


iASiS Open Data Graph: Automated Semantic Integration of Disease-Specific Knowledge

Anastasia Krithara



Introduction

01

iASiS Open Data Graph

02

Web Interface for the iASiS ODG

03

Drug-Drug Interaction Prediction

04

Error Detection

05

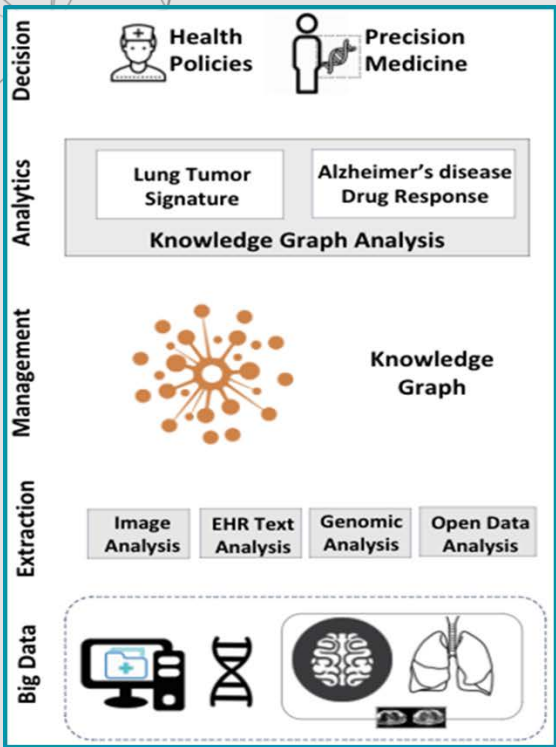


Conclusions & Future Work

06

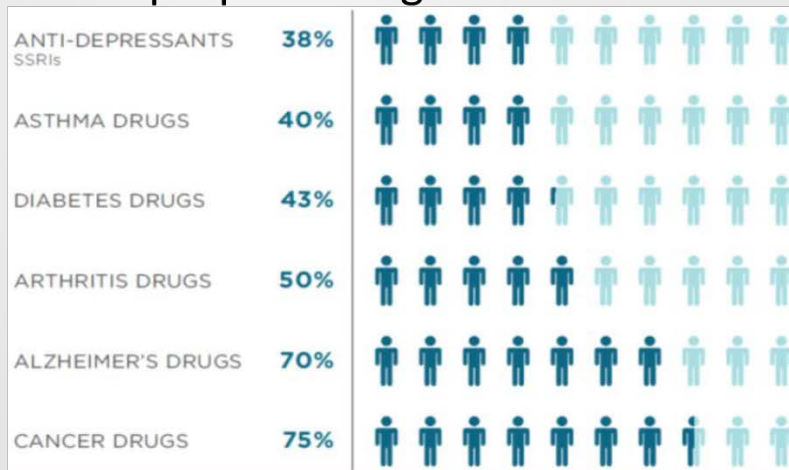
**Presentation
outline**

The iASiS framework

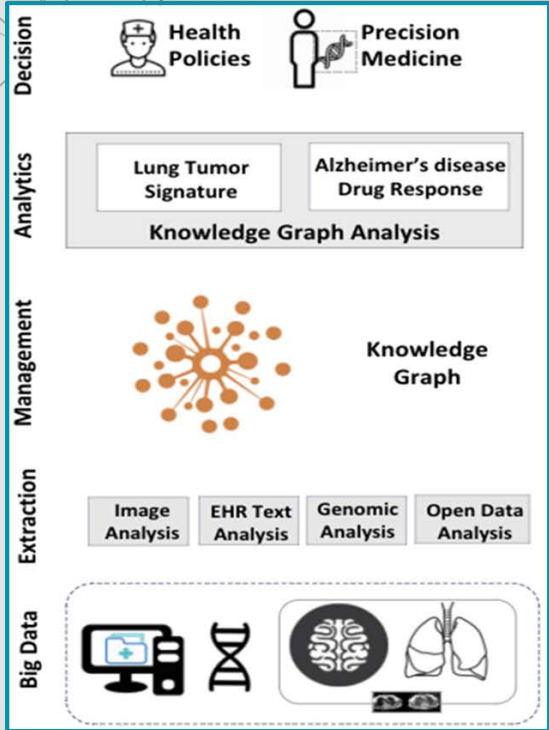


- iASiS focuses on two use cases:
 - Lung cancer
 - Alzheimer's disease

- General-purpose drugs are often ineffective



The iASiS framework



- iASiS analyses:
 - EHRs (English & Spanish)
 - MRI & PET/CT images
 - Genomic data (e.g. liquid biopsy samples)
 - Related bibliography (e.g. PubMed)
 - Biomedical databases (e.g. DrugBank)
 - Biomedical ontologies (e.g. GO, UMLS)

Available Data Sources

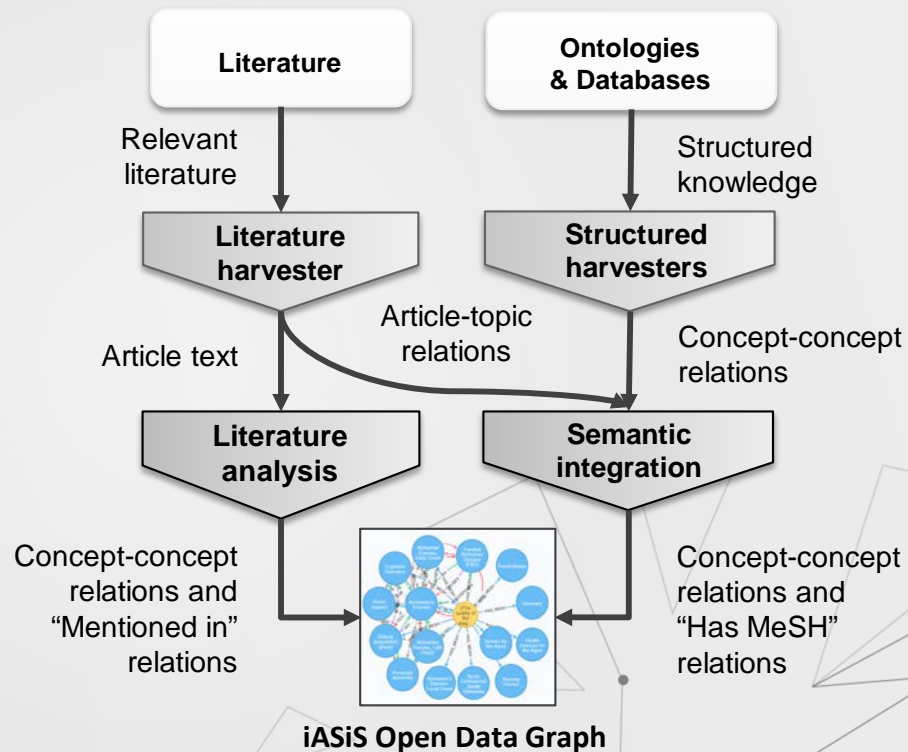
Openly available Biomedical Data sources

