Mining Literature-Based Knowledge Graph for Predicting Combination Therapeutics: A COVID-19 Use Case

Ahmed Abdeen Hamed
Clinical Data Science
Sano Centre for Computational Medicine
Kraków, Poland
a.hamed@sanoscience.org
* Corresponding author

Jakub Jonczyk
Department of Medicinal Chemistry
Jagiellonian University Medical College
Kraków, Poland
jakub.jonczyk@uj.edu.pl

Mohammad Zaiyan Alam
Information Sciences Institute
University of Southern California
Marina del Rey, California, USA
mzalam@isi.edu

Ewa Deelman
Information Sciences Institute
University of Southern California
Marina del Rey, California, USA
deelman@isi.edu

Byung Suk Lee
Department of Computer Science
University of Vermont
Burlington, Vermont, USA
byung.lee@uvm.edu

2022 IEEE International Conference on Knowledge Graph (ICKG) | 978-1-6654-5101-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICKG55886.2022.00018

***Abstract*—This paper presents a computational approach designed to construct and query a literature-based knowledge graph for predicting novel drug therapeutics. The main objective is to offer a platform that discovers drug combinations from FDA-approved drugs and accelerates their investigations by domain scientists. Specifically, the paper introduced the following algorithms: (1) an algorithm for constructing the knowledge graph from drug, gene, and disease mentions in the biomedical literature; (2) an algorithm for vetting the knowledge graph from drug combinations that may pose a risk of drug interaction; (3) and two querying algorithms for searching the knowledge graph by a single drug or a combination of drugs. The resulting knowledge graph had 844 drugs, 306 gene/protein features, and 19 disease mentions. The original number of drug combinations generated was 2,001. We queried the knowledge graph to eliminate noise generated from chemicals that are not drugs. This step resulted in 614 drug combinations. When vetting the knowledge graph to eliminate the potentially risky drug combinations, it resulted in predicting 200 combinations. Our domain expert manually eliminated extra 54 combinations which left only 146 combination candidates. Our three-layered knowledge graph, empowered by our algorithms, offered a tool that predicted drug combination therapeutics for scientists who can further investigate from the viewpoint of drug targets and side effects.**

***Index Terms*—Domain knowledge graph, drug repurposing, combination therapeutics, PubMed, ChEBI, disease ontology, gene ontology, drug interaction, MeSH terms, COVID-19**

I. Introduction

Since the time Coronavirus has become a global pandemic, the area of drug repurposing has attracted many players in the scientific community. A simple search in the PubMed online portal for ["Covid-19" and "drug repurposing"] results in 2,220 publications to date. Drug repurposing is known to accelerate the development process of the treatment by identifying existing FDA-approved drug(s) that may be used for the new disease [1], [2], [3]. Most certainly, this research could not be helped without the use of various computational methods that came to the rescue [4], [5], [6], [7], [8]. Indeed, some of the authors here have taken part in such research. The objective of this paper is to address the limitations of the previous research and to advance our previous findings. The current scope mandates working with different ontology (gene, disease, and chemical entities). It is a known fact that those ontologies are fragmented in nature. Such a challenge imposes the need to consolidate those different worlds. Hence the desperate need to construct a knowledge graph that connects all the pieces together, and answers the questions of the new direction of our investigations.

In our previous COVID-19 drug repurposing work [9], [10], we investigated the hypothesis of whether a combination of FDA-approved drugs may be considered a candidate treatment. In order for the hypothesis to be valid, two conditions must be satisfied: (1) evidence of such a combination must exist in the biomedical literature, (2) an exact match of the same combination must also exist in the clinical trials space. We pursued this hypothesis by extracting drug mentions in the PubMed abstracts and constructed a network of drug co-occurrences, where drugs mentioned in the same abstract are connected. We also used the same mechanism to construct a network of drugs that are mentioned in the description and indication of clinical trial records. A clique mining algorithm was applied to each of the networks to identify strongly-connected drugs as

This publication is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No. 857533 and carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund.

978-1-6654-5101-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICKG55886.2022.00018

a potential drug combination. We also presented an algorithm, namely "search-n-match", which compares cliques of the same size and their individual members to confirm the validity of such combinations. The research proved the hypothesis to be true and identified various drug combinations. For instance, the research reported that Nirmatrelvir and lopinavir were the most studied combination, and both are now commercialized under the name Paxlovid.

The previous work, however, was limited to testing the hypothesis stated above and validating the results. Though this work identified common combinations from the biomedical literature and the clinical trial space, it was limited in scope. Further research is needed from the following point of view: (1) the need to investigate whether the members of each combination pose a risk of drug-drug interaction when used together, (2) whether there are contraindications for the use of members of each combination due to the preexisting condition and the use of other drugs (e.g., hypertension, renal or liver dysfunction, simultaneous therapy with antipsychotics or corticosteroids), (3) whether the members of each combination act by the same biological target or by multiple targets. The scope presented here focuses on the first two items, while the third is up for future research.

In this paper, we use an ontology-based information extraction method against the PubMed abstracts. This process provides the necessary knowledge for drugs, diseases, and genes/proteins (as drug targets). The rest of this paper explains how such features form a knowledge graph that can be further explored and queried. Here, we use the COVID-19 domain as a use case; however, our knowledge graph framework may be instantiated in other diseases (e.g., Alzheimer's, asthma, or cancer therapy).

