A framework for disease-specific information extraction from biomedical literature and open databases, aiming at drug re-purposing

Stavroula Svolou[1,*,†], Fotis Aisopos[1,*,†], Anastasios Nentidis[1,*,†], Anastasia Krithara[1,†] and Georgios Paliouras[1,†]

[1]*Institute of Informatics and Telecommunications, National Centre of Scientific Research "Demokritos", Patr. Grigoriou & Neapoleos Str, Ag. Paraskevi, Athens, 15341, Greece*

Abstract

The information needed for complex biomedical tasks, such as drug re-purposing, is often scattered in different resources, such as biomedical ontologies, databases, and the scientific literature. Scientific literature, in particular, is often the most up-to-date resource, encapsulating the most recent information and latest findings. This work provides a framework for retrieving disease-specific literature and analysing their text with natural language tools for the extraction of concepts and semantic relations. The literature and the information extracted from it are then integrated with information from ontologies and databases, constructing an up-to-date knowledge graph. This graph is then further analysed, providing path-based feature representations for downstream tasks, such as link prediction. This framework is applied to nine neurological, neurometabolic, and neuromuscular disorders, aiming to identify re-purposed drug candidates as potential treatments. To this end, machine learning models are developed achieving promising results on three complementary link-prediction tasks related to drug re-purposing. The preliminary results reveal that both information extracted from the literature, such as concepts and relations, and document-level information, such as concept co-occurrence and document topics, are useful for these tasks.

Keywords

Literature mining, Biomedical ontologies, Knowledge Graph, Drug re-purposing

1. Introduction

The scientific literature is an indispensable source of information for biomedical research as indicated by the billions of annual searches on PubMed[1], the bibliographic database of the US National Library of Medicine (NLM). Moreover, PubMed, which currently consists of almost 31 million documents, is consistently growing with more than one million new documents added annually, during the last few years[2]. In this context, identifying information related to a specific disease is a challenging task per se. In particular, biomedical researchers, who often specialize in a specific disease or a group of diseases, need up-to-date access to all the information available in the relevant literature. Furthermore, they need to combine such information with knowledge located in different resources, such as biomedical ontologies and databases, to estimate the plausibility or the clinical potential of alternative scientific hypotheses and prioritize their experimental investigation.

In the field of drug development, for instance, drug re-purposing is a promising strategy for accelerating the identification of treatments by utilizing existing drugs for new therapeutic purposes. In this direction, existing drugs, already approved for some diseases, are considered as potential candidates

SCOLIA '25: First International Workshop on Scholarly Information Access (SCOLIA), April 10, 2025, Lucca, Italy

*Corresponding author.

†These authors contributed equally.

ssvolou@iit.demokritos.gr (S. Svolou); fotis.aisopos@iit.demokritos.gr (F. Aisopos); tasosnent@iit.demokritos.gr (A. Nentidis); akrithara@iit.demokritos.gr (A. Krithara); paliourg@iit.demokritos.gr (G. Paliouras)

0009-0006-5047-8506 (S. Svolou); 0000-0002-3942-0673 (F. Aisopos); 0000-0002-3782-4412 (A. Nentidis); 0000-0003-0491-4507 (A. Krithara); 0000-0001-9629-2367 (G. Paliouras)

[1]https://pubmed.ncbi.nlm.nih.gov/

[2]https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

for treating a new disease of interest. However, in order to decide which previously approved drug is more promising in a systematic way, one needs to compare many drugs identifying and considering any information available in different resources for each of them, as well as for the disease of interest. Therefore, the automation of the process is the only viable option for large-scale re-purposing studies, where many potential candidate drugs are available.

In this work, we present a software framework developed in the context of SIMPATHIC project[3] for mining and analysing information from scholarly articles, using predictive in-silico models, with the aim to identify, prioritize, and select re-purposing drug candidates, as well as druggable targets for such candidates. Initially, all the literature articles relevant to specific diseases of interest are retrieved from PubMed and PubMed Central (PMC) through semantic search based on Medical Subject Headings (MeSH)[4] topic annotations. Then, Entity Recognition (ER) and Natural Language Processing (NLP) are utilized, to extract knowledge from the raw text into a structured form. In specific, biomedical entities and relations of certain types are identified, presenting the information discussed within the text of those articles in the form of knowledge graph triples. This process also presupposes a fine-grained semantic indexing functionality, employing open and commonly accepted biomedical ontologies and vocabularies, such as the Unified Medical Language System (UMLS)[5]. This preliminary semantified literature knowledge graph is then further enriched with data coming from open databases such as Drugbank[6] and ontologies such as the OBO Foundry Human Disease Ontology[7], providing useful associations such as hypernymic relations and known Drug-Drug Interactions (DDIs) that are manually reported and documented by clinical experts. The integration of all the aforementioned datasets within the knowledge graph provides the ground for generating up-to-date and comprehensive feature representations for interactions among biomedical entities, such as drugs and genes, based on the paths that connect them. This paves the way for further AI-based analysis (e.g. link prediction) with machine learning models, resulting in candidate drug recommendations and drug prioritization.

To highlight the adequacy of these automatically generated feature representations in the field of drug re-purposing, we applied the framework for a group of rare neuromuscular diseases. In particular, we considered three alternative drug re-purposing scenarios, experimenting with three link prediction scenarios, namely predicting Drug-Disease, Drug-Gene, and Drug-Phenotype interactions. Our preliminary results suggest that the information extracted from literature, its integration into a knowledge graph, and the generated path-based feature representations can indeed be useful for link-prediction tasks related to drug re-purposing. In particular, the inclusion of document-level information, such as concept co-occurrence in documents and document-topic relations, in the feature representations appears to have a positive impact on the predictive performance of the machine learning models.