- **Biomedical Literature**
 - MEDLINE - 30 million journal citations
 - PubMed Central - over 6 million full-text records
- **Structured Resources**
 - Ontologies - over 900 only in BioPortal
 - Databases



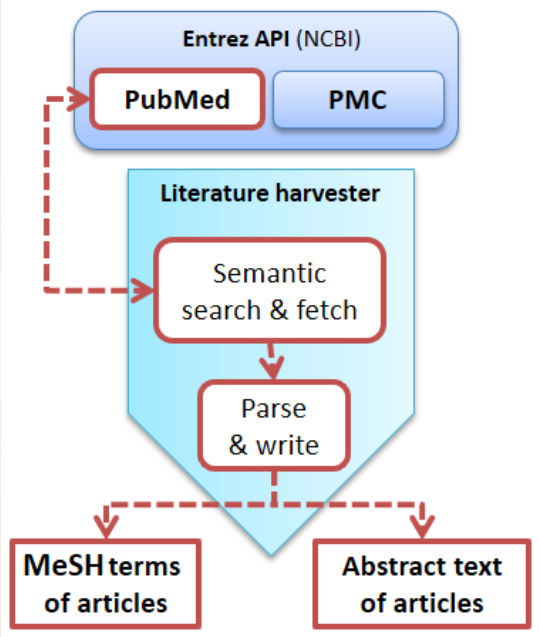
The iASiS Open Data Graph (iODG)

- An **open-source framework*** to combine available state-of-the-art tools
- **Automated** semantic retrieval and **integration** of disease-specific knowledge



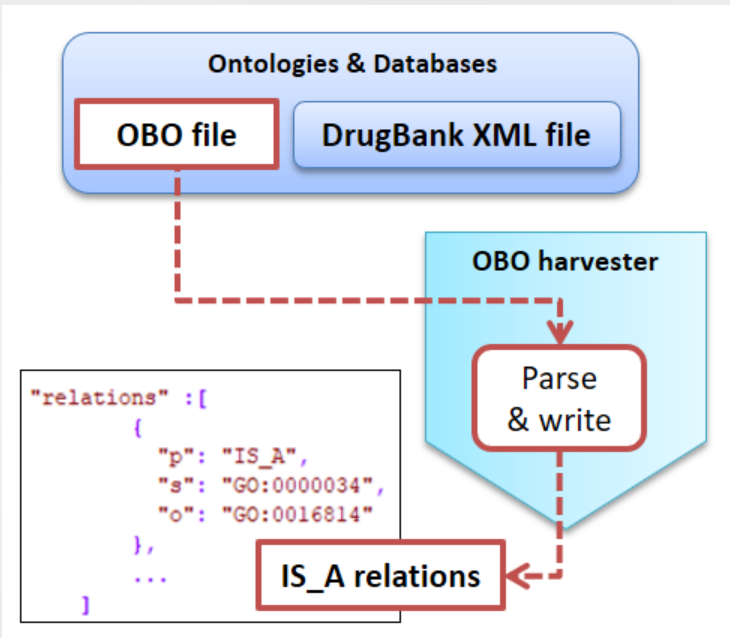
Harvesting Literature

- **Online semantic retrieval** of disease-specific literature
- **PubMed**: Abstract and MeSH topic annotations for each article
- **PubMedCentral (PMC)**: Article full text (when available)
- **Incremental** update




Harvesting Structured Resources

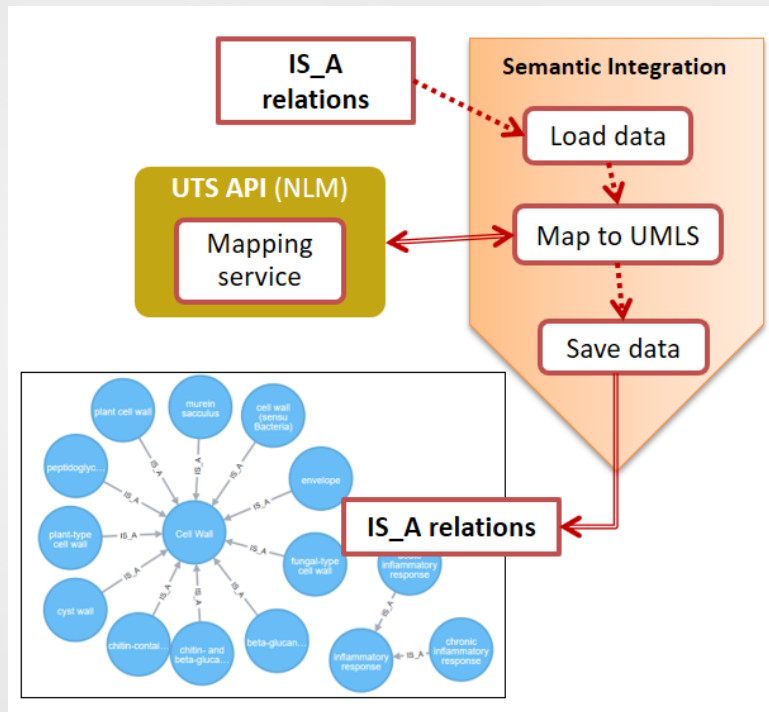
- Extract information of interest as relations between entities into a simple JSON-based representation
- Open Biomedical Ontologies (OBO): hypernymic relations (is-a)
- DrugBank: Drug-to-drug integrations



Semantic Integration (1/2)

Structured resources

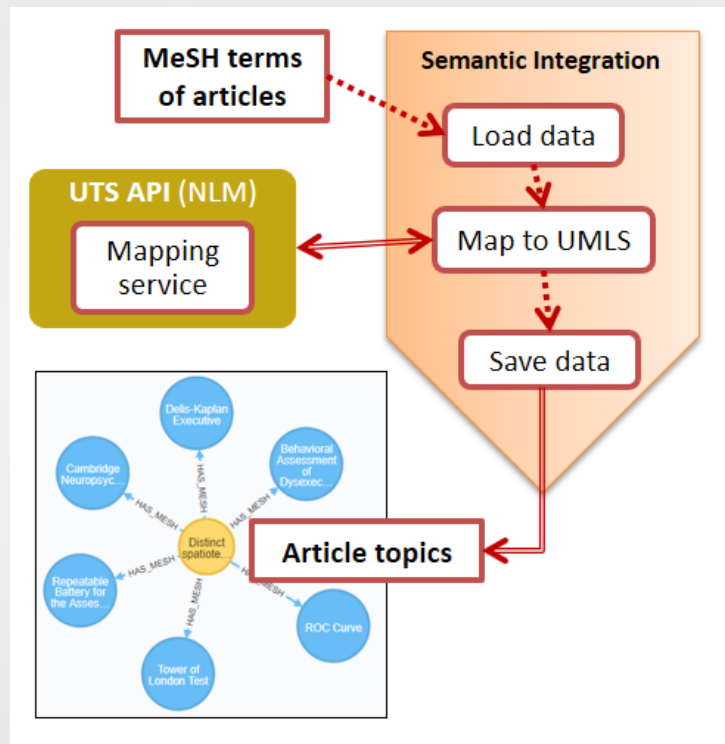
- Integrate harvested relations under a common schema
- Map entities to Unified Medical Language System (UMLS) concepts 
- Store as relations between nodes, based on the UMLS semantic network (SN)