II. Related Work

Drug repurposing is a domain currently established to benefit from the knowledge graph approach, as exemplified by publications starting to come out in the past few years. Zhu et al. [11] gave a comprehensive overview of knowledge graph construction methods and their use for drug repurposing studies.

A few published articles addressed the construction of a knowledge graph specifically for drug repurposing targeting *COVID-19*. Zhang et al. [12] used a neural network-based approach to identify drug candidates from PubMed and other research literature focused on COVID-19. In their conclusion, they recommended using a classifier developed on PubMedBERT (a variant of BERT which is a transformer-based machine learning technique) to construct a COVID-19 knowledge graph and then applying TransE (a neural knowledge graph completion algorithm) to predict drug repurposing candidates. Yan et al. [13] constructed a knowledge graph by integrating 14 public bioinformatic databases containing information on drugs, genes, proteins, viruses, diseases, and symptoms and developing their linkages; and then generated and ranked drug candidates for repurposing as treatments for COVID-19 by integrating motif scores, PageRank scores, and embedding scores for each drug. Al-Saleem et al. [14] constructed the "CAS Biomedical Knowledge Graph" using data from the "CAS Content Collection" and other public repositories; and then used their own result ranking method to predict potential drug repurposing candidates for COVID-19.

Other published articles addressed drug repurposing in general (beyond COVID-19) using a knowledge graph. SemaTyP by Sang et al. [15] is the first knowledge graph built based on PubMed abstract mining for drug discovery. They used a relation extraction tool to extract semantic predications from PubMed abstract texts. Gao et al. [16] constructed a knowledge graph based on associations and presented a computational approach to drug repurposing through lower-dimensional representation of entities and relations in the knowledge graph; they demonstrated the method for the case of Alzheimer's disease. Schartz et al. [17] proposed a new fact-checking mechanism to explaining drug discovery hypotheses using knowledge graph patterns; while interesting, this was not a computational work.

Ratajczak et al. [18] proposed a method to speed up searching a knowledge graph to predict drug repurposing by removing unnecessary facts tailored to the prediction task; knowledge graph construction was not part of the work.

While all these published articles carry significant contributions to the area of knowledge of graph-based drug repurposing, none of them considers drug combinations, differently from our clique-based knowledge graph. Du and Li [19] seem to be the only authors that considered drug combinations. They constructed a knowledge graph of combined therapies from PubMed abstracts manually selected for describing combined therapies. Discovery of drug combination repurposing, however, was not automated; it was done through manual investigations of overlapping semantic predications.

III. Literature-Mining for Drug Repurposing

A. Knowledge sources

The literature-mining for drug repurposing, centered on the medical publication records in PubMed, requires knowledge from multiple sources covering subdomains such as chemistry, human disease, and biology. For each of the subdomains, there are specialized ontologies (with domain-specific taxonomies), created and maintained independently by individual communities of interest. Since the scope of this paper is concerned with drug combinations for a specific disease, various drug targets, and potential drug interactions and side effects upon being combined, here we are using the following knowledge sources.

- PubMed [20]: an indexed database of 34 million citations for biomedical literature from specialized journals and online books. The articles are structured with elements in the MEDLINE format [21]. The journal abstracts are rich in knowledge that is inherently embedded in the text.
- The Chemical Entities of Biological Interest (ChEBI) [22]: a dictionary specialized in knowledge pertaining to "small" molecule chemical compounds,

which are the basis for many of the drugs designed by the Pharma industry.

- The Human Disease Ontology (DOID) [23]: a resource developed as a standardized ontology specialized in providing the research community with reliable knowledge of human disease terms.
- The Gene Ontology (GO) [24]: The world's largest gene functions knowledge-base and is the foundation for computational biology and genetics experiments in biomedical research.
- Medical Subject Heading (MeSH) [25]: a hierarchically organized dictionary used for indexing, cataloging, and searching biomedical and health-related information.

B. The Emerging Need for Knowledge Graphs

The notion of drug combinations as treatment starts with a homogeneous network of drugs. Mining the networks for cliques as a notion for "strongly connected" components offers insights into how the individual drugs that make up the clique may be used in combination for the treatment of a disease. However, this network is missing significant knowledge about whether some drugs interact with others. It is also missing knowledge about the disease linked, side effects, and drug targets. The embedding of such knowledge, however, may prove problematic as the notion of cliques constructed from drugs alone will dissolve. Having to maintain the network of drugs as an individual knowledge source is essential. This situation calls for another layer that integrates knowledge about the drugs with their connected features(side effects, drug targets, genes, and diseases).

The new layer naturally offers links to the network of cliques (via the drugs common to both layers), yet it also offers the connectivity missing in the first layer. The new features offer a wealth of knowledge from the perspective of the disease, and drug target (and hence the side-effect). The risk of drug interaction, however, remains missing and must be computed from another source. We have utilized the Medical Subject Headings as they provide direct pointers to the drug interaction links in the text of the PubMed Abstracts. This presents another fragmented layer of the drugs interacting with each other. Domain experts from the National Library of Medicine took on the task of manually annotating each PubMed Abstract with a "drug interactions" label whenever it applies. However, in order to harvest such knowledge, it must be extracted and stitched together as another layer of the drug interaction as part of the knowledge graph.