The contributions of this work are summarized below:

- A framework for generating comprehensive feature representations for alternative drug re-purposing hypotheses based on the retrieval and mining of biomedical literature and the creation of an up-to-date disease-specific knowledge graph.
- Investigation of using the generated representations in machine learning models for drug efficacy prediction, under three alternative scenarios.
- The application of the method for a group of rare diseases resulting in a list of potential drug re-purposing candidates for their treatment.

The rest of this paper is structured as follows: First, in section 2 we provide a brief introduction to relevant prior work. Then, in section 3 we describe the structure of the proposed framework. In section 4, we elaborated on the generation and analysis of the datasets, the development of the predictive models, and the respective preliminary results on prioritizing and selecting re-purposing drug candidates. Finally, in section 5 we conclude this work and discuss future directions.

[3] https://simpathic.eu/
[4] https://www.nlm.nih.gov/mesh/meshhome.html
[5] https://www.nlm.nih.gov/research/umls/index.html
[6] https://go.drugbank.com/
[7] http://obofoundry.org/ontology/doid.html

2. Related work

2.1. Literature Mining and Knowledge Graphs

Graphs have an intuitive and versatile structure that renders them adequate for integrating and representing information from different resources, including information extracted from the literature. In the biomedical domain, this is highlighted by the adoption of knowledge graphs by several frameworks and methods for integrating and analyzing health data for different diseases [1, 2, 3]. In this work, we build upon the iASiS Open Data Graph, an open-source framework for the automated retrieval and integration of disease-specific knowledge into an up-to-date knowledge graph, for any disease of interest [4]. The paths connecting different entities in such automatically-generated biomedical knowledge graphs for Alzheimer's disease and Lung Cancer have shown promising results as a resource for generating feature representations for downstream tasks such as drug-drug and drug-gene interaction prediction [5, 6]. In this work, we extended the prior work by introducing a new framework that considers three complementary scenarios related to drug re-purposing. Namely, the prediction of drug interactions with diseases, genes, and phenotypes. In addition, we experimentally investigate the effectiveness of this framework under the above scenarios on a larger scale, considering nine rare diseases.

2.2. Drug Re-purposing using Literature Information

Various works try to utilize the available biomedical literature to achieve drug re-purposing. GrEDeL [7] proposed a biomedical knowledge graph embedding-based recurrent neural network, trained with known drug therapies, in order to discover candidate drugs for diseases of interest. Similarly, works in [8], [9], [10] aim at drug re-purposing based on literature knowledge graph completion techniques, such as graph embedding methods.

The later [10] focuses on disease-specific (cancer) drug re-purposing, as well as works in [11], [12], [13] during the COVID-19 pandemic that tried to identify potential therapies for SARS-CoV-2 infection. Sosa, et. al [14] focus on rare diseases, using a literature-based knowledge graph embedding method to identify drug re-purposing candidates.

The added value of the framework presented in this work entails the addition of co-references and topic relations in the literature knowledge graph, which enhances the information extraction process and its predictive potential, as will be discussed in the experimental results.

3. Methods

The first parts of the proposed framework focus on the retrieval and mining of relevant literature (Sec. 3.1) and the semantic integration of literature-based information with information from relevant structured resources (Sec. 3.2), building upon the methods presented in [4]. Then, the generated Knowledge Graph is used to produce feature representations for three complementary link prediction tasks related to drug repurposing (Sec. 4.3).

3.1. Literature Retrieval and Mining

For the retrieval of relevant literature, we rely on a semantic search over PubMed/Medline. In particular, we use the annotations of PubMed/Medline articles with MeSH thesaurus terms, provided by NLM, to identify all the articles relevant to each disease of interest. MeSH, which stands for the Medical Subject Headings Thesaurus, is a controlled vocabulary developed and maintained by NLM. MeSH consists of more than thirty thousand topics (descriptors) for annotating the main subject of articles from the scientific literature, hierarchically organized primarily into broader and narrower topics [15]. In addition, these topic annotations with all the MeSH terms that represent the main subjects of each article are also retrieved. For these articles, the abstract text is also retrieved from PubMed/Medline, and the full text from PubMed Central, when available.

The text of each article, abstract and full text, is analyzed with concept and relation extraction tools, namely MetaMap [16] for the concepts and SemRep [17] for relations between these concepts. MetaMap and SemRep are two established literature mining tools developed by NLM that rely on a multi-stage NLP analysis. This analysis involves named entity recognition and disambiguation, and the use of syntactic and semantic rules. An important merit of MetaMap and SemRep is the adoption of the UMLS as their semantic reference schema. This renders them comprehensive supporting a wide range of more than three million concepts from the UMLS Metathesaurus[8] and more than thirty types of relations between them from the UMLS Semantic Network (SN)[9]. The precision and recall of MetaMap have been estimated to range from 84% to 93%, and from 61% to 89% respectively, for specific types of entities [18, 19]. The precision of extracting different types of relations between concepts has been estimated to range between 75% and 96%, and the recall between 55% and 70% [20]. Still, the vocabulary of these tools is directly extendible with additional concepts from particular vocabularies of interest. In particular, in this work, we extended the vocabulary considered by these tools with the NCI Metathesaurus (NCIm), which provides additional concepts from biomedical terminologies not available in the UMLS Metathesaurus[10].

