Information Modelling and Knowledge Bases XXXIII
M. Tropmann-Frick et al. (Eds.)

© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA210488

Towards Drug Repurposing for COVID-19 Treatment Using Literature-Based Discovery

Marina TROPMANN-FRICK [a,1] and Tobias SCHREIER [a]
[a] *Hamburg University of Applied Sciences*
Department of Computer Science
Hamburg, Germany

Abstract.
The ongoing COVID-19 pandemic brings new challenges and risks in various areas of our lives. The lack of viable treatments is one of the issues in coping with the pandemic. Developing a new drug usually takes 10-15 years, which is an issue since treatments for COVID-19 are required now. As an alternative to developing new drugs, the repurposing of existing drugs has been proposed. One of the scientific methods that can be used for drug repurposing is literature-based discovery (LBD). LBD uncovers hidden knowledge in the scientific literature and has already successfully been used for drug repurposing in the past. We provide an overview of existing LBD methods that can be utilized to search for new COVID-19 treatments. Furthermore, we compare the three LBD systems Arrowsmith, BITOLA, and SemBT, concerning their suitability for this task. Our research shows that semantic models appear to be the most suitable for drug repurposing. Nevertheless, Arrowsmith currently yields the best results, despite using a co-occurrence model instead of a semantic model. However, it achieves the good results because BITOLA and SemBT currently do not allow for COVID-19 related searches. Once this limitation is removed, SemBT, which uses a semantic model, will be the better choice for the task.

Keywords. literature-based discovery, drug repurposing, COVID-19, Arrowsmith, BITOLA, SemBT

1. Introduction

With the COVID-19 pandemic ongoing for some time, there is still a lack of viable treatments. Although a few promising drug candidates have been proposed, including remdesivir, chloroquine, and hydroxychloroquine [1], their efficacy and safety are still under investigation. What the proposed drugs have in common is that they have been repurposed for treating COVID-19. Remdesivir, for example, was initially developed for treating Ebola and Marburg virus disease but was found to be ineffective against these viral infections [1]. However, antiviral activity was demonstrated against SARS and MERS,

[1]Corresponding Author: Marina Tropmann-Frick, Hamburg University of Applied Sciences, Germany; E-mail: marina.tropmann-frick@haw-hamburg.de

two diseases caused by coronaviruses closely related to SARS-CoV-2 [2]. Likewise, chloroquine and hydroxychloroquine were initially developed for treating malaria [1].

It comes as no surprise that promising candidates for the treatment of COVID-19 are existing drugs. The development of a new drug typically takes 10-15 years, and costs between $500 million and $2 billion [3,4,5]. While the cost of developing a new drug for COVID-19 should be of secondary concern, considering that almost every country is affected by the virus, the development time of 10-15 years is not. The rapidly evolving situation demands new treatments as fast as possible. An alternative to developing new drugs is using existing drugs for new indications, a process known as drug repurposing or drug repositioning.

Literature-based discovery (LBD) is a cost- and time-efficient method that can be used for drug repurposing. It automatically or semi-automatically generates hypotheses for scientific research by finding hidden links in existing scientific literature [6,7]. LBD operates on large literature databases such as PubMed. Discoveries have the form of relations between two previously unrelated concepts, for example, a disease and a drug that treats the disease. Such relations are discovered by uncovering a third concept, like a physiologic function or gene expression, that relates to both the drug and the disease [6]. The discovery of the linking concept leads to the assumption that there may also be a link between the two primary concepts, which can then be investigated further.

This paper gives an overview of LBD methods and systems that could be used to search for possible COVID-19 treatments. Our objective is not to give a comprehensive overview of existing LBD methods. Such an overview can be found in [7]. Instead, we discuss the LBD methods that seem most suitable regarding the search for COVID-19 treatments. Our method selection is based on literature research about LBD with a focus on drug repurposing. First, section 2 presents related work on LBD and its application to drug repurposing. Then, section 3 gives an overview of LBD methods that can be used to search for drug candidates. Next, section 4 introduces three LBD systems and explains how they operate. Based on examples, we demonstrate how they could be used to search for COVID-19 treatments. Finally, section 5 concludes with a summary of our findings and discusses future work.

2. Related Work

LBD was first proposed by Swanson for uncovering hidden knowledge in scientific literature [8]. He read about Raynaud's disease increasing blood viscosity and platelet aggregation in one set of articles and fish oil reducing blood viscosity and platelet aggregation in another set of articles. However, he found no studies that reasoned that fish oil could treat Raynaud's disease. As a result, he proposed fish oil as a new treatment for Raynaud's disease. Another previously unexplored relationship he found was that magnesium might help against migraine [9]. These relationships were discovered manually by researching the literature. To automate the LBD process, Swanson and Smalheiser initiated the Arrowsmith project, a co-occurrence-based LBD system for automatic knowledge generation [10]. The system was later improved with the integration of Medical Subject Headings (MeSH)[11] and the Unified Medical Language System (UMLS)[12] to overcome some of the limitations of co-occurrence based models.