C. Knowledge Graph Construction

The preliminary work of this research presented a network of drug associations using the ChEBI ontology. The associations were originally co-occurrences of two or more drugs that appeared in the same abstract. The association analysis extracted the most frequent set of drugs. The associations among drugs lent themselves as a network where the nodes are the drugs and links are the associations extracted using the Apriori [26] algorithm. Here we use the same computational method of constructing the network of drugs for further clique mining. Figure 1 gives an overview of the steps in the computational workflow of constructing the knowledge graph.

Algorithm 1 outlines the steps for constructing the layers of the knowledge graph (the layer of drugs, and the heterogeneous layer of all the other features including drugs).

Algorithm 1 Knowledge graph feature layer construction.

```
Load the ChEBI ontology drug terms into memory.
Load the GO ontology and the DOID ontology into memory.
for each PubMed article's abstract text A do
    Create two empty graphs, D (for "drugs") and H (for "heterogeneous").
    Initialize an empty list l_1 of co-occurring terms.
    for each drug term d in the ChEBI ontology do
        if d is found in the abstract then
            Add d to l_1.
        end if
    end for
    for each combination pair in the list l_1 do
        add to D an edge whose end nodes are the pair.
    end for
    Initialize an empty list l_2 of co-occurring terms.
    for each of ontologies ChEBI, DOID, and GO do
        for each term t in the current ontology do
            if t is found in the abstract A then
                Add t to l_2.
            end if
        end for
    end for
    for each combination pair in the list l_2 do
        add to H an edge whose end nodes are the pair.
    end for
    Output two individual layers of D (for drugs) and H (for all the features including drugs).
end for
```

Here, we provide the details of constructing the drug-interaction layer and how it is inferred. As stated earlier, our approach for computing the drug interactions is designed around the processing of the Medical Subject Headings (MeSH). This step mandates the scanning of the PubMed records to establish whether the MeSH field exists. Algorithm 2 outlines the steps for constructing this layer from the "Drug Interactions" MeSH terms.

D. Vetting and Searching the Knowledge Graph

Now, we have established the need for a fragmented three-layered knowledge graph and demonstrated the computational steps for constructing them, here we present the general framework for vetting the knowledge graph and making it ready for querying. Starting with the layer of drugs, we compute the cliques with a maximum size of five (as was concluded in our previous study). For the computed cliques to

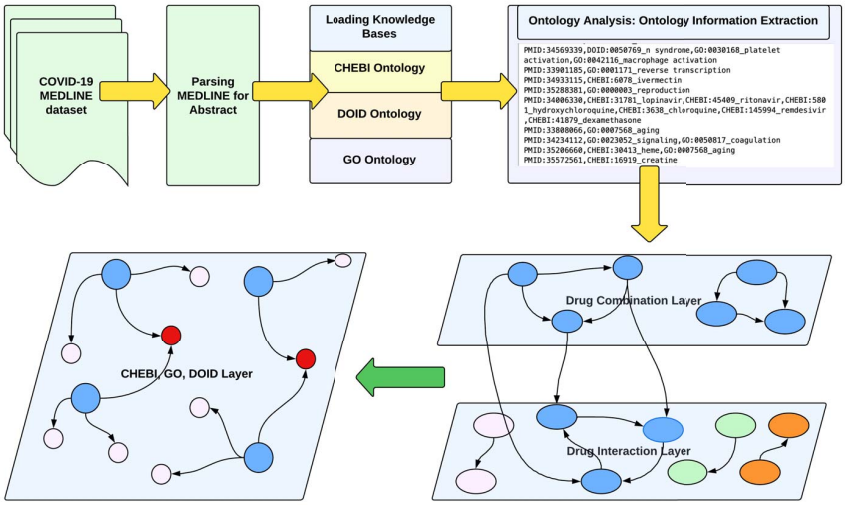

Fig. 1. An overview of the computational workflow steps. Starting with a set of publications in the MEDLINE format, (1) the PubMed abstract texts are extracted along with the PMID, and (2) the various ontologies (CHEBI, GO, DOID) are loaded; then, (3) features are extracted using the ontology terms, constructing network layers comprising the drug combination layer, the drug interaction layer, and the associated heterogeneous ontology layer; then, (4) drug-drug interactions are vetted, and (5) there emerges the final knowledge graph ready for answering queries.

Algorithm 2 Drug interaction network layer construction.

```
Load the ChEBI ontology drug terms into memory.
for each PubMed publication record in the data set do
    if a MeSH field exists and contains the "Drug Interaction" label then
        Extract the abstract text.
        Extract the drug terms mentioned in the abstract as interacting and add them to the drug interaction layer as nodes.
        Add an edge between the newly added drug nodes.
    end if
end for
Output is a network of interacting drugs.
```

be considered, they must pass the vetting process. Specifically, all cliques must be checked for any drug interactions among any two components of a clique. That entails that each pair of drugs of a clique must be checked against the layer of drug interactions in the knowledge graph. If any pair in the drug combination has an exact match in the drug interactions layer, the entire clique is removed. Once all the cliques are vetted, the knowledge graph is ready for querying.