3.2. Semantic integration into a Knowledge Graph

The semantic schema of the proposed framework for integrating information from different resources is based on the UMLS and NCIm. In particular, the information retrieved and mined is integrated into a knowledge graph with two basic types of nodes: a) nodes representing articles of the literature, and b) nodes representing concepts from the UMLS or the NCIm metathesauri. Each article node is linked with each concept node corresponding to a concept recognized in its text with an incoming directed edge. These edges representing concept-article relations are labeled as "MENTIONED_IN" edges. A concept node is also linked to any other concept node for which a relation has been extracted from the text of some article. These concept-to-concept edges are labeled with the respective UMLS SN relation type. For instance, we suppose that the relation *"Aspirin"(CUI: C0004057) TREATS "Myocardial Infarction (CUI: C0027051)"* was extracted from the text of some article. In this case, the node for "Aspirin" will be linked with the node for "Myocardial Infarction" with an edge labeled as "TREATS".

In order to integrate the topic annotations of the articles in the same knowledge graph, we link each article node with each concept node corresponding to a MeSH topic of the article with an outgoing edge labeled as "HAS_MESH". The MeSH thesaurus is one of the vocabularies included in the UMLS, hence the mapping from MeSH topics to respective UMLS concepts is available in the UMLS metathesaurus. Beyond topic annotations for the relevant literature, hierarchical relations between MeSH topics were also extracted from the MeSH thesaurus. These binary relations were integrated into the same graph as edges between corresponding concept nodes, labeled after the respective UMLS SN relation type, that is "IS_A".

Hierarchical relations between concepts were extracted from other ontologies as well, namely the Gene Ontology [21] and the Disease Ontology [22], enriching the integrated knowledge graph with more edges labeled as "IS_A". Finally, relations representing the chemical interaction between drugs were extracted from DrugBank [23]. These relations were also integrated into the same knowledge graph, as edges labeled after the respective UMLS SN relation type, that is "INTERACTS_WITH". As with MeSH, Gene Ontology, Disease Ontology, and DrugBank are resources included in the vocabularies of the UMLS metathesaurus, hence the mapping from the original-resource identifiers to UMLS concepts is also available in the UMLS metathesaurus.

[8] In this work, we used the UMLS 2023 release. https://web.archive.org/web/20230710090306/https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

[9] https://www.nlm.nih.gov/research/umls/knowledge_sources/semantic_network/index.html

[10] https://ncimetathesaurus.nci.nih.gov/ncimbrowser/

3.3. Link Prediction and Drug-repurposing model

Using the generated knowledge graph described above and an external data set of known drug relations as groundtruth, we now aim to apply link prediction for three use case scenarios: (a) drug-disease, (b) drug-gene, and (c) drug-phenotype relations. Those use cases have been identified as the most interesting ones towards drug re-purposing by the SIMPATHIC medical experts. Link prediction is addressed as a binary classification problem (relation/no-relation) between respective nodes, using a "white-box" semantic path analysis method presented in [6].

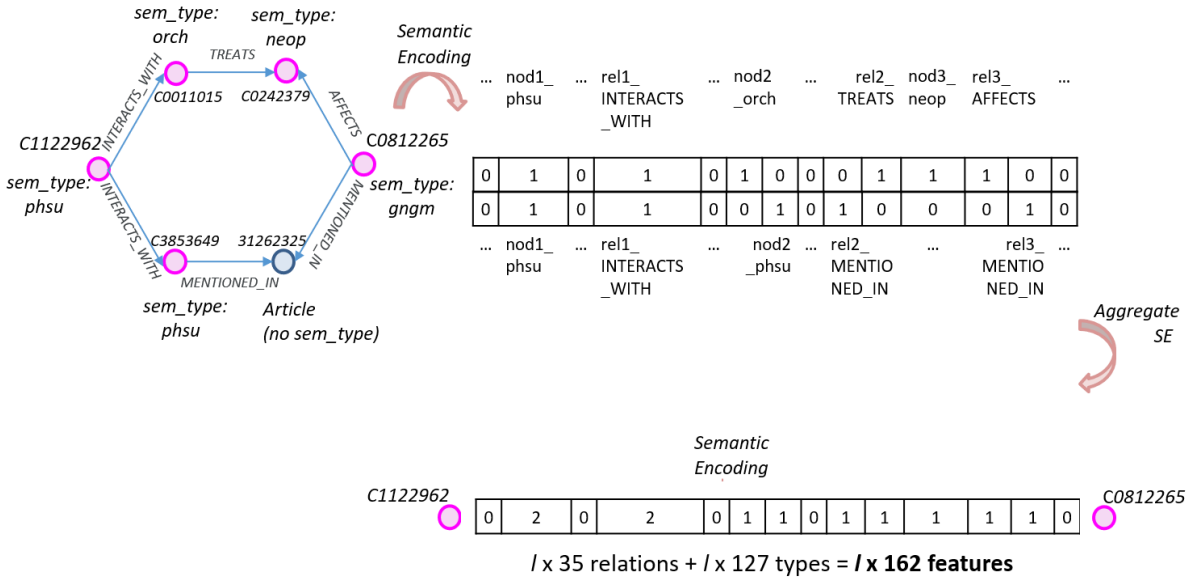

Figure 1: Feature generation process for a drug (Gefitinib) - gene (ERBB3) sample pair with only 2 paths connecting its nodes.

