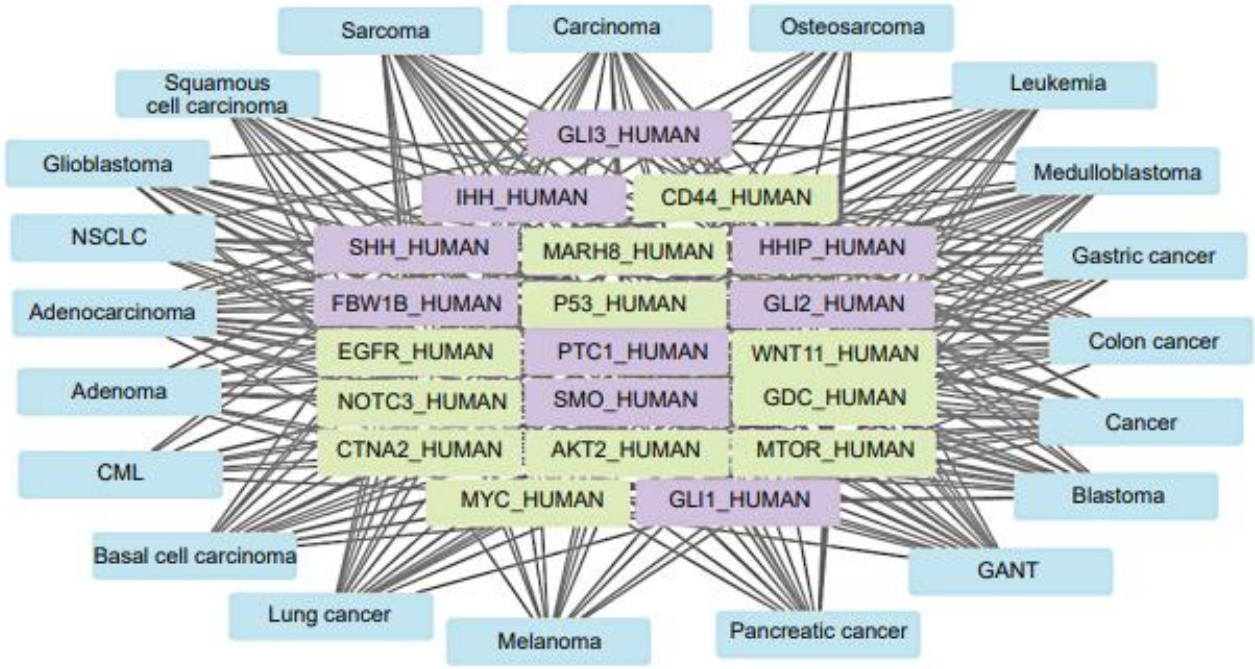
Mechanisms and effects of ACE inhibitors

- >20 000 of titles and abstracts, dedicated to known ACE inhibitors
- Relation types: mechanisms (chemical-gene), effects (chemical-disease), co-prescriptions (chemical-chemical)

Key findings

- The method recovered the canonical ACE inhibitor mechanism + RAS-related context
- Repurposing-relevant signals: nephroprotective effects, antifibrotic activity, antitumor mechanisms, antidepressant effects.

Application



The role of Heghehog pathway in neoplastic diseases

- >10 000 of titles and abstracts, dedicated to role of Hedgehog pathway in neoplastic diseases
- Relation types: regulation on protein level (gene-gene), involvement in particular neoplastic diseases (gene-disease)

Key findings

- 97% of genes matched differentially expressed genes in OncoDB.
- Overlap counts were statistically significant across cancer types (χ^2, $p < 0.05$).
- Text-mined proteins formed a dense PPI subnetwork around the Hedgehog pathway.

Biziukova N. et al., Big Data Mining and Analytics, 2024

Web-interface

Associations

PMID 41695673 Transcriptomic profiling of porcine duodenal, jejunal, and i...

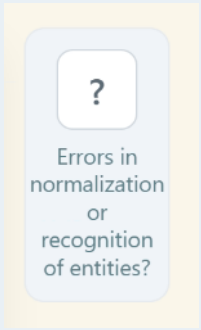
Relation ID	Sentence	Interaction rule	Confidence (task)	Confidence (DB)
11	Porcine epidemic diarrhea virus (PEDV) is a highly pathogenic virus that causes severe gastrointestinal disease in neonatal piglets, often leading to high mortality.	Cause of the disease	0.803	0.759
12	Porcine epidemic diarrhea virus (PEDV) is a highly pathogenic virus that causes severe gastrointestinal disease in neonatal piglets, often leading to high mortality.	Cause of the disease	1.000	1.000
11	Porcine epidemic diarrhea virus (PEDV) is a highly pathogenic virus that causes severe gastrointestinal disease in neonatal piglets, often leading to high mortality.	Main disease-symptom	0.803	0.811
13	Further analysis identified key upregulated genes involved in immune modulation, inflammation regulation, and tissue integrity, such as SLIT2, MMD2, and PKHD1, along with downregulated genes, including IL-1A, MMP13, and GNA15, that help control inflammation and minimize tissue damage.	Impact on protein/gene	0.718	0.692