To issue queries against the knowledge graph, we need two different inputs: (1) a vetted list of cliques from the first layer and (2) the heterogeneous layer of all the features including the drugs. A query of one or more drugs may be issued. In the event of searching for a single drug, this can be done as outlined in Algorithm 3. In the event of multiple drugs as input from a query, it is done as outlined in Algorithm 4. Later in this paper, we will present concrete examples of two drugs as

Algorithm 3 Knowledge graph search for single-drug repurposing.

Require: a single drug d

```
Initialize an empty result-set R.
for each clique in the list of cliques do
    for each member m in the clique do
        if the input d matches the member m then
            Find the matching member m in the heterogeneous layer.
            for each edge connected to m do
                Traverse the edge and identify the target node on the other end of the edge.
                Insert the target node into R as a neighbor.
            end for
        end if
    end for
end for
Output the result-set R.
```

a search query and show the matching combinations resulting from such a search.

IV. Knowledge Graph for COVID-19 Use Case

A. The COVID-19 Knowledge Graph Construction

We queried PubMed for the search keyword "COVID-19". The search resulted in 311,456 relevant articles and their corresponding Abstract[AB] field. This is the basic entry point for our work forward. The articles are provided in a format known as MEDLINE [21], which is a record-based plain-text format where an article is described by predefined fields related to authors (AU), PubMed article ID (PMID), title (TI), abstract (AB), and Medical Subject Headings (MeSH)

Algorithm 4 Knowledge graph search for multi-drug repurposing.

Require: a list of multiple drugs

Initialize an empty result-set R.
for each element d in the list of drugs to search **do**
  Initialize an empty list L of drugs matching.
  for each clique in the list of cliques **do**
    if the element d is in the clique **then**
      Insert d into L.
    end if
  end for
  for each element d in L **do**
    Find the matching member m in the heterogeneous ontology layer.
    for each edge connected to m **do**
      Traverse the edge and identify the target node on the other end of the edge.
      Insert the target node into R as a neighbor.
    end for
  end for
end for
Output the result-set R.

among many other metadata. The fields of interest to this study are PMID, AB, and MeSH. For the first two layers of the knowledge graph, we need only the PMID and AB fields. While the PMID is needed to link the ontology features by the notion of co-occurrence, the AB field provides the text to the drug and related features to be extracted.

The ChEBI ontology features contributed the nodes of drugs in the drug combination layer (see Figure 1). The links are derived from the drugs that co-occurred in the same abstract. The three ontologies (ChEBI, DOID, and GO) together contribute the nodes in the heterogeneous layer. We enforce the same notion as before to establish links among all the three types of knowledge features together. When analyzing the 311,456 articles using the three ontology features, it produced a viable 121,483 records of features. Each of these records constitutes the nodes and edges to be contributed to the heterogeneous layer of the knowledge graph (the three types of features together). Table I shows the summary statistics for each of the three layers.

B. Vetting and Querying for Drugs

Vetting Drugs Layer: We stated previously the need to construct a vetting mechanism for interacting drugs. Here, we provide concrete details related to the COVID-19 dataset. We queried the PubMed online portal using the search keywords "drug interaction". The search resulted in 333,006 records (but only 10,000 were the allowed limit). Then, we parsed the MeSH term metadata (which is part of the MEDLINE record) and identified all the records that met the search criteria. If an article has a MeSH annotated as [MH]"Drug interactions', we processed the associated abstract text and extracted the mentioned drugs. Using the same notion of co-occurrences,

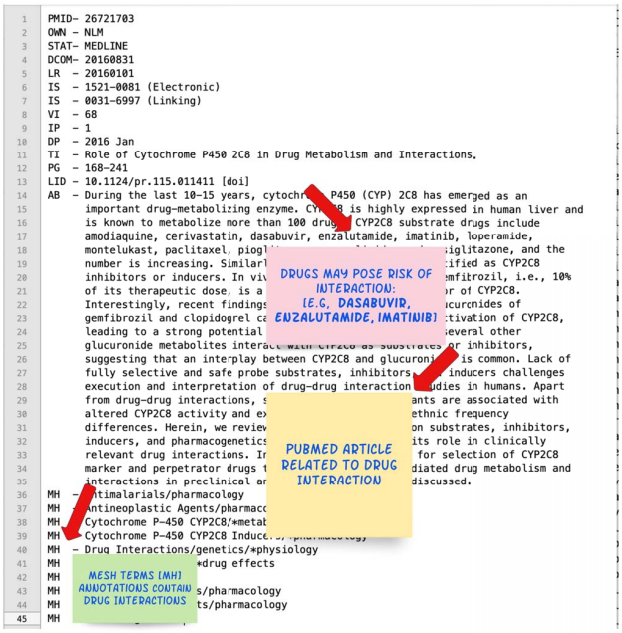

Fig. 2. An example of a PubMed article (PMID:26721703)(from the dataset of 10,000 articles), which is related to drug interaction. As shown, provides information about drugs that may interact (e.g., dasabuvir, enzalutamide, and imatinib). The role of the MeSH term meeting the search criterion [MH]"Drug interactions' is to be noted.

each pair of the drugs was connected as a potential risk of a drug interaction. This dataset provided a knowledge layer that enabled a much-needed vetting mechanism to avoid the risk of drug interactions. Figure 2 shows how a sample of a MEDLINE record for a given abstract that meets our criteria. It is important to note that eliminating all the drugs from the abstracts in this fashion is conservative but also effective in removing any doubt about any candidate combinations.