In specific, a set of features is extracted by all existing paths between pairs of interesting nodes in each scenario (e.g. drugs and diseases). Those features are constructed by aggregating the occurrences of subject/object node semantic types and relation types in each hop h of a path, denoted as nod$h$_semanticType and rel$h$_relationType. A schematic presentation of this process for an example drug-gene pair can be seen in Figure 1.

Summing individually the different semantic types (127) and different relation types (35) of each hop for paths of length l, we end up with $lx162$ features in total. Note that, in the current settings, we set maximum path length $l = 3$, as longer paths do not seem to improve the link prediction accuracy. Thus, the feature size finally equals 468.

After generating features for each scenario, a random forest classifier is trained using the labels of the external groundtruth to learn to discriminate between correct and incorrect relations. As a final step, in order to address the drug re-purposing task, the trained classifier in each scenario is asked to score negative pairs (i.e. cases for which no external evidence of relation exists), using all drugs existing in our knowledge graph. This results into a ranked list of drug candidates, each one with a certain confidence, for every disease, gene or phenotype of interest, according to what scenario we want to investigate. The highest ranked candidates can be used as recommendations for drug re-purposing, based on the confidence score accompanying each prediction.

4. Experiments

4.1. Methods Implementation

The proposed framework was implemented as an open-source library of two distinct modules that can be used independently. The Knowledge Graph generation module[11], for the retrieval and integration of up-to-date information from disease-specific literature and selected structured resources has been developed with Java, building upon the iASiS Open Data Graph framework [4]. Regarding the literature mining tools, we have used SemRep (release 1.7) and MetaMap 2020 configured to use the UMLS2023 Metahesaurus vocabulary extended with the vocabulary of the NCIm Metahesaurus. The resulting knowledge graph is saved as a Neo4j[12] graph database. The link prediction method[13] has been implemented with Python 3.12, using the scikit-learn[14] library.

4.2. Data

4.2.1. Scholarly Data Retrieval

As mentioned, in the context of our experiments, we are focusing on 9 rare neurological, neurometabolic and neuromuscular syndromes, namely:

- Spinocerebellar Ataxia type 3 (SCA3)
- Leigh syndrome (Leigh)
- Congenital Neurotransmitter defects (CNT)
- Pyridoxine Dependent Epilepsy (PDE)
- Glutaric Aciduria (GA1)
- PMM2-Congenital Disorder Glycosylation (PMM2)
- Zellweger Spectrum Disorders (ZSD)
- Myotonic Dystrophy type 1 (DM1)
- Congenital Myasthenic Syndrome (CMS)

As advised by medical experts, the Guanosine Triphosphate Cyclohydrolase (GTPCH) Deficiency and the Succinic Semialdehyde dehydrogenase (SSADH) Deficiency for CNT, as well as the Peroxisome Biogenesis Disorder 1A (PBD1A) and the Peroxisome Biogenesis Disorder 6A (PBD6A) for ZSD, were considered as the most interesting sub-syndromes to investigate. Thus, we ended up with 11 specific rare syndromes at hand.

As a first step towards obtaining our data, we need to decide the MeSH Headings, based on which PubMed and PubMed Central must be queried. Some syndromes such as GA1 do not correspond to a MeSH Heading, in which cases we decided to use a more general term (e.g. “Brain Diseases, Metabolic”). Table 1 presents an overview of the specific syndromes, along with the corresponding UMLS, OMIM[15], ORPHA[16] and MeSH terms.

Using the framework presented in Section 3 and the MeSH Headings of Table 1, a total of **34,712** scientific articles from PubMed and PMC have been harvested and analyzed, resulting in a knowledge graph of approximately **215 thousand** nodes and **5.5 million** relations. A small sample of our knowledge graph is available via GitHub.[17]

[11] https://github.com/SSvolou/SIMPATHIC_SCOLIA_2025/tree/main/Harvesting/Literature_Harvester
[12] https://neo4j.com/
[13] https://github.com/SSvolou/SIMPATHIC_SCOLIA_2025/tree/main/Link_Prediction
[14] https://scikit-learn.org/
[15] https://omim.org/
[16] https://www.orpha.net/en/disease
[17] https://github.com/SSvolou/SIMPATHIC_SCOLIA_2025/blob/main/Knowledge%20Graph%20Sample.csv

Table 1
The OMIM terms, Orphanet identifier, UMLS Concept Unique Identifiers and MeSH Headings used for each syndrome in SIMPATHIC.

SYNDROME	OMIM	ORPHA	CUI	MeSH Heading
SCA3	109150	98757	C0024408	"Machado-Joseph Disease"
Leigh	252010	506	C0023264	"Leigh Disease"
GTPCH	233910; 128230	2102; 98808	C0268467	"Phenylketonurias "
SSADHD	271980	22	C0268631	"Amino Acid Metabolism, Inborn Errors"
PDE	266100	3006	C1849508	"Pyridoxine-dependent epilepsy"
GA1	231670	25	C0268595	"Brain Diseases, Metabolic"
PMM2	212065	79318	C0349653	"congenital disorder of glycosylation"
PBD1A	214100	912	C0043459	"Zellweger Syndrome"
PBD6A	614870	912, 79189	C3553947	"Zellweger Syndrome"
DM1	160900	273	C3250443; C0027126	"Myotonic Dystrophy 1"
CMS	620451	590	C0751882	"Myasthenic Syndromes, Congenital"

4.2.2. Drug indication datasets

Following the creation of the knowledge graph, the aim is to experiment with three separate link prediction scenarios, identifying (a) new drug-disease treatment relations, (b) new drug-gene interactions, and (c) drug-phenotype relations. To this end, a groundtruth of approved relations of each type is required, in order to train the machine learning classifier described in Section 3.