Semantic Integration (2/2)

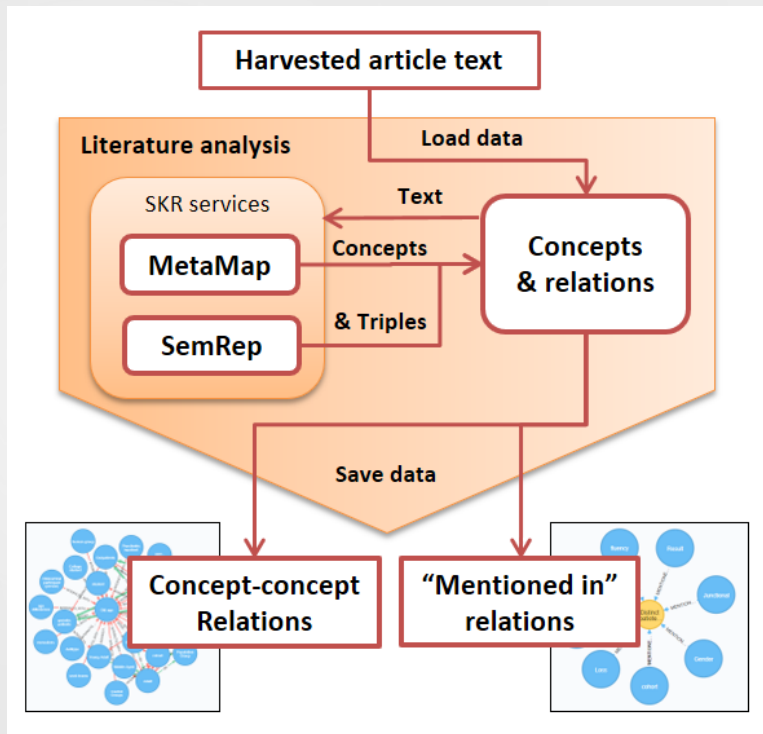
Disease-specific Literature

- Integrate **MeSH** topic annotations under the same common schema
- Map MeSH labels to **UMLS** concepts
- Add **articles** as nodes and “**Has MeSH**” relations connecting them with the nodes of the labels



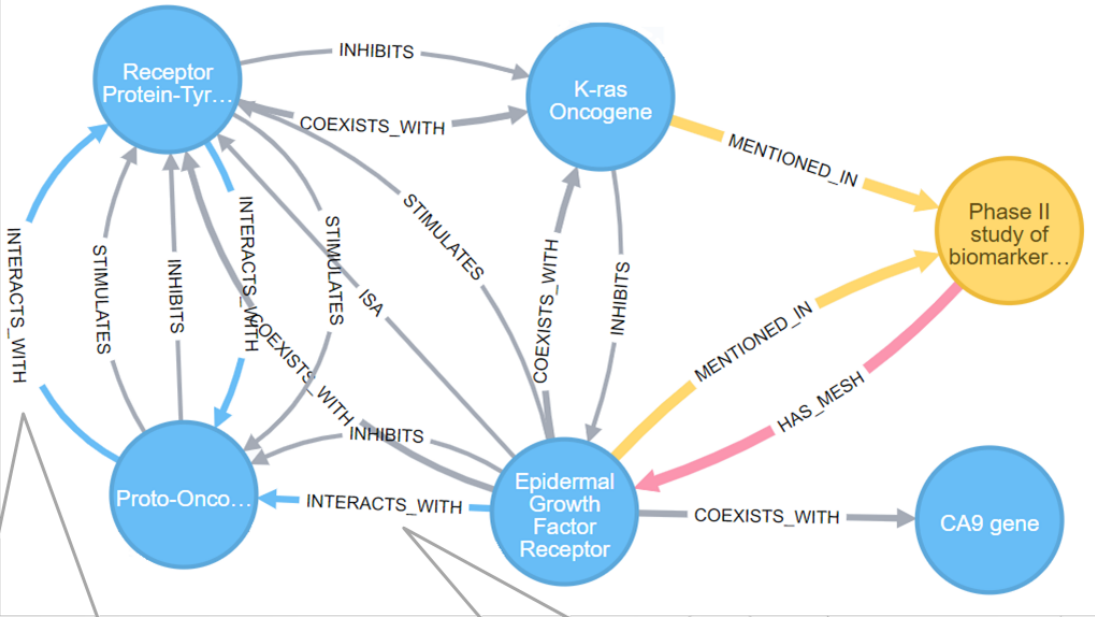
- Extract UMLS concepts with MetaMap and UMLS SN relations between them with SemRep 
- Integrate extracted UMLS SN relations under the **common schema** in the graph
- Add “**Mentioned In**” relations connecting UMLS concepts with corresponding articles

Literature Analysis



The Integrated Graph

- Detailed provenance information is kept for each relation
- The graph is incrementally updated when new knowledge is harvested from any resource



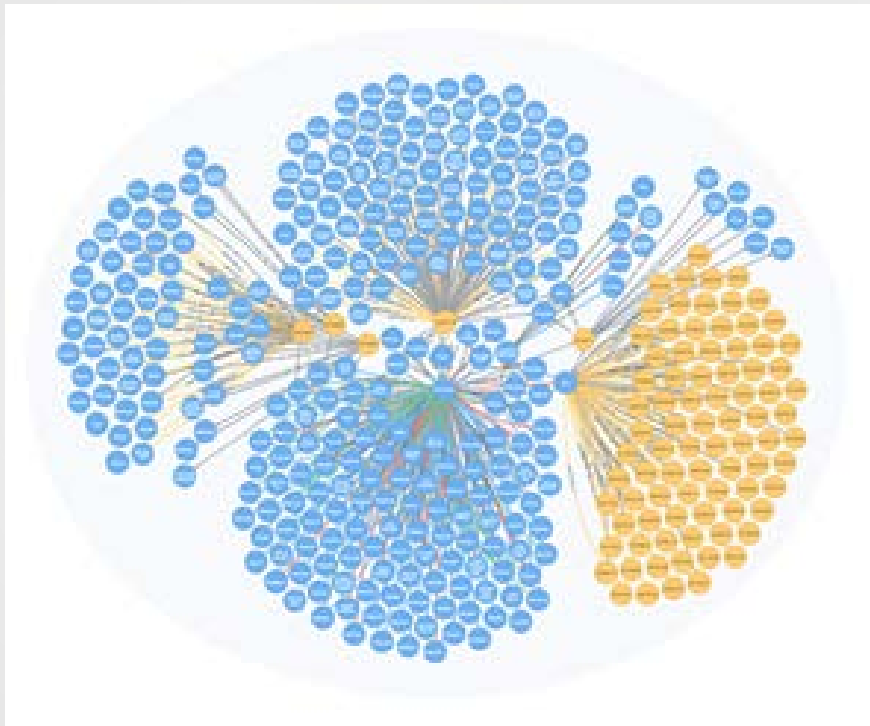
resource: text, text
 sent_id: 21215704_abstract_2,
 17322284_abstract_7