The algorithmic process of constructing the knowledge graph (Algorithm 1) produced 614 combinations. Using the drug interaction vetting algorithm (Algorithm 2) we successfully eliminated 414 combinations. The remaining 200 combination candidates were further vetted manually by our domain experts. Figure 3 is visually showing a sample part of the network that resulted from the vetting process and the potential risk of drug interaction links is eliminated. In the figure, we observe common drugs that have been investigated as COVID-19 treatment (hydroxychloroquine, chloroquine, and remdesivir).

Any combination that contained a chemical that is not a drug was vetted out manually by our domain expert author — an example is the term "ligand" that appeared in one or more of the publications being analyzed. Concretely, we subjected the cliques proposed by our algorithm to deeper analysis from the point of view of medical chemistry. It helped us remove groups of non-drug substances (including metabolites, solvents, and neurotransmitters) present in the ChEBI ontology. This shows the direction for the further development of the algorithm to reduce the noise resulting from the presence of such

TABLE I
SUMMARY STATISTICS OF EACH LAYER IN THE KNOWLEDGE GRAPH.

Layer	# of nodes	# of edges	# of nodes from: ChEBI	DOID	GO	# of cliques
Drug combination	844	2,450	844	N/A	N/A	2,001
Drug Interaction	1,044	12,159	1,044	N/A	N/A	N/A
ChEBI, GO, DOID	1,800	9,843	844	19	306	N/A

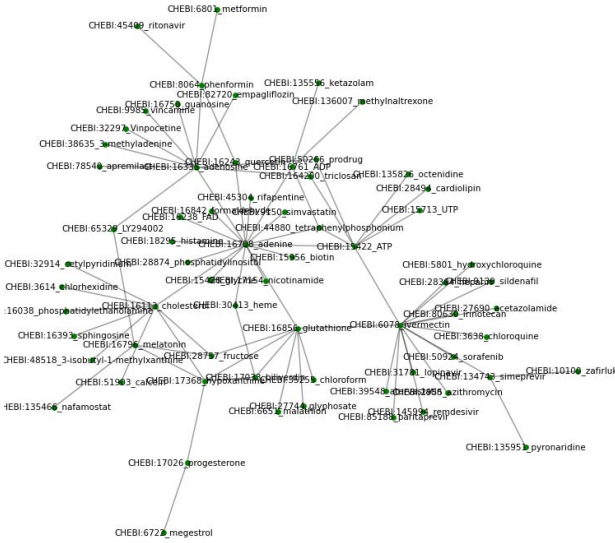

Fig. 3. A network sample of 100 drugs and their connection extracted after being vetted through the drug interaction layer. The sample preserves all the drugs and their connections. However, it is also clear that the network needs more vetting as some invalid drugs are still not eliminated (e.h., CHEBI:15713_UTP, and CHEBI:50266_prodrugs).

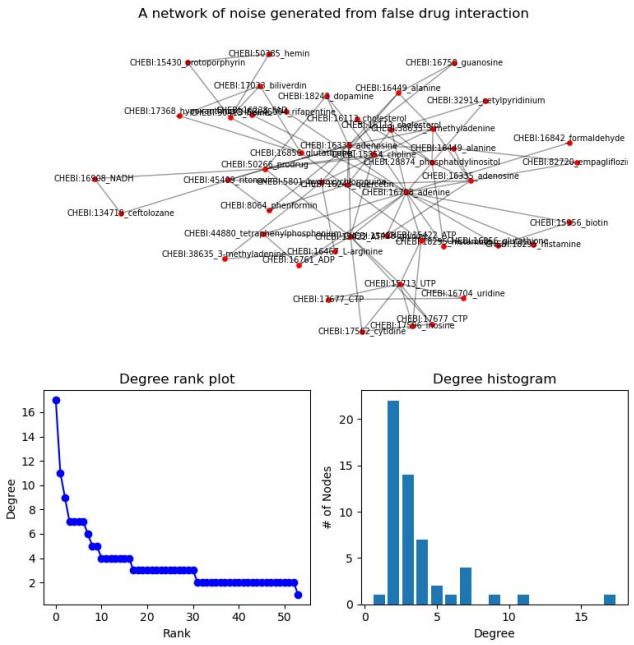

Fig. 4. A network of false drug interactions generated from non-drug nodes (which were valid chemical entities that took some part of the experiments) presented in the abstract. Those nodes introduced many links and enabled the generation of false combinations. Such invalid cliques were identified by our domain expertise. The histogram plot is showing that some drugs have a degree of 20 members. Removing such drugs is necessary for more viable results.

substances in the ontology used. This resulted in 54 drugs that formed 30 false cliques It is essential to eliminate such noise because they introduce many links that produce invalid drug combinations. Figure 4 demonstrates the significance of this vetting mechanism. The figure shows how 54 invalid drugs may cause noise and produce false combinations; although the number of those nodes in the knowledge graph is fairly small, it does create a serious issue of noise because of the high degrees of some drug nodes.