For links of type (a), drug indications related to the 11 aforementioned syndromes were extracted and unified from TTD [24], DrugCentral [25], Open Targets [26] and Drugbank [27] repositories. On the other hand, a set of documented drug-gene interactions was retrieved from TTD, KEGG, Drugbank and DGIdb [28] for scenario (b). To proceed with scenario (c), since no open database with structured drug-phenotype relations could be found, we have decided to apply a simple inductive rule on the datasets obtained for (a) and (b): If a drug treats a syndrome related to a phenotype or interacts with a gene related to this phenotype, then we consider the drug-phenotype relation as positive. For each scenario, we consider all possible pairs that do not have a known positive relation as negatives, resulting in the highly imbalanced groundtruth datasets depicted in Table 2.

Table 2
Labeled samples used as groundtruth for Drug-Disease, Drug-Gene and Drug-Phenotype link prediction scenarios.

	Drug-Disease	Drug-Gene	Drug-Phenotype
Positive samples	217	106	1414
Negative samples	19415	20543	7665

4.3. Link Prediction Model Evaluation

After extracting the features[18] from our knowledge graph using the collected ground truth samples, we apply a Random Forest Classifier from scikit-learn[19]. To minimize the risk of over-fitting to specific patterns in the graph, we set the number of decision trees to 100. By using this ensemble learning approach, where each tree is trained on a random subset of the data, we are able to identify the most important features for each use case we study with greater confidence.

To evaluate our model's performance, we implemented a nested Cross Validation (CV) strategy specifically developed to further address the challenges of imbalanced datasets. The outer loop performs a 10-fold CV, while the inner loop runs a 5-fold CV to determine the best under-sampling approach for each fold. In particular, in each iteration of the outer loop, we split the training set into 5 folds,

[18]https://github.com/SSvolou/SIMPATHIC_SCOLIA_2025/tree/main/Extracted_Features
[19]https://scikit-learn.org/

where one of these folds is used as the validation set, and the remaining four are utilized for the model's training.

In the inner loop, we predefined a list of promising sampling ratios (i.e. ratio = [0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.225]) and tested two under-sampling strategies, i.e. RandomUnderSampler[20] and NearMiss[21]. RandomUnderSampler randomly selects samples from the Negative samples class to address the imbalance, while NearMiss selects samples from the Negative samples class based on their proximity to the Positive samples class. We tested all possible combinations of sampling ratios and under-sampling strategies. For each combination and fold, we trained the model and calculated the F1-Score, the harmonic mean of Precision and Recall, using the following formulas:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where TP, TN, FP, and FN denote the numbers of true positive, true negative, false positive and false negative samples, respectively.

Based on the highest average F1-score across the 5 folds in the inner loop, we determine the best under-sampling strategy and ratio, which we then use to train our model on the outer training dataset. By applying this nested CV approach, we aim to handle the imbalanced nature of the tasks we investigate, leading to meaningful predictions.

4.4. Preliminary Results and Lessons Learned

Table 3 summarizes the performance results for each link prediction scenario at hand.

Table 3
Macro-average metric values in a 10-fold Cross Validation of Drug-Disease, Drug-Gene and Drug-Phenotype link prediction scenarios.

	Drug-Disease	Drug-Gene	Drug-Phenotype
Precision	0.497	0.730	0.579
Recall	0.626	0.680	0.157
F1-Score	0.549	0.698	0.246

Looking at the results of Table 3 it is obvious that the performance of the classifier is much worse for scenario (c). This suggests that the groundtruth generated for drug-phenotypes as discussed in Section 4.2.2 is not of high quality. Moreover, examining the disease nodes in the knowledge graph, we identify some syndromes that have a few or even no relationships at all. The lack of context in such parts of the graph is evident due to the lack of a substantial number of research publications in PubMed that focus on such syndromes, which finally causes the classification model performance to drop.

In terms of these performance results, it should be noted that some of the links that are tested may already have been extracted by SemRep. In those cases, the relation extraction tool has already identified the relationships examined in scientific articles, making, thus, the role of the link prediction model trivial. In all three link prediction scenarios, the percentage of such "obvious" predictions was no more than 3% of the overall links (7 out of 217 of Drug-Disease, 1 out of 106 Drug-Gene, and 34 out

[20] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html
[21] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.NearMiss.html

of 1414 Drug-Phenotype positive samples), so we consider the effect of such cases in the overall task performance evaluation as minimal.

Figure 2 illustrates the ten most important features for the classifier of each link prediction scenario. As can be observed, co-mention and topic relationships in various positions of paths are of utmost importance for a classification decision. This finding highlights the significance of the extraction and inclusion of such relation types and instances into our knowledge graph, in order to enhance the performance of the classifier.

A closer examination of the feature sets across the three scenarios reveals that several features consistently play a crucial role in classification. In particular, "MENTIONED_IN" appears in all three feature sets, highlighting the importance of co-mention relationships in identifying relevant connections. Additionally, "INTERACTS_WITH", "HAS_MESH", "humn" (Human), and "PROCESS_OF" are present in at least two scenarios, suggesting that they hold broad predictive value across multiple link prediction tasks. These common features may indicate fundamental relationships between entities that are universally relevant, making them key components for the classification model.