Another co-occurrence-based LBD system named BITOLA was developed by Hristovski et al. [13]. They proposed the use of discovery patterns [6] to search for drugs

that may be repurposed for different indications. To apply the discovery patterns, semantic knowledge had to be derived from the literature. Hristovski et al. used the semantic parsers BioMedLEE and SemRep for this task. Later they integrated SemRep with BITOLA and named the new system SemBT (Semantic BITOLA).

Ahlers et al. adapted the discovery patterns proposed by Hristovski et al. and developed their own discovery pattern [14]. While the discovery patterns proposed by Hristovski et al. are aimed at drug repurposing, the discovery pattern derived by Ahlers et al. is focused on investigating hitherto unknown mechanisms of action involved in existing drug applications. Henry and McInnes provide a comprehensive overview of current and past LBD methods and systems [7]. Thilakaratne et al. [15] give a systematic review of existing LBD literature.

3. Methods

This section addresses important aspects of LBD in the context of drug repurposing. When using LBD, it must first be decided what model should be used to extract knowledge from the literature. Section 3.1 discusses semantic models, which are the most qualified models for LBD. Semantic models extract semantic predications from the scientific literature, which represent relations between two terms. The semantic predications are extracted with semantic parsers. Section 3.2 introduces the two semantic parsers SemRep and BioMedLEE. The process of extracting the semantic predications is explained in section 3.3. For extracting new knowledge from the literature, most LBD systems rely on Swansons's ABC model, which is discussed in section 3.4. If a semantic model is used, different discovery patterns may be applied, which are tailored to detect certain relation types. Two different discovery patterns which are aimed at drug repurposing are described in section 3.5.

3.1. Semantic Models

The most promising LBD models for drug repurposing are semantic models. Other LBD models include co-occurrence models and distributional models. They will not be discussed here, but an explanation of these models and literature for further reading is provided in [7]. Semantic models use semantic predications that are extracted from biomedical literature using semantic parsers [7]. These semantic predications reflect relations assumed between two terms [14], for example, "chloroquine treats malaria". The two terms are chloroquine and malaria, and the relation assumed between them is treats.

Using semantic predications increases the quality of the extracted relations at the risk of missing relations. Thus, it increases the model's precision at the cost of recall. Another benefit of semantic predications is that the extracted relations are labeled, allowing the user to remove uninteresting relations. This reduces the amount of reading required by the user. The model's precision can be increased further by manually eliminating relations that have been wrongly identified (false positives). Furthermore, using predications can explain potential discoveries [6]. Other benefits of semantic predications include normalization, stop word removal, and identification of multi-word terms [7].

3.2. Semantic Parsers

Semantic parsers extract semantic predications from literature. The most popular semantic parser in the biomedical domain is SemRep [16]. Semantic predications in SemRep are three-part propositions, which consist of a subject argument, an object argument, and a relation that binds them. For example, from the sentence

- "We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia."

SemRep would extract the predications

1. Hemofiltration-TREATS-Patients
2. Digoxin overdose-PROCESS_OF-Patients
3. hyperkalemia-COMPLICATES-Digoxin overdose
4. Hemofiltration-TREATS(INFER)-Digoxin overdose

SemRep is based on the UMLS, a comprehensive collection of biomedical vocabularies and standards [17]. The UMLS consists of three knowledge resources, the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon and Lexical Tools. The Metathesaurus, the biggest component of the UMLS, is a large biomedical vocabulary organized by concept or meaning. It links similar names for the same concepts from nearly 200 different vocabularies and identifies useful relationships. The Semantic Network consists of semantic types that categorize concepts from the Metathesaurus and useful semantic relations between them. The SPECIALIST Lexicon and Lexical Tools include a large syntactic lexicon of biomedical and general English terms and tools for NLP tasks such as normalizing strings, generating lexical variants, and creating indexes.

SemRep has been used to extract about 94 million semantic predications from 27.9 million PubMed articles, which are stored in SemMedDB [18]. The subject and object arguments in SemRep semantic predications are concepts from the Metathesaurus, while their relationships are semantic relations from the Semantic Network. Another semantic parser is BioMedLEE, which is a knowledge-based phenotype organizer system that extracts genotype-phenotype relations from biomedical text [19].

3.3. Semantic Predication Extraction

Following the approach of a semantic model, the first step in the LBD process is the extraction of semantic predications from the literature. Semantic predications reflect known facts that are contained explicitly in the literature. Hristovski et al. use the *Associated_with_change* relation from BioMedLEE to extract predications where one concept (e.g., a disease) is associated with a change in another concept (e.g., a pathological function). In addition to the *Associated_with_change* relation from BioMedLEE, they use the *Treats* relation from SemRep to extract drugs that are known to treat certain diseases. BioMedLEE is not required to extract semantic predications that reflect a change in a concept provoked by another concept. Ahlers et al. entirely rely on SemRep for the extraction of semantic predications from biomedical literature. To represent the inhibitory action of one bioactive substance on another, they use the INHIBITS relation. Etiological relations between a bioactive substance and a pathological process are represented with the relations CAUSES, PREDISPOSES, and ASSOCIATED_WITH. Known drug-disease relationships are extracted using the TREATS and PREVENTS relations [14].