Ultimately, 146 connections were selected. Many of them include drugs approved for COVID-19 therapy, such as ritonavir (in 5 connections), remdesivir (in 21 connections), or dexamethasone (in 14 connections). Many of the proposed drug combinations comprise substances with a different mechanisms of action. These were both combinations that could enhance the effect of antiviral therapy (remdesivir + simeprevir + pyronaridine) and combinations with a broader therapeutic spectrum (conivaptan + dexamethasone + azithromycin). Some of the proposed combinations, however, contained drugs with a repetitive mechanism of action, which limits or prevents the therapeutic use of such clique. Further development of the algorithm should allow us to assess the degree of divergence in the mechanisms of action of drugs in the selected groups.

Querying: The following is an example of how to query the network.

```
CHEBI:31781_lopinavir
```

Here is an example of how to query for drug combinations containing using a single drug named "lopinavir" with the CHEBI term ID 31781. This query returned six different combinations, all of which included the drug lopinavir (the input query). Table II shows the size (i.e., number of drugs) of each combination and lists the drugs constituting it. Each query result can be further explored using (Algorithm 3) by the neighboring drugs, diseases, genes, and proteins. Here is a sample of the gene terms associated with lopinavir:

```
GO:0046697_decidualization
GO:0000792_heterochromatin
GO:0005488_binding
```

C. Implementation Considerations

In order for any therapeutic prediction system to deliver high-efficacy combinations, it must have an ever-evolving knowledge graph that can support such investigations for various diseases. This need calls for a scalable and extensible

TABLE II
THE RESULT OF QUERYING THE KNOWLEDGE GRAPH FOR THE CHEBI:31781_LOPINAVIR DRUG. THE TABLE SHOWS SIX DRUG COMBINATIONS OF SIZES THREE, FOUR, OR FIVE. (BECAUSE OF SPACE LIMITATION, THE DRUG NAMES IN THE LAST TWO ROWS ARE SHORTENED.)

	Size	Member 1	Member 2	Member 3	Member 4	Member 5
Query result: 1	3	CHEBI:135466_nafamostat	CHEBI:5001_fenofibrate	CHEBI:31781_lopinavir	N/A	N/A
Query result: 2	3	CHEBI:39548_atorvastatin	CHEBI:6078_ivermectin	CHEBI:31781_lopinavir	N/A	N/A
Query result: 3	3	CHEBI:4781_emetine	CHEBI:145994_remdesivir	CHEBI:31781_lopinavir	N/A	N/A
Query result: 4	3	CHEBI:82960_raltegravir	CHEBI:31781_lopinavir	CHEBI:4781_emetine	N/A	N/A
Query result: 5	4	CHEBI:5801_hydroxychl.	CHEBI:80630_irinotecan	CHEBI:6078_ivermectin	CHEBI:31781_lopinavir	N/A
Query result: 6	5	CHEBI:135466_nafamostat	CHEBI:5138_fluvoxamine	CHEBI:31781_lopinavir	CHEBI:135632_cam..	CHEBI:45409_ritonavir

scientific workflow environment that provides such support. Indeed, we have implemented our system and executed using the Pegasus Workflow Management System (WMS) [27], [28]. In the current phase of this research, the main objective of this workflow is to perform the various information extraction tasks to process COVID-19 biomedical publications. However, the utilization of a workflow management system makes it possible to process any other input that supports the research of any other disease. Following are the steps that Pegasus made possible to make the final recommendation of 170 combinations: (1) the first step comprises parsing the various ontologies (e.g., disease, gene, and drug-related ontologies), and, (2) the second step is happening concurrently while parsing, is partitioning the input COVID-19 publication dataset into chunks of 3,000 publications each to achieve parallel execution, (3) thirdly, Once the ontology-based features are extracted, each partition of the dataset is paired with all three aforementioned ontologies and all the extracted feature outputs are combined. Pegasus WMS provides end-to-end data management for the workflow, exhibiting the capability of portable execution environments (e.g., containers used for the tasks in the workflow) and, thus, enables us to scale the workflow further. Such scalability enables the construction and maintenance of big knowledge graphs of any disease or multiple various related diseases (e.g., COVID-19 and asthma). Figure 5 shows the information extraction stages started from the MEDLINE format until the ontology features are extracted and the knowledge graph layers are ready to be constructed.

V. CONCLUSION AND FUTURE DIRECTION

In this paper, we presented a combination therapeutics knowledge graph that is implemented and executed by a workflow management system for scalability and extensibility. Using a COVID-19 dataset of biomedical abstracts, we have shown the steps of extracting the domain knowledge using specialized ontologies (of drugs, genes/proteins, and diseases). The knowledge extracted presented naturally-formed three different layers of knowledge (about drug combinations, drug interactions, and a heterogeneous ontology layer of drugs, diseases, and drug targets). The way the three layers of knowledge needed to interact inspired the construction of the knowledge graph. The algorithms we presented have demonstrated promise in both predicting drug combinations and eliminating those that are false.