resource: text
 sent_id: 23270470_abstract_1

Using an iASiS Open Data Graph

An **intuitive** UMLS-based **graph** representation accessible by

- **Domain experts** using interactive visualization interfaces such as Neo4J Bloom
- **Knowledge analysis** studies and applications through graph queries, as automatically developed and updated datasets



Case Studies

Three distinct disease-specific **data graphs** were developed from the literature for **three case studies**

Case Study	Duchene Muscular Dystrophy (DMD)	Alzheimer's Disease and Dementia (ADD)	Lung Cancer (LC)
Articles	4,403	108,458	141,712
Articles with full-text	1,075	6,000	10,000
UMLS concepts extracted	21,982	75,985	92,846
UMLS SN relation extracted	27,954	392,421	608,759
"Has MeSH" relations	113,136	3,651,698	4,228,785
"Mentioned in" relations	335,071	7,587,772	9,940,847

Structured resources

Knowledge harvested from four **structured** resources was also integrated in each disease-specific data graph.

Structured resource	UMLS concepts	Relation type	Distinct relations
Disease Ontology (DO)	5,307	Hypernymic (“is a”)	5,129
Gene Ontology (GO)	64,751	Hypernymic (“is a”)	125,629
MeSH hierarchy	55,400	Hypernymic (“is a”)	123,287
DrugBank	3,642	Drug interactions	1,628,077

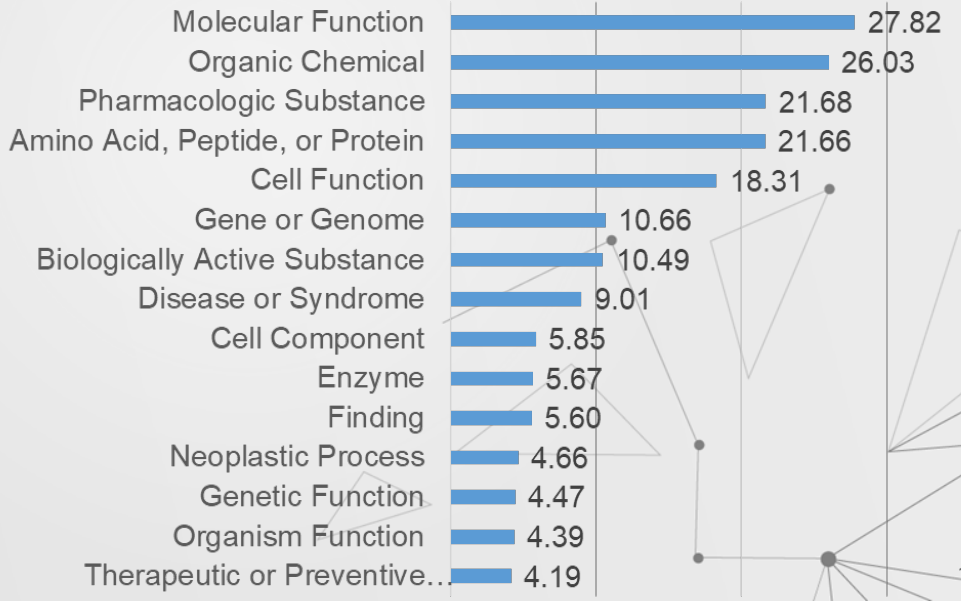
The Lung-Cancer iODG

Top 15 most abundant UMSL SN relation types in the LC iODG

Top 15 most abundant UMLS semantic types in the LC iODG

Distinct relations in millions

Distinct entities in thousands



- **Broad overview** of all knowledge available for each use case to select areas to focus
- E.g. ranking the **semantic types** by the number of distinct **literature-extracted** concepts for each use case

Query examples (1/3)

Semantic Type	Disease Rank			STDV
	LC	ADD	DMD	
Plant	23	22	79	32.62
Bacterium	53	72	104	25.78
Neoplastic process	9	36	59	25.03
Fungus	72	94	120	24.03
Mental of Behavioral Dysfunction	65	28	53	18.88
Eukaryote	39	43	70	16.86
Activity	81	76	51	16.07
Mental process	62	39	34	14.93
Hazardous or poisonous Substance	32	45	61	14.53
Organism Attribute	77	71	50	14.18

Top 10 semantic types with the highest standard deviation (STDV) of ranks for the three case studies

- Precise access to highly **detailed information** for specific entities of interest
- E.g. **drugs** directly **related** with patient Long Term Survivorship (**LTS**) in LC
- Knowledge from **literature** and **structured resources** (“is a” and “interacts with”)

Query examples (2/3)

Concept Label	Interacting concepts		
	All	DrugBank	Enzymes
Cisplatin	991	538	45
Antineoplastic Agents	210	0	24
Aim	243	0	11
Melphalan	58	51	0
everolimus	640	627	2
cetuximab	71	31	4
C3a des-Arg77	6	0	0
Interferons	12	0	0
animal allergen extracts	87	0	5
gefitinib	985	760	39
Altretamine	110	101	0
Topotecan	561	532	2
Paclitaxel	1,518	1,356	17
Carboplatin	147	97	1

Web User Interface - Graph Browser

The screenshot shows a web browser window with the URL localhost:3000/browse. The page title is "Open Data Graph" with navigation links for "Home" and "Browse", and a "Login" link in the top right. The main heading is "Graph Browser" followed by an "Input search term" search bar. On the left side, there are three filter sections: "Relationship types" with an "Expand node" button and a "Get next nodes" link; "Semantic types"; and "Number of neighbors" with a text input containing "10" and an "Apply" button. On the right side, a blue header bar is followed by a "Related Articles" section. This section contains a table with columns for "PMID", "Title", "Journal", and "Mentions". Below the table, a note states: "* highlighted articles are annotated with the term as topic".