3.4. Swanson's ABC Model

After extracting explicit semantic predications that reflect known facts in the literature, the next step in the LBD process is retrieving previously unknown implicit knowledge. Almost all LBD systems rely on Swanson's ABC model to discover new knowledge from research literature [8]. It builds on the assumption that when term A is connected to term B, and term B is connected to term C, it can be assumed that there also is a link between terms A and C. For discoveries to be new, the terms A and C must occur in different literature sets. Figure 1 shows Swanson's ABC-model in the biomedical domain.

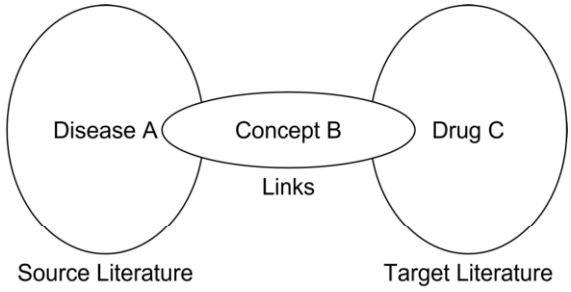

Figure 1. Swanson's ABC-model [20]

For example, Swanson proposed fish oil as a new treatment for Raynaud's disease. The A term in his discovery was Raynaud's disease. Two of the B terms that co-occurred with Raynaud's disease were blood viscosity and platelet aggregation, which are both increased in Raynaud's disease (AB literature). Blood viscosity and platelet aggregation co-occurred with a C term, fish oil (rich in eicosapentaenoic acid), which reduces blood viscosity and platelet aggregation (BC literature). Consequently, fish oil was proposed to treat Raynaud's disease (newly found AC relationship). Swanson's ABC model can be used for open and closed discovery, which are depicted in Figure 2.

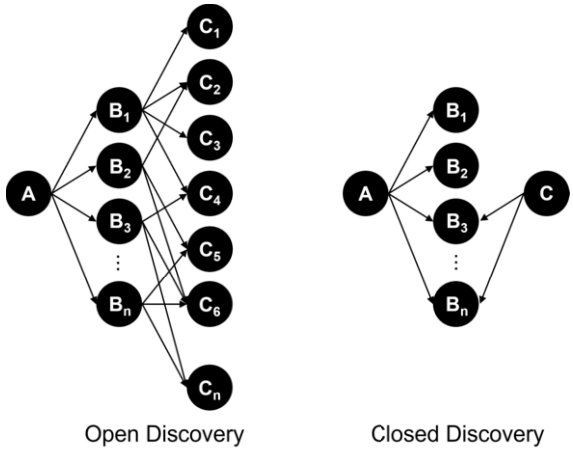

Figure 2. Open and closed discovery [7]

Open discovery is used to generate new hypotheses [7]. To perform open discovery, the user must provide the LBD system with a start term. The system then generates a list of linking terms that co-occur with the start term. Based on the linking terms, the system generates a list of target terms that co-occur with the linking terms. The result is a list of previously unknown relations between the start term and the target terms.

Closed discovery, on the other hand, is primarily used to explain correlations or observations [7]. For example, it may be used to examine hypotheses previously generated with open discovery. To perform closed discovery, the user must enter a start term and a target term. The system then generates a list of linking terms related to the start term and the target term. The result is a list of linking terms that may explain the relationship between the start term and the target term.

3.5. Discovery Patterns

Based on Swanson's ABC model, different discovery patterns have been proposed. Discovery patterns are used to extract potentially new treatments from biomedical literature and to explain existing drug-disease relationships by identifying previously unknown pathways. Hristovski et al. [6] proposed the *Maybe_Treats* discovery pattern, which has the forms *Maybe_Treats1* and *Maybe_Treats2* depicted in Figure 3.

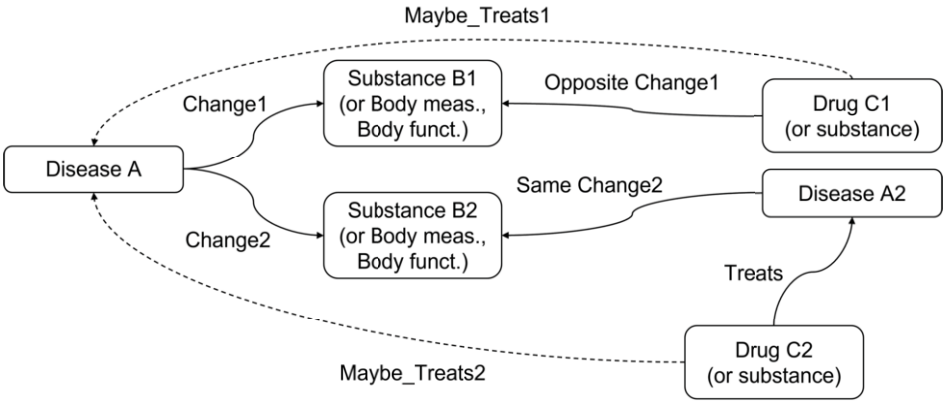

Figure 3. Two forms of the *Maybe_Treats* discovery pattern *Maybe_Treats1* and *Maybe_Treats2* [6].