With the Pegasus workflow being the backbone support for the knowledge graph, we can extend the knowledge graph by introducing new ontologies (e.g., drug target ontology (DTO) [29], [30], bioassay ontology (BOA) [31], cell line ontology [32]). The workflow also offers the flexibility to include various other datasets (e.g., lab notes, doctor's notes, clinical observations, and real-world evidence data). The knowledge graph offers a platform for answering an unlimited number of questions and makes the knowledge easy to query using any open-source engine.

As a natural extension of this ongoing research, the authors intend to further investigate the combinations to develop better knowledge of the mechanism of action. This addresses the level of toxicity for each drug and, hence, sheds insights about the dosage. We will also continue to investigate the drug target for each of the members of a drug combination. The members of a drug combination may negatively or positively interact and may also share the same target with others. This may have a significant impact on the side-effect which must be further investigated in our future research.

ACKNOWLEDGMENTS

This publication is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No. 857533 and carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. This research used the Pegasus Workflow Management Software funded by the National Science Foundation under grant No. 1664162.

REFERENCES

[1] M. A. Farha and E. D. Brown, "Drug repurposing for antimicrobial discovery," *Nature microbiology*, vol. 4, no. 4, pp. 565–577, 2019.
[2] J. Langedijk, A. K. Mantel-Teeuwisse, D. S. Slijkerman, and M.-H. D. Schutjens, "Drug repositioning and repurposing: terminology and definitions in literature," *Drug discovery today*, vol. 20, no. 8, pp. 1027–1034, 2015.
[3] E. L. Tobinick, "The value of drug repositioning in the current pharmaceutical market," *Drug News Perspect*, vol. 22, no. 2, pp. 119–125, 2009.
[4] Y. Zhou, F. Wang, J. Tang, R. Nussinov, and F. Cheng, "Artificial intelligence in covid-19 drug repurposing," *The Lancet Digital Health*, vol. 2, no. 12, pp. e667–e676, 2020.
[5] X. Wang and Y. Guan, "Covid-19 drug repurposing: a review of computational screening methods, clinical trials, and protein interaction assays," *Medicinal research reviews*, vol. 41, no. 1, pp. 5–28, 2021.

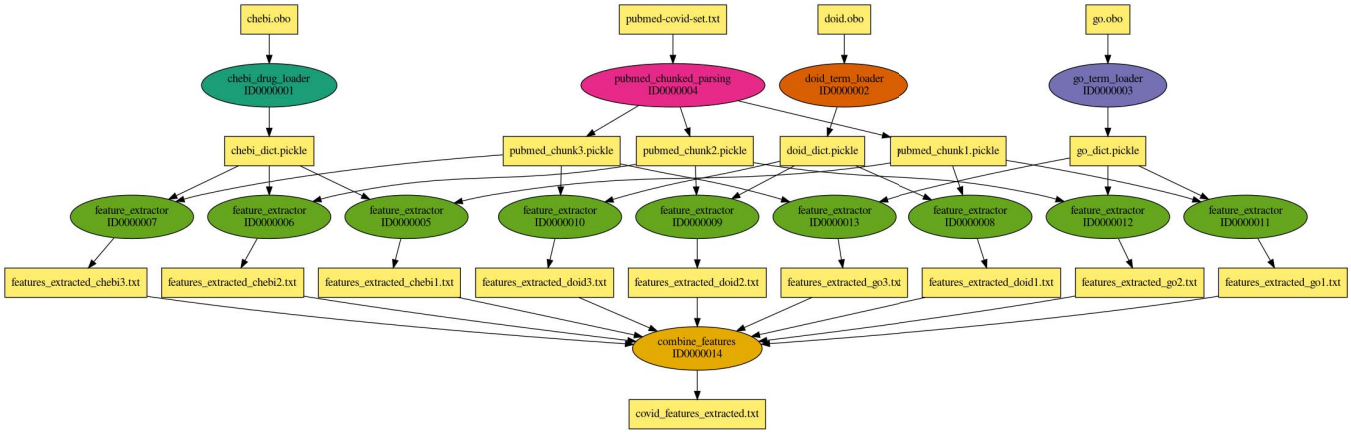

Fig. 5. A snapshot of the complexity of the information extraction steps using the Pegasus Workflow Management System.

[6] J. Wang, "Fast identification of possible drug treatment of coronavirus disease-19 (covid-19) through computational drug repurposing study," *Journal of chemical information and modeling*, vol. 60, no. 6, pp. 3277–3286, 2020.

[7] X. Li, J. Yu, Z. Zhang, J. Ren, A. E. Peluffo, W. Zhang, Y. Zhao, J. Wu, K. Yan, D. Cohen *et al.*, "Network bioinformatics analysis provides insight into drug repurposing for covid-19," *Medicine in Drug Discovery*, vol. 10, p. 100090, 2021.