Web User Interface - Graph Browser

The screenshot shows a web browser window titled "Welcome to Open Data Graph" with the URL "/browse". The page header includes "Open Data Graph", "Home", "Browse", and a "Login" link. The main content area is titled "Graph Browser" and features a search input field containing "dystri". Below the search field, a list of search results is displayed, categorized into "Relationship types" and "Semantic types".

Relationship types

- dystrophin-associated glycoprotein complex
- dystroglycan complex
- Dystrophin-Associated Protein Complex
- Dystrophin
- dystrobrevin
- Dystrophin-Associated Glycoprotein 1
- Dystrophin genes
- Dystrophin-Associated Proteins
- Dystroglycans

Semantic types

- dystroglycan binding
- dystrobrevin complex
- Dystrophy
- Dystrophia myotonica 2
- Dystrophia myotonica 1
- Dystrophic cardiomyopathy

At the bottom left, there is a "Number of neighbors" section with a dropdown menu set to "10" and an "Apply" button.

Web User Interface - Graph Browser

Welcome to Open Data Graph

Open Data Graph Home Browse

Graph Browser

Relationship types

- ADMINISTERED_TO (1)
- AFFECTS (31)
- ASSOCIATED_WITH (15)
- AUGMENTS (10)
- CAUSES (34)
- COEXISTS_WITH (20)
- CONVERTS_TO (2)
- DISRUPTS (6)

Semantic types

- Amino Acid, Peptide, or Protein (65)
- Acquired Abnormality (4)
- Biologically Active Substance (52)
- Biologic Function (1)
- Body Part, Organ, or Organ Component (31)
- Cell Component (17)
- Cell Function (7)

Number of neighbors

Expand node Get next nodes

Dystrophin

Semantic Type:

- Substance -> Amino Acid, Peptide, or Protein
- Substance -> Biologically Active Substance

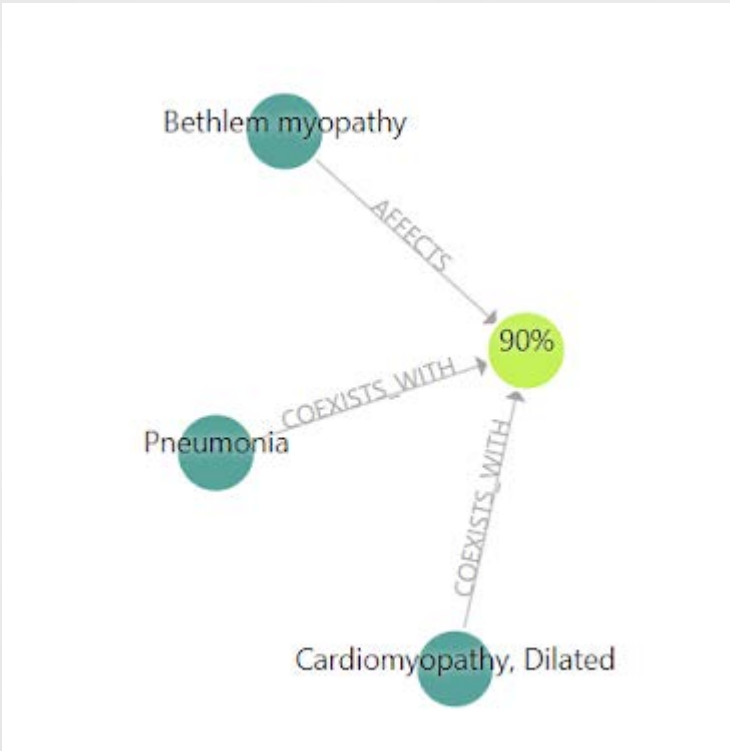
Related Articles

PMID	Title	Journal	Mentions
28869484	Efficient Skipping of Single Exon Duplications in DMD Patient-Derived Cell Lines Using an Antisense Oligonucleotide Approach.	Journal of neuromuscular diseases	0
10573008	Identification of point mutations in Turkish DMD/BMD families using multiplex-single stranded conformation analysis (SSCA).	European journal of human genetics : EJHG	1
10573008	Identification of point mutations in Turkish DMD/BMD families using multiplex-single stranded conformation analysis (SSCA).	European journal of human genetics : EJHG	0

* highlighted articles are annotated with the term as topic

- Some nodes are:
 - Too Generic
 - Diagnosis
 - Sports
 - Patient
- Extraction Errors
 - 90%
 - 2mm

Are all nodes useful?





Feature engineering

Node features

- UMLS Semantic types
- Pagerank
- Node2vec

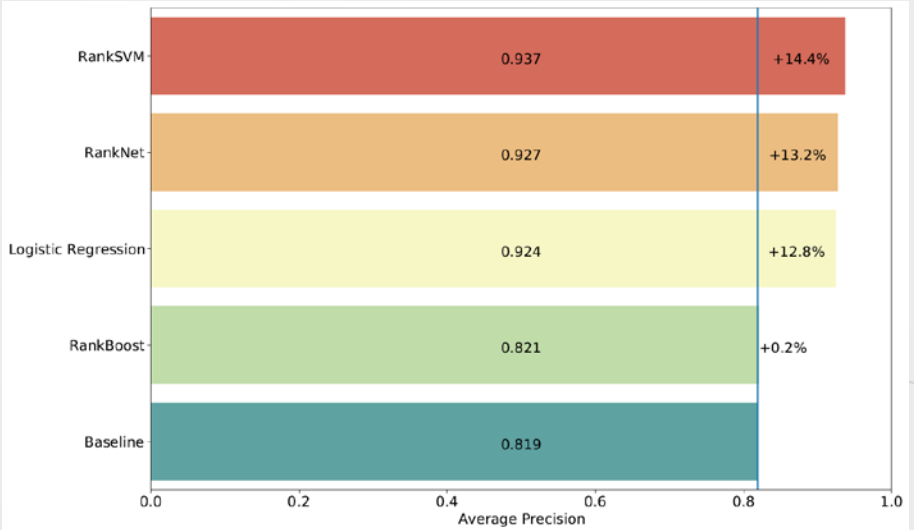
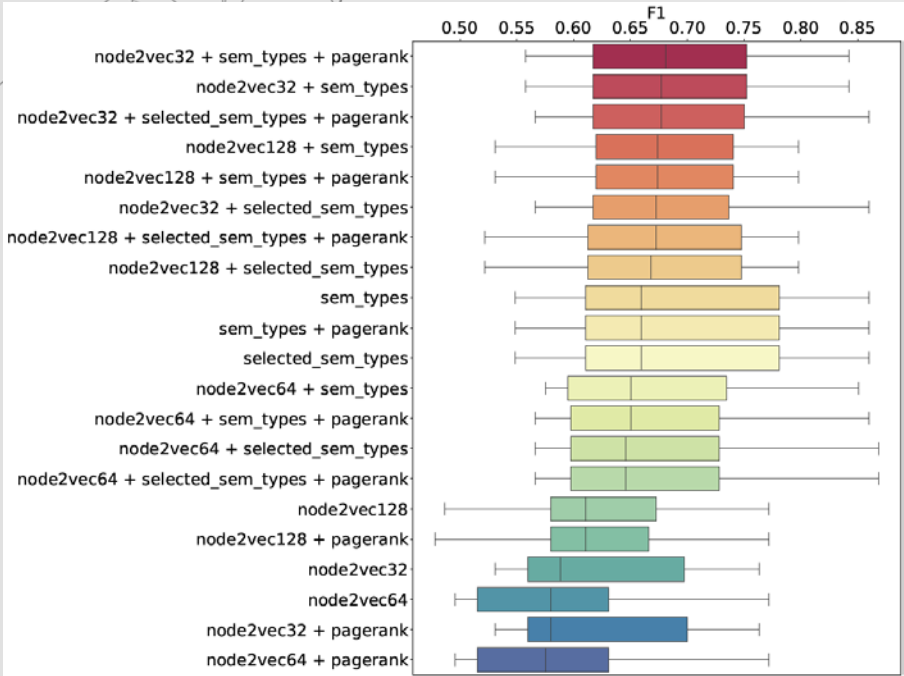
Node ranking