The *Maybe_Treats1* discovery pattern considers drug C1 to possibly treat disease A if the following applies. First, the disease is associated with changing a B1 concept, which might be a substance, a body measure, or a body function. Second, the drug provokes an opposite change of this substance. To consider this a discovery, the drug and the disease can not co-occur in the literature. For example, in Swanson's discovery of fish oil as a new treatment for Raynaud's disease, the drug C1 was fish oil, while disease A was Raynaud's disease. One of the B1 concepts that linked Raynaud's disease to fish oil was blood viscosity. Blood viscosity is a body measure, which is increased in patients with Raynaud's disease and reduced by fish oil.

The *Maybe_Treats2* discovery pattern follows a different approach. Other than the *Maybe_Treats1* discovery pattern, it does not search for a drug that causes an opposite change of a B concept changed by the disease A. Instead, it searches for different diseases A2 that cause similar changes of a B2 concept as the disease A. Drugs C2 that treat the disease A2 are assumed to possibly treat disease A as well. As with the *Maybe_Treats1* discovery pattern, the drug and the disease must not co-occur in the literature for the discovery to be new. Using this approach, Hristovski et al. observed that insulin levels are decreased in patients with Huntington's disease. Insulin levels are also decreased in patients with Diabetes Mellitus (type 2 diabetes). Therefore, they proposed drugs for Diabetes Mellitus for treating Huntington's disease.

Another discovery pattern was developed by Ahlers et al. They used the discovery pattern *May_Disrupt* to investigate the mechanisms underlying drug therapies that are currently used but poorly understood [14]. Other than the *Maybe_Treats* discovery pattern, which searches for a wide range of linking concepts that may connect a drug to a disease, the *May_Disrupt* discovery pattern concentrates on pharmacogenomics, the relationship among drugs, genes, and diseases. The *May_Disrupt* discovery pattern consists of the following three parts:

1. Substance A <inhibits> Substance B
2. Substance B <causes> Pathology C
3. Substance A <may_disrupt> Pathology C

To apply the pattern, first, the relations must be extracted from the literature. Ahlers et al. use SemRep for this task. First, "drug A inhibits substance B" relations are extracted using the INHIBITS relation. Second, "substance B causes disease C" relations are extracted using the CAUSES, PREDISPOSES, and ASSOCIATED_WITH relations. Third, "drug A may disrupt disease C" relations are extracted using the TREATS relation. When the pattern is used for open discovery, it states the following: if drug A inhibits substance B and substance B causes disease C, drug A may disrupt disease C. If it is known that drug A disrupts disease C, the mechanism of action can be examined with closed discovery. When the pattern is used for closed discovery, it states that if drug A inhibits substance B and substance B causes disease C, then substance B is involved in the mechanisms of action of drug A disrupting disease C.

4. LBD Systems

Several LBD systems have been developed in the past. We compared the three LBD systems Arrowsmith, BITOLA, and SemBT, concerning their suitability to search for COVID-19 drug repurposing candidates. The systems share the following properties. They are publicly available, support open and closed discovery, use PubMed as their literature database, and MeSH for filtering terms. An overview of other LBD systems is provided in [15]. We briefly explain the functionality of each system and demonstrate open and closed discovery for each system based on example experiments. Additionally, we provide a conclusion for each system that outlines its limitations.

4.1. Arrowsmith

Based on Swanson's manual discoveries of the connections between Raynaud's disease and fish oil [8], and between migraine and magnesium [9], Swanson and Smalheiser developed the Arrowsmith LBD system [10]. Arrowsmith is the first semi-automatic LBD system and uses a co-occurrence model to find relationships between terms. To overcome some of the limitations of co-occurrence-based models, Smalheiser later improved the system by integrating biomedical knowledge resources, including the UMLS [12] and MeSH [11]. MeSH is a hierarchically organized vocabulary controlled by the National Library of Medicine (NLM). It includes subject headings appearing in PubMed, the NLM catalog, and other NLM databases and is used to index, catalog, and search biomedical information. Arrowsmith allows the user to input PubMed queries in order to define the A and C literature. For that purpose, a simplified version of the PubMed query box was integrated into the system.

For closed discovery with Arrowsmith, the user is asked to input two separate PubMed queries that define the A and C literature. We used closed discovery to examine linking concepts that may explain the mechanism of action of remdesivir concerning COVID-19. To define the A literature, we used the start term COVID-19. The query yielded 25,000 articles dealing with COVID-19, the maximum number of articles considered by Arrowsmith for the A literature. To define the C literature, we chose remdesivir as the target term. That resulted in 526 articles about remdesivir. Theoretically, up to 25,000 articles could be considered by Arrowsmith for the C literature, so the total number of articles considered for closed discovery adds up to 50,000 articles [21]. Arrowsmith always considers the latest 50,000 articles. Articles that occur in both literature sets are removed. Thus only indirect relations between the literature sets via linking concepts are captured. However, the removed articles are kept and can be inspected by the user. For COVID-19 and remdesivir, this affected 457 articles, leaving 68 articles that deal with remdesivir but not COVID-19.