On the other side, certain features appear only in specific scenarios, indicating their case-specific significance. For instance, "phsu" (Pharmacological Substance) and "USES" are unique to the top-10 features of scenario (a), while "LOCATION_OF" is only present in scenario (b), and "podg" (Patient or Disabled Group) is only found in scenario (c). These case-specific features may capture unique aspects of the relationships within their respective datasets, but their impact do not generalize across all tasks.

A major advantage of the employed methodology is its transparency. Unlike other methods, such as Large Language Models (LLMs), which often function as "black-boxes" with limited insight into their decision-making process, our feature-based approach provides meaningful justifications for classification outcomes. By explicitly identifying the most influential features in each scenario, we facilitate a deeper understanding of the underlying relationships within the data, where transparency is particularly valuable in order to make our biomedical predictions trustworthy.

Using the models tested for each scenario, the next step towards drug re-purposing is the scoring of all possible drug candidates for each syndrome of interest (Table 1) and the identification of the top ranking cases. The resulting csv files, available via Github[22], list all initial candidate drugs per syndrome, related gene or phenotype.

As can be observed in the candidate list of scenario (a), for some syndromes (e.g. PMM2-Congenital disorder glycosylation), no candidate drugs have been identified. This happens because of the lack of literature that focuses on such conditions, resulting in an insufficient related context in the knowledge graph. As a mitigation measure, we can identify potential candidates for genes (e.g. PMM2 gene) or phenotypes (e.g. Dysarthria) related to those syndromes, through the other two candidate lists.

5. Conclusions and future work

This paper introduced a framework that generates comprehensive and up-to-date feature representations for alternative drug re-purposing hypotheses, considering three complementary link prediction tasks. Namely, the prediction of drug interactions with diseases, genes, or phenotypes. To do so, this framework is based on retrieving and mining disease-specific literature and the automated generation of a knowledge graph, where information from disease-specific literature is integrated with the information from structured resources. We experimentally investigate the use of this framework to provide drug re-purposing suggestions for a real-world scenario, concerning nine rare neurological, neurometabolic, and neuromuscular syndromes. In this direction, beyond the application of the proposed framework that led to the generation of the respective knowledge graph, we also developed three groundtruth drug indication datasets for model development and evaluation, based on data obtained from publicly available repositories.

A key challenge encountered during this process was the highly imbalanced nature of these datasets, as rare diseases, by definition, have fewer or even no known drug indications at all. Furthermore, no