Ranking models

- Logistic Regression
- RankSVM
- RankBoost
- RankNet

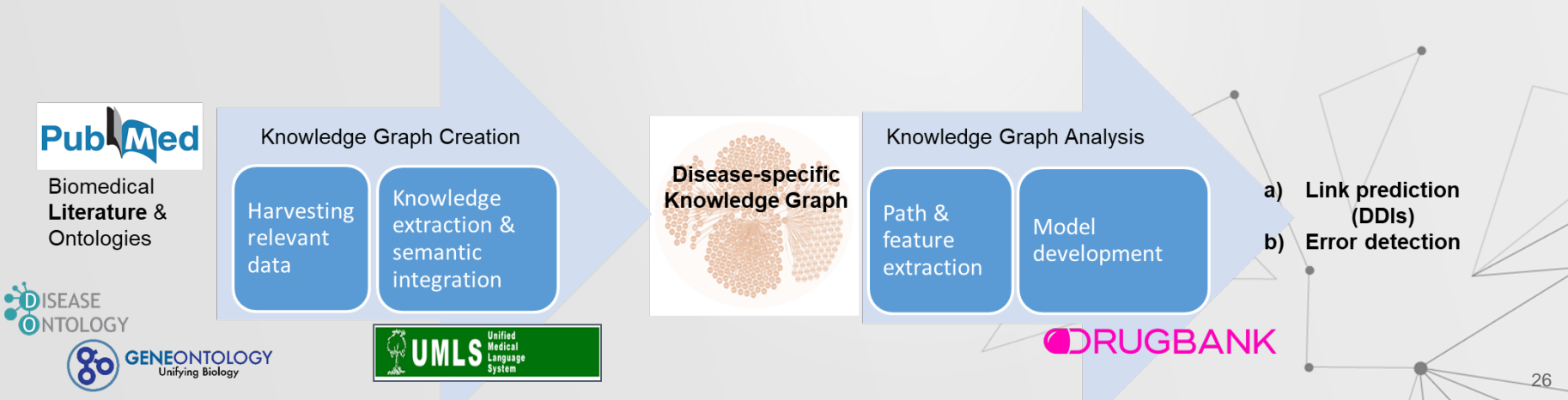


Node ranking



iASiS ODG Analysis

- **Link Prediction (Drug–drug interactions)**
 - Predicting interactions between drugs in the Knowledge Graph
- **Error detection**
 - Identifying noise and errors in biomedical Knowledge Graphs



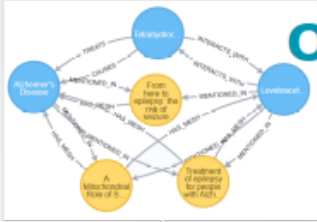
Link prediction

- A change in a **drug's effect** on the body when the **drug is taken together with another drug**.
- More than **190,000 US hospital admissions** each year due to DDIs.
- Standard approaches focus on structural, chemical or other drug similarities (targets, pathways, transporters etc.)
- **Limitation:** Exclusion of heterogeneous data sources, rich in context



Link prediction

- **Final Result:** Creation of a Machine Learning model that can **predict non-obvious links**
- iASiS use case: Predict probable **drug-drug interactions (DDIs)**
- **Evaluation** using DrugBank (“golden” dataset)
- ✓ A relevant patent application submitted (EPO)



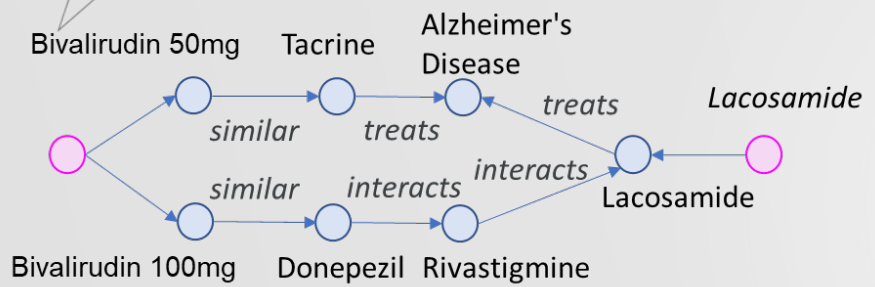
Open Data Graph

Retrieve Paths between Entities (e.g. drugs)