Next, the system searches for words and two- and three-word phrases in the A and C literature article titles. The resulting linking terms are processed and ranked according to the predicted probability of being relevant to the user. Torvik and Smalheiser integrated this feature into Arrowsmith to solve the problem of predicting which of the hundreds to thousands of linking terms returned for a single query are most likely to be relevant to the user [12]. They developed a logistic regression model that estimates the probability for each linking term to be relevant. Based on their predicted relevance, the linking terms can be ranked. Furthermore, the model estimates the total number of relevant linking terms. For COVID-19 and remdesivir, 498 linking terms were generated, 89 of which were predicted to be relevant. We limited the linking terms to concepts that may explain the mechanism of action of remdesivir regarding COVID-19. For that, we restricted the linking terms to the semantic types anatomy, chemicals & drugs, genes & molecular sequences, gene & protein names, and physiology. That decreased the number of linking terms to 96. 13 of these terms were predicted to be relevant, the first ten of which are shown in Table 1.

Table 1. First ten linking terms generated by Arrowsmith using closed discovery for the start term COVID-19 and the target term remdesivir.

Rank	Probability	Linking Term
1	0.99	ritonavir
2	0.99	respiratory syndrome coronavirus
3	0.99	lopinavir ritonavir
4	0.98	ebola
5	0.97	cov
6	0.96	rna dependent rna
7	0.94	dependent rna polymerase
8	0.94	dependent rna
9	0.74	antiviral strategy
10	0.74	chloroquine

If the user selects a linking term, Arrowsmith presents the articles linked to the start and target term through the linking term. Multiple linking terms may be selected as well. Of the first ten linking terms shown in Table 1, only rna dependent rna, dependent rna polymerase, and dependent rna explain the physiologic link between COVID-19 and remdesivir. All these terms are synonyms for RNA-dependent RNA polymerase and thus refer to the same concept. Remdesivir inhibits the RNA-dependent RNA polymerase of MERS [22], various flaviviruses [23], Ebola virus [24], and human endemic and zoonotic deltacoronaviruses [25]. Since SARS-CoV-2 also relies on an RNA-dependent RNA polymerase for the catalyzation of the RNA replication process [26], this could explain the mechanism of action of remdesivir regarding COVID-19.

To perform open discovery with Arrowsmith, the user must enter a single PubMed query to define the A literature. We used the start term COVID-19 to limit the A literature to articles dealing with COVID-19. Next, the user chooses a MeSH category to filter the C literature searched for target terms. Since the objective is to find existing drugs that may be repurposed for COVID-19 treatment, the target terms should be drugs. Therefore, we chose the MeSH category molecular mechanisms of pharmacological action, which includes 20 classes of drugs. Given a MeSH category, the system performs a closed discovery search between the defined A literature and all subcategories of the defined MeSH category (C literature). For each subcategory, several metrics are calculated that quantify the result:

- nC = number of articles in the C literature
- nAC = number of articles in both the A and C literature
- nTot = total number of linking terms in the subcategory search
- nR = number of linking terms predicted to be relevant
- pR = percentage of linking terms predicted to be relevant (high pR values indicate that the A and C literature share a lot of implicit information. $pR < 0.1$ is near chance level whereas $pR > 0.3$ is a relatively high value.)

Table 2 shows the metrics computed for each subcategory. The user may click each of the Job IDs to examine the linking terms that have been generated for the start term and the respective subcategory. Peptidomimetics scored the highest pR and were ranked first. Clicking on a Job ID opens a similar interface as for closed discovery. Arrowsmith

Table 2. Result of open discovery with Arrowsmith for the start term COVID-19 and the target MeSH category molecular mechanisms of pharmacological action.

Rank	Job ID	C-query	nC	nAC	nTot	nR	pR
1	1901117	Peptidomimetics [mh]	1160	3	2621	481	0.184
2	190118	Enzyme Inhibitors [mh]	50000	381	28846	5081	0.176
3	190111	Angiotensin Receptor Antagonists [mh]	17334	141	13995	2165	0.155
4	1901110	Fibrin Modulating Agents [mh]	38061	33	22974	3385	0.147
5	1901118	Radiopharmaceuticals [mh]	50000	7	22273	3184	0.143
6	19011	Alkylating Agents [mh]	17397	2	13492	1793	0.133
7	1901112	HIV Fusion Inhibitors [mh]	1213	0	2931	384	0.131
8	190114	Antimetabolites [mh]	50000	28	28015	3561	0.127
9	1901115	Neurotransmitter Agents [mh]	50000	16	27309	3355	0.123
10	1901113	Membrane Transport Modulators [mh]	50000	5	24221	2870	0.118
11	1901111	Heparin Antagonists [mh]	1288	0	2651	270	0.102
12	1901114	Mitosis Modulators [mh]	15530	4	11378	1122	0.099
13	190117	Enzyme Activators [mh]	2610	0	4604	453	0.098
14	190112	Antacids [mh]	6292	0	6840	657	0.096
15	1901116	Nitric Oxide Donors [mh]	6946	0	6697	591	0.088
16	190115	Antioxidants [mh]	50000	20	19749	1673	0.085
17	190113	Antifoaming Agents [mh]	192	0	793	59	0.074
18	190119	Enzyme Reactivators [mh]	1749	0	2742	191	0.070
19	1901119	Sequestering Agents [mh]	31538	2	16122	1044	0.065
20	190116	Cerumenolytic Agents [mh]	27	0	132	2	0.015

presents the user with the generated linking terms between the start term and the selected target term. As with closed discovery, the user may select one or more linking terms and inspect the articles containing the start and linking term and the linking and target term. Unfortunately, unlike with closed discovery, the linking terms can not be filtered by their semantic type. Clicking on the respective button results in an internal server error.