[22]https://github.com/SSvolou/SIMPATHIC_SCOLIA_2025/tree/main/Drug_Candidates

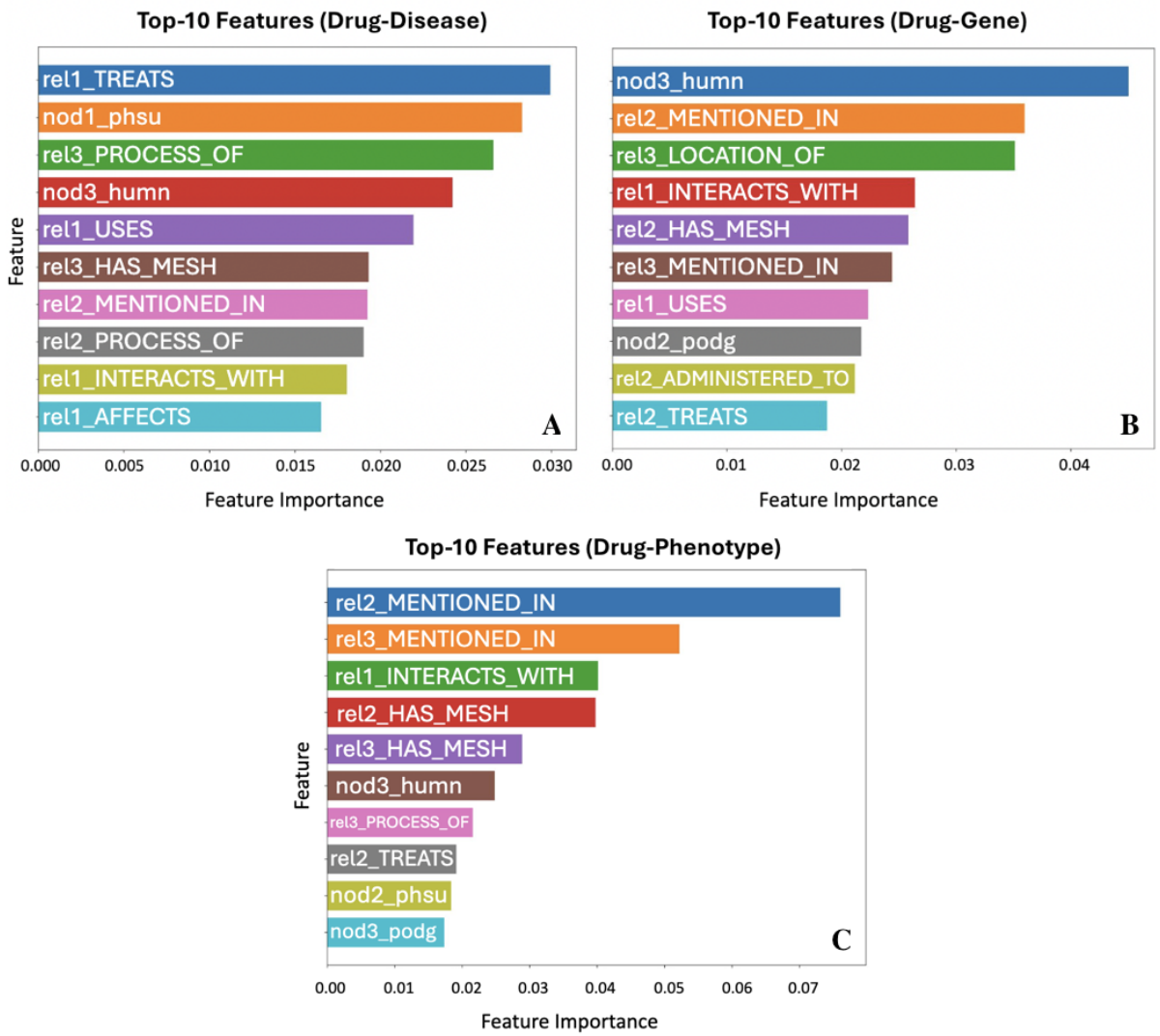

Figure 2: Top-10 most important features identified by our classifier for each one of the studied link prediction scenarios: (A) Drug-Disease, (B) Drug-Gene and (C) Drug-Phenotype.

publicly available database with structured drug-phenotype relationships was found, prompting us to address this gap by applying an inductive rule to the datasets for the drug-disease and drug-gene cases. Addressing the imbalance in the training datasets is crucial, and future work involves data augmentation incorporating up-to-date information from medical experts.

The preliminary results of this work are promising, in particular as regards to the predictive performance achieved in the two scenarios concerning drug-disease and drug-gene link prediction. In particular, the inclusion of information about the relevant articles in the feature representation, that is concept co-occurrence and article topics, seems to be useful for all three scenarios, as revealed by the high importance of the respective feature based on "MENTIONED_IN" and "HAS_MESH" edges respectively. In addition, the lack of predicted drug indications for specific rare syndromes highlights the importance of the complementary scenarios adopted in this framework, as drug interactions with disease-specific genes or phenotypes can be used to identify re-purposing candidates for low-resource diseases.

This work has led to a ranking of alternative drug re-purposing hypotheses, providing transparent predictions, the quality of which, however, still requires validation by wet-lab experiments. The

granularity of the concepts proved to be another important challenge that needs careful consideration, predominantly for document retrieval. In some cases, such as GA1, we may need to consider a broader concept due to the lack of sufficient syndrome-specific context and exact identifier alignment across vocabularies. In other cases, however, important sub-diseases of a main disease concept may be of interest as well.

Our future research plans concerns the investigation of improvements in each part of the proposed framework. For literature retrieval and mining, we built upon traditional biomedical semantic search and NLP approaches, which are often consistent and explainable. However, exploiting modern methods that rely on LLMs is a direction that could improve the quality of the generated knowledge graph and respective feature representations. We also plan to explore approaches for the detection and removal of noisy or not useful parts of the Knowledge Graph, leading to quality improvement and potentially more manageable size. In addition, we plan to enhance the framework by considering additional information, such as author, journal, or year information of articles, semantic descriptions of concepts, and pre-trained embeddings for concept terms. Finally, we also research alternative link prediction methods based on Graph Neural Networks (GNNs).

Acknowledgments

This work was funded by the SIMPATHIC project, in the context of European Union's Horizon 2020 research and innovation programme under grant agreement No 101080249.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. Aisopos, S. Jozashoori, E. Niazmand, D. Purohit, A. Rivas, A. Sakor, E. Iglesias, D. Vogiatzis, E. Menasalvas, A. Rodriguez Gonzalez, G. Vigueras, D. Gomez-Bravo, M. Torrente, R. Hernández López, M. Provencio Pulla, A. Dalianis, A. Triantafillou, G. Paliouras, M.-E. Vidal, Knowledge graphs for enhancing transparency in health data ecosystems1, Semantic Web 14 (2023) 943–976. doi:10.3233/sw-223294.

[2] A. Sakor, S. Jozashoori, E. Niazmand, A. Rivas, K. Bougiatiotis, F. Aisopos, E. Iglesias, P. D. Rohde, T. Padiya, A. Krithara, G. Paliouras, M.-E. Vidal, Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities, Journal of Web Semantics 75 (2023) 100760. URL: https://linkinghub.elsevier.com/retrieve/pii/S1570826822000440. doi:10.1016/j.websem.2022.100760.

[3] A. Krithara, F. Aisopos, V. Rentoumi, A. Nentidis, K. Bougatiotis, M.-E. Vidal, E. Menasalvas, A. Rodriguez-Gonzalez, E. Samaras, P. Garrard, M. Torrente, M. Provencio Pulla, N. Dimakopoulos, R. Mauricio, J. Rambla De Argila, G. Gaetano Tartaglia, G. Paliouras, iASiS: Towards Heterogeneous Big Data Analysis for Personalized Medicine, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), volume 2019-June, IEEE, 2019, pp. 106–111. URL: https://ieeexplore.ieee.org/document/8787467/. doi:10.1109/CBMS.2019.00032.

[4] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, iASiS Open Data Graph: Automated Semantic Integration of Disease-Specific Knowledge, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), volume 2020-July, IEEE, 2020, pp. 220–225. URL: https://ieeexplore.ieee.org/document/9183291/http://arxiv.org/abs/1912.08633. doi:10.1109/CBMS49503.2020.00049. arXiv:1912.08633.