Semantic Path Features

Frequent Pattern Features

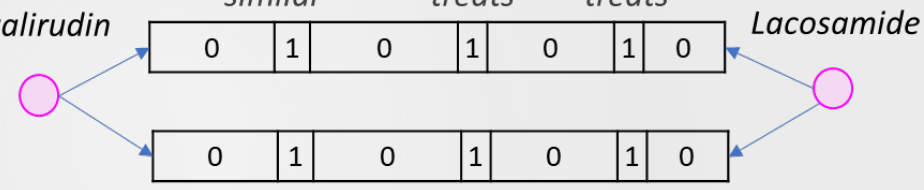
ML MODEL



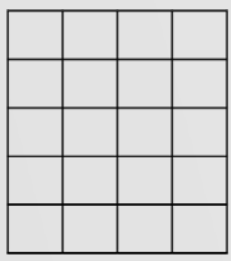
DDI path features

3 hops x 35 relations = **105 features**

One-hot encoding



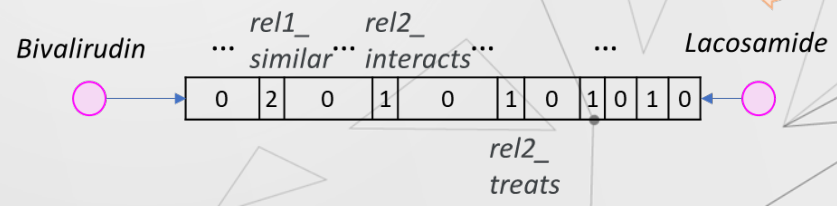
Drug Pairs



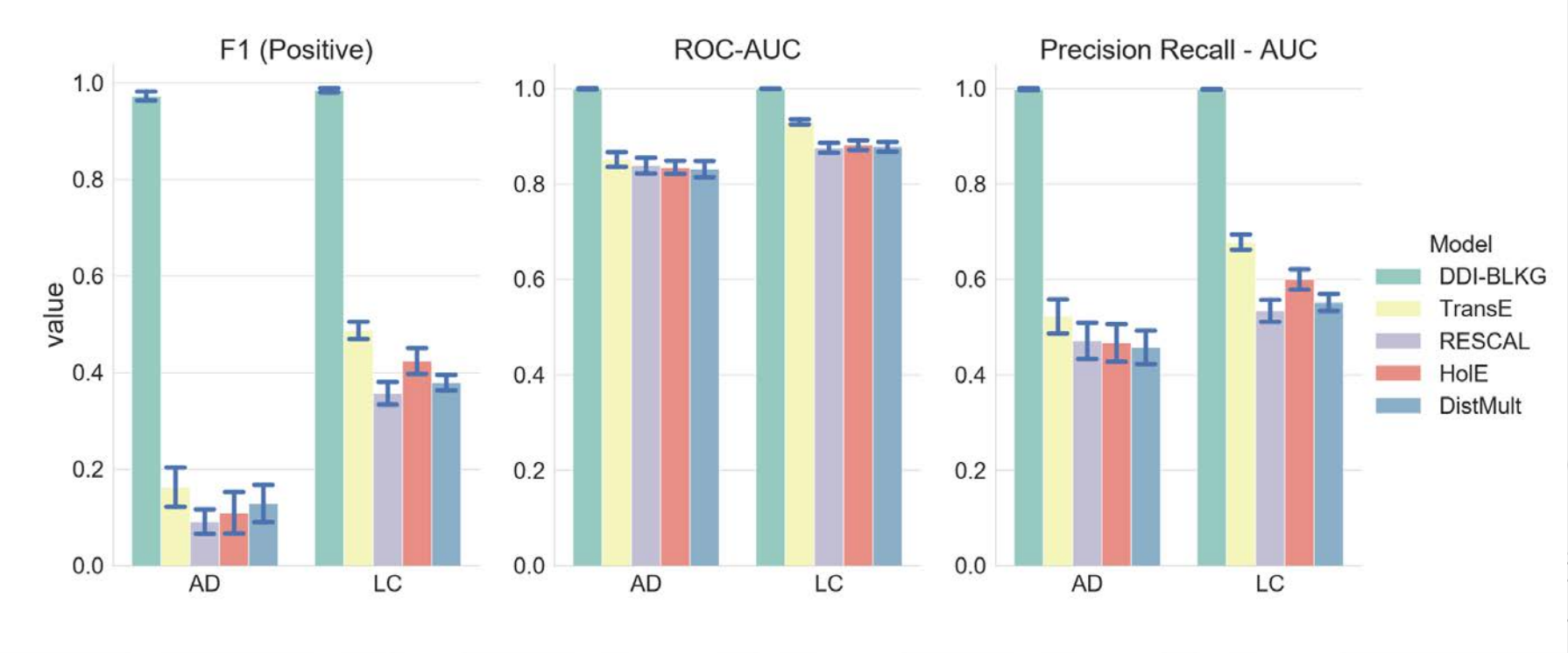
105 features

All pairs

Aggregate Paths



Experiments - Results



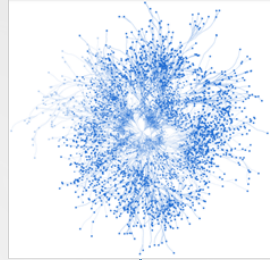
Drug-Target interactions

Drug-target Interaction (DTI) identification

- “Targets” refer to proteins
- Hot research topic
- Used in drug discovery, drug repositioning and other fields
- Documented in repositories like repositories KEGG, DrugBank, ChEMBL, and STITCH
- New DTI predictions help **discover new drugs** and **new uses of existing drugs**
- target some specific genes causing diseases

DTI prediction

- Similar approach with DDI pipeline
 - Training of Machine Learning model on Path-based features
 - Output: **Drug-target interaction prediction**
- **Technical work needed**
 - Path and Feature extraction from Causaly* Knowledge graph
 - Find and integrate a **“golden” dataset** (like Drugbank for DDIs)



Open data
Knowledge
Graph

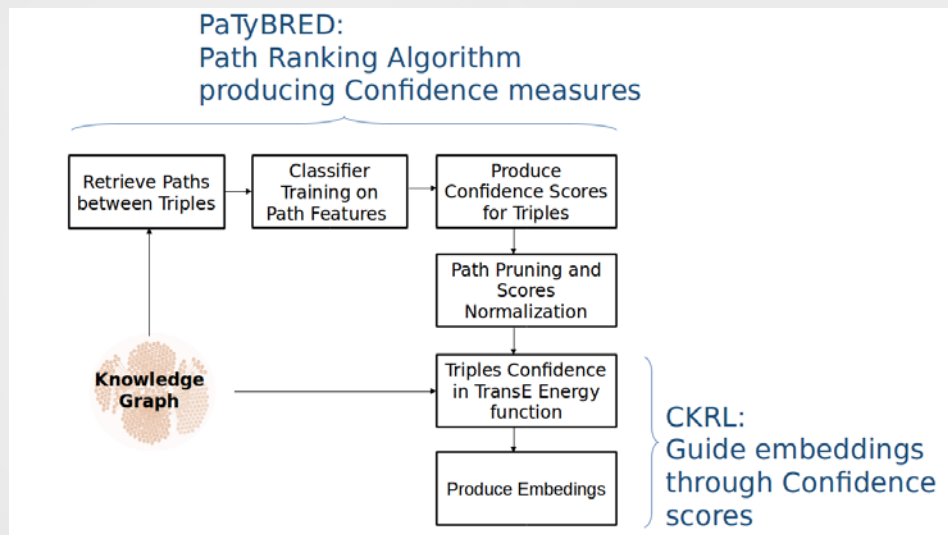
Retrieve Paths between Entities
(e.g. drug-proteins)