For COVID-19 and peptidomimetics, 2,621 linking terms were generated, 481 of which were predicted to be relevant. 445 articles occurred in both literature sets and were not included in the search for linking terms. Table 3 shows the first ten linking terms predicted to be relevant by Arrowsmith. Linking terms that could stimulate further research include 3c protease, furin, and proteasome inhibitor. For example, ten articles investigate the effects of 3C-like protease inhibition on SARS-CoV-2 and related coronaviruses (e.g., [27]). Four studies research peptidomimetics as 3C-like protease inhibitors [28,29,30,31].

The results demonstrate Arrowsmiths potential regarding the search for COVID-19 treatments. However, there are limitations to Arrowsmith that result from the underlying co-occurrence model. Although the system was improved by integrating MeSH categories and UMLS semantic type filtering, the system still generates many irrelevant linking terms. This is because co-occurrence-based models do not harness known semantic knowledge in biomedical literature. Furthermore, finding linking terms worth researching in open discovery is aggravated by semantic type filtering for linking terms not working at the moment.

Table 3. First ten linking terms generated by Arrowsmith using open discovery for the start term COVID-19.

Rank	Probability	Linking Term
1	0.99	respiratory syndrome coronavirus
2	0.99	molecular dynamic simulation
3	0.99	molecular docking
4	0.99	syndrome coronavirus
5	0.99	3c protease
6	0.99	dynamic simulation
7	0.99	docking study
8	0.99	docking molecular
9	0.99	furin
10	0.99	proteasome inhibitor

4.2. BITOLA

BITOLA was developed by Hristovski et al. [13] and uses association rule mining, a variant of the co-occurrence model [7], to identify relationships. BITOLA computes frequency and confidence measures for restricting and ranking terms. The retrieved terms can be filtered using MeSH semantic groups and types and the computed frequency and confidence. The use of MeSH turned out to be a limiting factor concerning the search for COVID-19 treatments. For the sake of comparability, we intended to examine the relationship between COVID-19 and remdesivir, as we did with Arrowsmith. However, although COVID-19 and remdesivir are MeSH, BITOLA does not allow to use them as start or target terms. We suspect that this is because COVID-19 and remdesivir are currently classified as MeSH Supplementary Concept Data, which BITOLA may not recognize. For example, SARS and chloroquine, which are classified as MeSH Descriptor Data, are recognized by BITOLA.

Therefore, we used SARS and chloroquine to demonstrate the use of BITOLA for closed discovery. SARS and chloroquine appeared together in two PubMed articles. In total, 1,676 linking terms were generated. We restricted the linking terms to enzymes, as we were interested in a potential inhibitory effect of chloroquine on enzymes involved in the replication of SARS. This decreased the number of linking terms to 32. Table 4 shows the ten linking terms that were ranked first.

Clicking on Frequency AB or Frequency BC performs a PubMed query for the A and B or B and C terms. Unfortunately, the user cannot determine which of the articles that the PubMed query returned linked the terms together. For example, clicking on Frequency AB of lactate dehydrogenase leads to a PubMed search that returns 246 articles. However, it cannot be determined which of these articles are part of the 19 articles that linked SARS to lactate dehydrogenase. This limits the tool's usability for further research, especially compared to Arrowsmith, which presents the user exactly the articles that linked two terms together.

To perform open discovery with BITOLA, the user must enter a MeSH as the start term. Because COVID-19 is not recognized as a MeSH by BITOLA, we used SARS instead. Without further restriction, BITOLA generated 2,848 linking terms. Since this is way too much for manual evaluation, we limited the linking terms to enzymes. This reduced the number of linking terms to 52, the first ten of which are shown in Table 5.

Table 4. First ten linking terms generated by BITOLA using closed discovery for the start term SARS and the target term chloroquine.

Linking Term	Semantic Type	Frequency AB	Frequency BC
Lactate Dehydrogenase	Enzyme	19	42
Cysteine Protease	Enzyme	8	31
Endopeptidases	Enzyme	4	41
Cathepsins	Enzyme	1	43
Alanine Transaminase	Enzyme	12	11
Aspartate Transaminase	Enzyme	7	20
RNA-Directed RNA Polymerase	Enzyme	7	1
paired basic amino acid cleaving enzyme	Enzyme	1	3
Peptide Hydrolases	Enzyme	2	46
ALANINE AMINOPEPTIDASE	Enzyme	5	1

Next, the user must select the linking terms that should be used to search for target terms. Unfortunately, there is no option to select or deselect all generated linking terms at once. Instead, they must be selected or deselected one by one. We selected lactate dehydrogenase, which was ranked first.

Table 5. First ten linking terms generated by BITOLA using open discovery for the start term SARS.