[8] T. Muramatsu and M. Tanokura, "A novel method of literature mining to identify candidate COVID-19 drugs," *Bioinformatics Advances*, vol. 1, no. 1, 07 2021, vbab013. [Online]. Available: https://doi.org/10.1093/bioadv/vbab013

[9] A. A. Hamed, T. E. Fandy, K. L. Tkaczuk, K. Verspoor, and B. S. Lee, "Covid-19 drug repurposing: A network-based framework for exploring biomedical literature and clinical trials for possible treatments," *Pharmaceutics*, vol. 14, no. 3, 2022. [Online]. Available: https://www.mdpi.com/1999-4923/14/3/567

[10] L. E. Gates and A. A. Hamed, "The anatomy of the SARS-CoV-2 biomedical literature: Introducing the CovidX network algorithm for drug repurposing recommendation," *Journal of Medical Internet Research*, vol. 22, 2020.

[11] Y. Zhu, C. Che, B. Jin, N. Zhang, C. Su, and F. Wang, "Knowledge-driven drug repurposing using a comprehensive drug knowledge graph," *Health Informatics Journal*, vol. 26, no. 4, pp. 2737–2750, 2020, pMID: 32674665. [Online]. Available: https://doi.org/10.1177/1460458220937101

[12] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu, "Drug repurposing for COVID-19 via knowledge graph completion," *Journal of Biomedical Informatics*, vol. 115, p. 103696, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046421000253

[13] V. K. C. Yan, X. Li, X. Ye, M. Ou, R. Luo, Q. Zhang, B. Tang, B. J. Cowling, I. Hung, C. W. Siu, I. C. K. Wong, R. C. K. Cheng, and E. W. Chan, "Drug repurposing for the treatment of COVID-19: A knowledge graph approach," *Advanced Therapeutics*, vol. 4, no. 7, jul 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/adtp.202100055

[14] J. Al-Saleem, R. Granet, S. Ramakrishnan, N. A. Ciancetta, C. Saveson, C. Gessner, and Q. Zhou, "Knowledge graph-based approaches to drug repurposing for COVID-19," *Journal of Chemical Information and Modeling*, vol. 61, no. 8, pp. 4058–4067, 2021, pMID: 34297570. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00642

[15] S. Sang, Z. Yang, L. Wang, X. Liu, H. Lin, and J. Wang, "SemaTyP: a knowledge graph based literature mining method for drug discovery," *BMC Bioinformatics*, vol. 19, no. 193, 2018. [Online]. Available: https://d-nb.info/116397949X/34

[16] Z. Gao, P. Ding, and R. Xu, "KG-Predict: A knowledge graph computational framework for drug repurposing," *Journal of Biomedical Informatics*, vol. 132, p. 104133, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422001496

[17] K. Schatz, C. Melo-Filho, A. Tropsha, and R. Chirkova, "Explaining drug-discovery hypotheses using knowledge-graph patterns," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 3709–3716.

[18] F. Ratajczak, M. Joblin, M. Ringsquandl, and M. Hildebrandt, "Task-driven knowledge graph filtering improves prioritizing drugs for repurposing," *BMC Bioinformatics*, vol. 23, no. 84, 2022. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04608-y

[19] J. Du and X. Li, "A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development," *JMIR Medical Informatics*, vol. 8, no. 4, p. e18323, Apr 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32343247

[20] NIH-NLM. PubMed Central. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[21] ——. MEDLINE/PubMed Data Element (Field) Descriptions. [Online]. Available: https://www.nlm.nih.gov/bsd/mms/medlineelements.html

[22] EMBL-EBI. ChEBI (Chemical Entities of Biological Interest). [Online]. Available: https://www.ebi.ac.uk/chebi/

[23] OntologySearch. Human Disease Ontology. [Online]. Available: https://www.ebi.ac.uk/ols/ontologies/doid

[24] GeneOntology. The Gene Ontology resource. [Online]. Available: http://geneontology.org/

[25] NIH-NLM. MeSH (Medical Subject Headings). [Online]. Available: https://www.ncbi.nlm.nih.gov/mesh/

[26] Agrawal, Rakesh and Srikant, Ramakrishnan and others, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215. Citeseer, 1994, pp. 487–499.

[27] ISI. Pegasus: Makes the Work Flow. [Online]. Available: https://pegasus.isi.edu/

[28] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, , and D. S. Katz, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," *Scientific Programming*, vol. 13, no. 128026, 2005. [Online]. Available: https://www.hindawi.com/journals/sp/2005/128026/

[29] BioPortal. Drug Target Ontology. [Online]. Available: https://bioportal.bioontology.org/ontologies/DTO

[30] Y. Lin, S. Mehta, H. Küçük-McGinty, J. P. Turner, D. Vidovic, M. Forlin, A. Koleti, D.-T. Nguyen, L. J. Jensen, R. Guha, S. L. Mathias, O. Ursu, V. Stathias, J. Duan, N. Nabizadeh, C. Chung, C. Mader, U. Visser, J. J. Yang, C. G. Bologa, T. I. Oprea, and S. C. Schüer, "Drug target ontology to classify and integrate drug discovery data," *Journal of Biomedical Semantics*, vol. 8, no. 50, 2017. [Online]. Available: https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0161-x

[31] NCBO. BioAssay Ontology. [Online]. Available: http://bioassayontology.org/

[32] BioPortal. Cell Line Ontology. [Online]. Available: https://bioportal.bioontology.org/ontologies/CLO