Decide Path
Features

ML MODEL

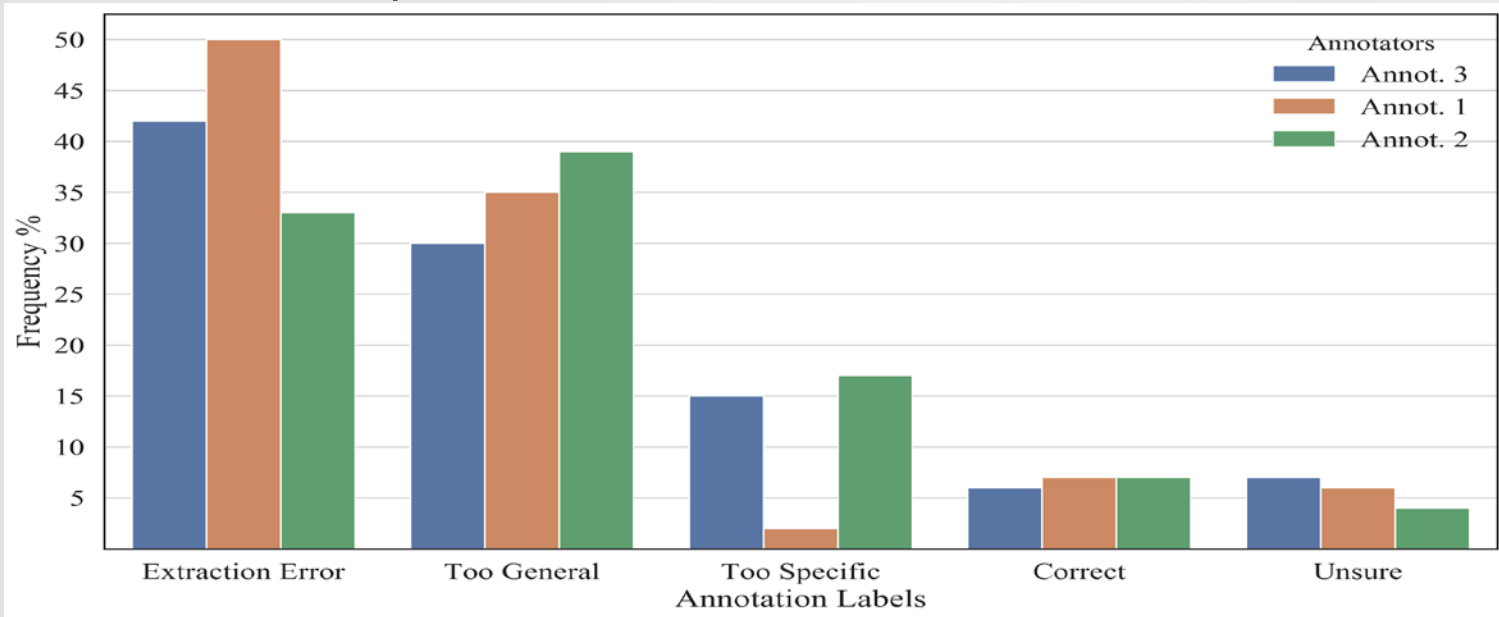
- Aim: Identify/remove **noisy and erroneous data** from a **Knowledge Graph**
 - Small variety of error detection methods for KGs existing (low accuracy)
- A hybrid approach (**Path Ranking Guided Embeddings (PRGE)**) has been developed in IASIS combining different mainstream methods
 - Providing low confidence scores for (potentially) erroneous triples in a Knowledge Graph
 - Marginally improving State of the Art
- **Technical work needed**
 - PRGE code should be automated
 - Use of graph Embeddings other than TransE

Error detection



Error detection evaluation in iASiS KG

Proof of Concept: PRGE method was applied over iASiS Open Data graph to detect erroneous triplets



Annotator's Evaluation of the top-100 most erroneous triples in the iASiS Open Data Graph

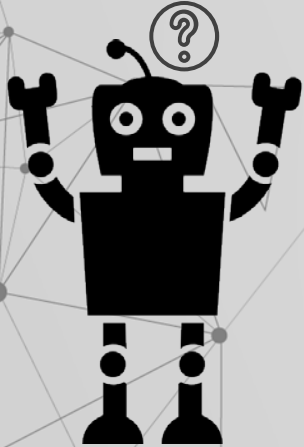
Conclusions

- An **open-source** framework for the semantic **retrieval** and **integration** of disease-specific knowledge into graph
 - Focus on **automated** construction and **incremental** update
 - The graph can be linked to interactive **user interfaces** or provide up-to-date **datasets** for knowledge discovery methods
- **Open Data Graph Analysis**
 - **Link Prediction (Drug–drug interactions)**
 - Predicting interactions between drugs in the Knowledge Graph
 - **Error detection**
 - Identifying noise and errors in biomedical Knowledge Graphs

Future Plans

Our future plans focus on the quality and extension of knowledge

- Detailed **evaluation** of the integrated graph
- Automated **noise detection** approaches
- Structured harvesters for more relation **types** and **resources**



谢谢你

Relevant Publications

- Nentidis, A., Bougiatiotis, K., Krithara, A., & Paliouras, G. (2020). iASiS Open Data Graph: Automated Semantic Integration of Disease-Specific Knowledge. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (Vol. 2020-July, pp. 220–225). IEEE. <https://doi.org/10.1109/CBMS49503.2020.00049>
- Gyftakis, K. (2020). A system for automated disease-specific knowledge integration.
- Bougiatiotis, K., Aisopos, F., Nentidis, A., Krithara, A., & Paliouras, G. (2020). Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 12299 LNAI). https://doi.org/10.1007/978-3-030-59137-3_12
- Bougiatiotis, K., Fasoulis, R., Aisopos, F., Nentidis, A., & Paliouras, G. (2020). Guiding Graph Embeddings using Path-Ranking Methods for Error Detection in noisy Knowledge Graphs. ArXiv. Retrieved from <http://arxiv.org/abs/2002.08762>
- Romanos Fasoulis, Konstantinos Bougiatiotis, Fotis Aisopos, Anastasios Nentidis, Georgios Paliouras, Error detection in Knowledge Graphs: Path Ranking, Embeddings or both?, In CoRR, volume abs/2002.08762, 2020

Unified Medical Language System

UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems

It consists of:

- **Metathesaurus:** Terms and codes from many vocabularies and mappings between them
- **Semantic Network:** Broad categories (semantic types) and their relationships (semantic relations)
- **SPECIALIST Lexicon:** A large syntactic lexicon of biomedical and general English



SemRep

An NLP system designed to recover semantic propositions from biomedical text

Input	<i>We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.</i>
Output	<i>Hemofiltration - TREATS - Patients Digoxin overdose - PROCESS_OF - Patients hyperkalemia - COMPLICATES - Digoxin overdose Hemofiltration - TREATS(INFER) - Digoxin overdose</i>

{Object,Subject} -> Concepts from UMLS metathesaurus
Relationship -> Relation from UMLS semantic network