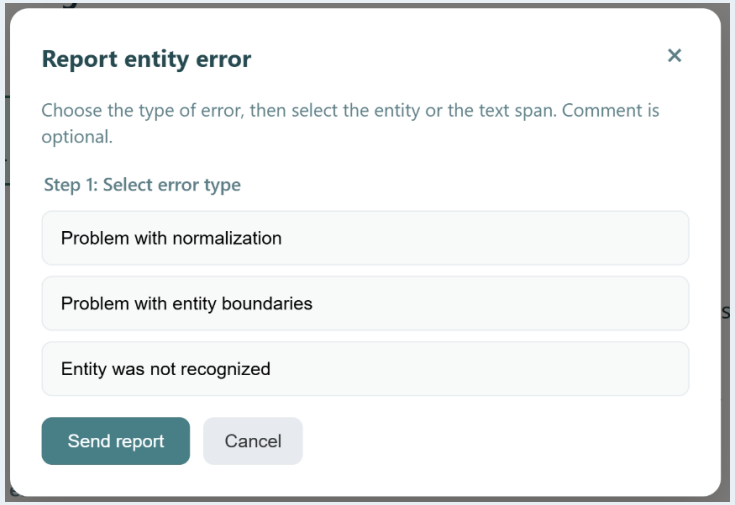
LitBiome

https://way2drug.com/textomics/index.php

Updates

User feedback



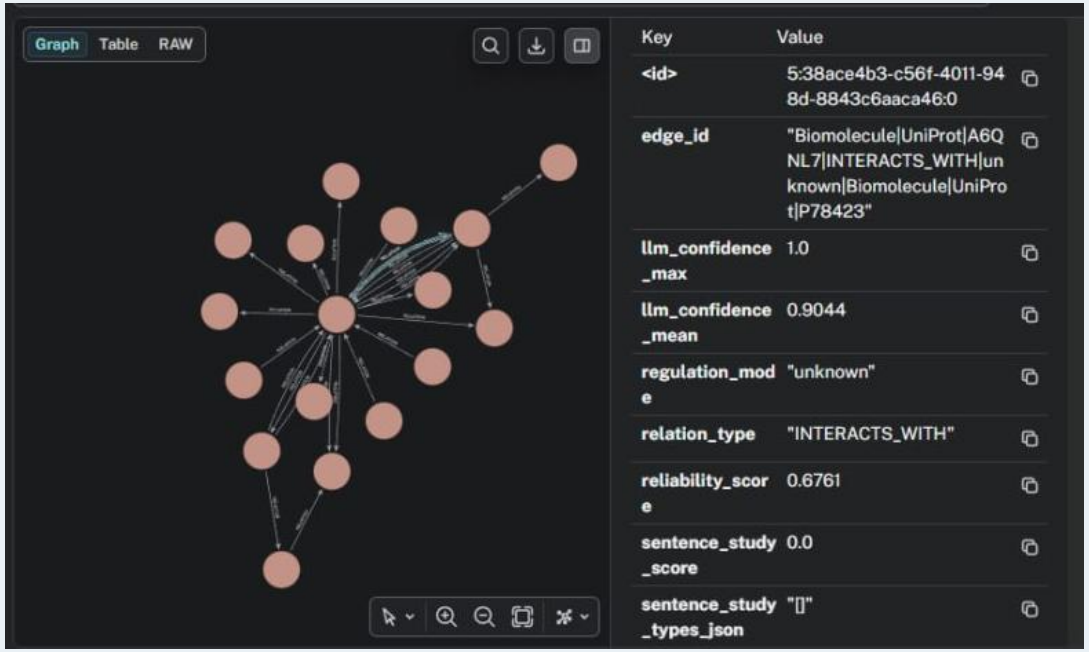


Normalization errors

Click a row to open the correction form: enter the correct identifier in the same DB and submit.

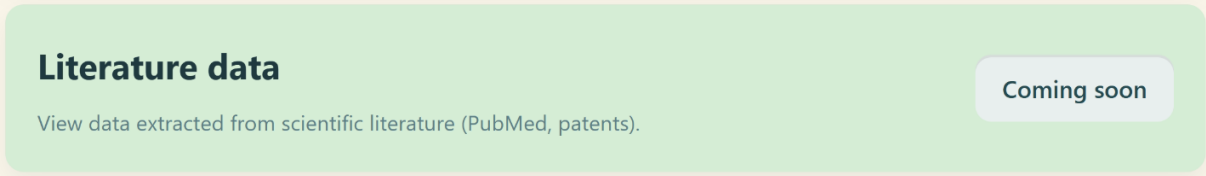
ID	Task	Text	Type	Surface / selected	Comment	Reported by (name)	Reported by (email)
1	16	16	normalization	VKORC1		Nadezhda Biziukova	nad.smol@ibmc.msk.ru

Knowledge graphs instead of flat html-graphs



Neo4j

Option to explore already analyzed text, included in the database



Thank you for your attention!

Acknowledgments

- Olga A. Tarasova, Anastasia V. Rudik, Vladimir V. Poroikov, Dmitry A. Filimonov, *Institute of Biomedical Chemistry, Moscow, Russia*