Linking Term	Semantic Type	Frequency	Confidence (%)
Lactate Dehydrogenase	Enzyme	19	0.768
ACE2 enzyme	Enzyme	12	0.485
Carboxypeptidase	Enzyme	12	0.485
Alanine Transaminase	Enzyme	12	0.485
Cysteine Protease	Enzyme	8	0.323
Creatine Kinase	Enzyme	8	0.323
Aspartate Transaminase	Enzyme	7	0.283
RNA-Directed RNA Polymerase	Enzyme	7	0.283
3C-like proteinase, Coronavirus	Enzyme	6	0.243
ALANINE AMINOPEPTIDASE	Enzyme	5	0.202

Without further restriction, the search generated 22,069 target terms. Therefore, we limited the target terms to pharmacologic substances, which include enzyme inhibitors. This reduced the number of linking terms to 2,538. Unfortunately, there is no option to further restrict the target terms, for example, to include enzyme inhibitors exclusively. Table 6 shows the first ten target terms generated for lactate dehydrogenase.

The example of COVID-19 as a search term has shown that the use of MeSH can limit the usability of BITOLA since the system did not recognize COVID-19. Another drawback to BITOLA is that terms may only be restricted using a limited subset of MeSH semantic groups and types. The system does not allow for further restriction using lower levels of the MeSH hierarchy. The open discovery search for enzyme inhibitors, which, despite the restrictions put in place, returned 451 articles dealing with lactate dehydrogenase and enzyme inhibitors, demonstrated this issue. An option to further limit the enzyme inhibitors to agents specifically targeting lactate dehydrogenase would have been helpful.

Table 6. First ten target terms generated by BITOLA using open discovery for the start term SARS and the linking term lactate dehydrogenase.

Target Term	Rank Freq	Rank Conf	Count Bs	Freq	Conf	"Discovery?"
Lactate	26068	3,4285	1	1372	4.463	YES
Adenosine Triphosphate	20368	2,6789	1	1072	3.487	YES
Lactic acid	9481	1,247	1	499	1.623	YES
Enzyme Inhibitors	8569	1,127	1	451	1.467	NO
Amino Acids	7676	1,0096	1	404	1.314	NO
Antioxidants	7296	,9596	1	384	1.249	NO
Superoxide Dismutase	7011	,9221	1	369	1.200	YES
Recombinant Insulin	6479	,8521	1	341	1.109	NO
Amylases	6213	,8172	1	327	1.064	YES
Hydrogen Peroxide	5662	,7447	1	298	0.969	YES

Furthermore, BITOLA's usability is limited because the user is not presented with the articles involved in term-linking. Instead, the user is referred to a general PubMed search for the terms involved, impairing the system's transparency. Finally, the most significant limitation to BITOLA remains its underlying co-occurrence model. Without the use of semantic knowledge, the system generates too many unrelated terms. To increase the quality of the generated terms, Hristovski et al. proposed to use discovery patterns. Although this is a promising approach, it requires external semantic parsers like SemRep and BioMedLEE to extract semantic predications from the scientific literature.

4.3. SemBT

To address some of BITOLA's issues, Hristovski developed SemBT (Semantic BITOLA), the semantic version of BITOLA, which takes advantage of semantic knowledge extracted from biomedical text with SemRep. Search queries are not entered in natural language but instead as "questions", consisting of subject, relation, and object. These questions refer to the different components of SemRep's semantic predications. At least one of the components must be specified, but two or all three may also be specified.

1. `Chloroquine`: Simple question with only one component specified. The concept chloroquine may be either the subject or the object. The question will return any biomedical concepts related to chloroquine.
2. `Chloroquine TREATS`: More specific question, where both a concept and a relation are specified. The concept chloroquine may be either the subject or the object, regardless of whether it is placed before or after the relation TREATS. The question will return any biomedical concepts that are related to chloroquine via the TREATS relation.
3. `Chloroquine TREATS Malaria`: Concrete question where all three components are specified. Both concepts may be either subject or object. The question will return any semantic relations that match the specified criteria.

The question is forwarded to Lucene, which means that full Lucene query syntax is allowed. Unfortunately, like BITOLA, SemBT currently neither recognizes COVID-19 nor related terms, limiting the tool's usability for searching for COVID-19 treatments.

The generated subjects and objects and the relations may be filtered using semantic types and relations. The semantic types must be abbreviated. The subject and object, and their semantic type, may be referred to explicitly using qualifiers. When it should not be distinguished between subject and object, the `arg` qualifiers may be used. The relation may also be referred to explicitly.

- `sub_name`: subject name
- `sub_semtype`: subject semantic type abbreviation
- `obj_name`: object name
- `obj_semtype`: object semantic type abbreviation
- `arg_name`: subject or object name
- `arg_semtype`: subject or object semantic type abbreviation
- `relation`: relation name

A fully qualified question making use of the qualifiers is shown below:

- `sub_name:Chloroquine sub_semtype:phsu relation:TREATS obj_name:Malaria obj_semtype:dsyn`: `phsu` refers to the abbreviated semantic type pharmacologic substance, `dsyn` to disease or syndrome.