[5] K. Bougiatiotis, F. Aisopos, A. Nentidis, A. Krithara, G. Paliouras, Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 12299 LNAI, 2020, pp. 122–132. URL: http://link.springer.com/10.1007/978-3-030-59137-3_12https://link.springer.com/10.1007/978-3-030-59137-3_12. doi:10.1007/978-3-030-59137-3_12.

[6] F. Aisopos, G. Paliouras, Comparing methods for drug–gene interaction prediction on the biomedical literature knowledge graph: performance versus explainability, BMC Bioinformatics 24 (2023) 1–21. URL: https://doi.org/10.1186/s12859-023-05373-2. doi:10.1186/s12859-023-05373-2.

[7] S. Sang, Z. Yang, X. Liu, L. Wang, H. Lin, J. Wang, M. Dumontier, Gredel: A knowledge graph embedding based method for drug discovery from biomedical literatures, Ieee Access 7 (2018) 8404–8415.

[8] Y. Zhu, W. Jung, F. Wang, C. Che, Drug repurposing against parkinson's disease by text mining the scientific literature, Library Hi Tech 38 (2020) 741–750.

[9] X. Dong, W. Zheng, Emerging technologies for drug repurposing: Harnessing the potential of text and graph embedding approaches, Artificial Intelligence Chemistry (2024) 100060.

[10] A. Daowd, S. Abidi, S. S. R. Abidi, A knowledge graph completion method applied to literature-based discovery for predicting missing links targeting cancer drug repurposing, in: International Conference on Artificial Intelligence in Medicine, Springer, 2022, pp. 24–34.

[11] V. N. Ioannidis, D. Zheng, G. Karypis, Few-shot link prediction via graph neural networks for covid-19 drug-repurposing, arXiv preprint arXiv:2007.10261 (2020).

[12] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, et al., Covid-19 literature knowledge graph construction and drug repurposing report generation, arXiv preprint arXiv:2007.00576 (2020).

[13] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, H. Kilicoglu, Drug repurposing for covid-19 via knowledge graph completion, Journal of biomedical informatics 115 (2021) 103696.

[14] D. N. Sosa, A. Derry, M. Guo, E. Wei, C. Brinton, R. B. Altman, A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases, in: Pacific Symposium on Biocomputing 2020, World Scientific, 2019, pp. 463–474.

[15] S. J. Nelson, W. D. Johnston, B. L. Humphreys, Relationships in Medical Subject Headings (MeSH), 2001, pp. 171–184. URL: http://link.springer.com/10.1007/978-94-015-9696-1_11. doi:10.1007/978-94-015-9696-1_11.

[16] A. R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association 17 (2010) 229–236. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2009.002733. doi:10.1136/jamia.2009.002733.

[17] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (2020) 188. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3517-7. doi:10.1186/s12859-020-3517-7.

[18] R. Reátegui, S. Ratté, Comparison of MetaMap and cTAKES for entity extraction in clinical notes., BMC medical informatics and decision making 18 (2018) 74. URL: http://www.ncbi.nlm.nih.gov/pubmed/30255810http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6157281. doi:10.1186/s12911-018-0654-2.

[19] D. B. Hier, R. Yelugam, M. D. Carrithers, D. C. Wunsch II, High Throughput Neurological Phenotyping with MetaMap, European Scientific Journal, ESJ 18 (2022) 37. doi:10.19044/esj.2022.v18n4p37.

[20] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160. doi:10.1093/bioinformatics/bts591.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology., Nature genetics 25 (2000) 25–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/

10802651http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3037419. doi:10.1038/75556. arXiv:10614036.

[22] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, Disease Ontology: a backbone for disease semantic integration., Nucleic acids research 40 (2012) D940–6. URL: http://www.ncbi.nlm.nih.gov/pubmed/22080554http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3245088. doi:10.1093/nar/gkr972.

[23] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets., Nucleic acids research 36 (2008) D901–6. URL: http://www.ncbi.nlm.nih.gov/pubmed/18048412http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2238889. doi:10.1093/nar/gkm958.

[24] Y. Zhou, Y. Zhang, D. Zhao, X. Yu, X. Shen, Y. Zhou, S. Wang, Y. Qiu, Y. Chen, F. Zhu, Ttd: Therapeutic target database describing target druggability information, Nucleic acids research 52 (2024) D1465–D1477.

[25] S. Avram, T. B. Wilson, R. Curpan, L. Halip, A. Borota, A. Bora, C. G. Bologa, J. Holmes, J. Knockel, J. J. Yang, et al., Drugcentral 2023 extends human clinical data and integrates veterinary drugs, Nucleic acids research 51 (2023) D1276–D1287.

[26] A. Buniello, D. Suveges, C. Cruz-Castillo, M. B. Llinares, H. Cornu, I. Lopez, K. Tsukanov, J. M. Roldán-Romero, C. Mehta, L. Fumis, et al., Open targets platform: facilitating therapeutic hypotheses building in drug discovery, Nucleic Acids Research 53 (2025) D1467–D1475.

[27] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. A. Strawbridge, et al., Drugbank 6.0: the drugbank knowledgebase for 2024, Nucleic acids research 52 (2024) D1265–D1275.

[28] M. Cannon, J. Stevenson, K. Stahl, R. Basu, A. Coffman, S. Kiwala, J. F. McMichael, K. Kuzma, D. Morrissey, K. Cotto, et al., Dgidb 5.0: rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms, Nucleic acids research 52 (2024) D1227–D1235.