Since SemBT uses SemRep for semantic predication extraction, the discovery patterns described in section 3.5 may be applied. To demonstrate the use of SemBT for closed discovery, we chose the *May_Inhibit* discovery pattern. Because SemBT does not allow for COVID-19 or related terms as arguments, we used chloroquine and malaria instead. For the question defining the AB literature, we qualified chloroquine as subject and INHIBITS as relation. As semantic type for chloroquine, we used organic chemical (`orch`), which includes drugs such as chloroquine. We limited the objects to the semantic types amino acid, peptide, or protein (`aapp`), and gene or genome (`gngm`). For the question that defines the BC literature, we set malaria as object with semantic type disease or syndrome (`dsyn`). We specified CAUSES as relation and limited the subjects to `aapp` and `gngm`. The AB and BC literature were specified using the following fully qualified questions:

1. `sub_name:Chloroquine sub_semtype:orch relation:INHIBITS obj_semtype:(aapp OR gngm)`
2. `sub_semtype:(aapp OR gngm) relation:CAUSES obj_name:malaria obj_semtype:dsyn`

SemBT found 215 AB relations, 112 BC relations, and 24 common linking terms. The first ten linking terms are shown in Table 7. The generated linking terms appear to be relevant for connecting chloroquine to malaria. However, tumor necrosis factor (TNF) and the CD4 gene are included redundantly in the list. The user may click any generated linking term to inspect the articles linking the start term to the target term. Other than BITOLA, SemBT presents the user the articles responsible for the linkage instead of referring to a general PubMed search for the terms involved.

With SemBT, open discovery must be performed in a different way than with BITOLA. While BITOLA allows the user to search for linking terms related to the start term and subsequently search for target terms related to one or more linking terms,

Table 7. First ten linking terms generated by SemBT using closed discovery for the start term chloroquine and the target term malaria.

Rank	Linking Term	Count AB	Count BC
1	TNF	4	4
2	Tumor Necrosis Factor-alpha	2	4
3	Antibodies	1	4
4	CD4	1	2
5	TNF gene	1	2
6	CD4 gene	1	2
7	cytokine	1	2
8	Proteins	1	2
9	Peptide Hydrolases	2	1
10	NOS2	1	1

SemBT requires the user to perform two separate searches. When searches performed with SARS produced no meaningful results, we turned to chloroquine and malaria again. To search for linking terms related to the start term malaria, we set malaria as object with semantic type `dsyn`. The subjects we limited to `aapp` and `gngm`. As the relation, we defined CAUSES. The full query was:

- `sub_semtype:(aapp OR gngm) relation:CAUSES obj_name:Malaria obj_semtype:dsyn`

Table 8. First ten linking terms generated by SemBT using open discovery for the start term malaria.

Linking Terms	Relation Type	Start Term	Frequency
cytokine	CAUSES	Malaria, Cerebral	12
cytokine	CAUSES	Malaria	10
Antibodies	CAUSES	Malaria	8
Tumor Necrosis Factor-alpha	CAUSES	Malaria	5
Tumor Necrosis Factor-alpha	CAUSES	Malaria	4
Intercellular adhesion molecule 1	CAUSES	Malaria, Cerebral	4
chemokine	CAUSES	Malaria, Cerebral	3
Genes	CAUSES	Malaria	3
Proteins	CAUSES	Malaria	3
TNF—TNF gene	CAUSES	Malaria	3

The search generated 22 linking terms. Table 8 shows the ten linking terms ranked first. We chose the linking term TNF to search for target terms. To search for agents that may inhibit TNF, we qualified it as object and set its semantic type to `gngm`. The subjects we limited to `orch`. As the relation, we specified INHIBITS. The resulting question was:

- `sub_semtype:orch relation:INHIBITS obj_name:Tumor Necrosis Factor-alpha obj_semtype:gngm`

SemBT generated 239 target terms. As shown in Table 9, chloroquine showed up on the fourth place. This example demonstrates SemBTs potential in searching for drug-

Table 9. First ten target terms generated by SemBT using open discovery for the linking term TNF.

Target Term	Relation Type	Linking Term	Frequency
Pentoxifylline	INHIBITS	TNF	32
Methotrexate	INHIBITS	TNF	9
Curcumin	INHIBITS	TNF	7
Chloroquine	INHIBITS	TNF	6
Ketamine	INHIBITS	TNF	5
vesnarinone	INHIBITS	TNF	4
Rolipram	INHIBITS	TNF	4
Aspirin	INHIBITS	TNF	4
triptolide	INHIBITS	TNF	4
Ethanol	INHIBITS	TNF	4

disease associations. However, the fact that COVID-19 and related terms are currently not allowed as arguments limits the tool's usability for searching COVID-19 treatments. Once this limitation is removed, SemBT could be a valuable tool to search for COVID-19 treatments.

5. Conclusion

Regarding LBD, we conclude that semantic models are the most suitable for searching for COVID-19 drug repurposing candidates. They use semantic knowledge derived from research literature with semantic parsers. The use of semantic models becomes especially appealing in the biomedical domain, which provides several knowledge resources like the UMLS and MeSH, as well as semantic parsers for knowledge extraction, such as SemRep and BioMedLEE.

We compared three existing LBD systems concerning their suitability for searching for novel COVID-19 treatments, Arrowsmith, BITOLA, and SemBT. Although Arrowsmith is based on a co-occurrence model, we believe it is the best choice at the moment. This is because both BITOLA and SemBT currently do not recognize COVID-19 or related terms, which makes them virtually useless for COVID-19 related searches. This limitation is probably caused by MeSH, which all systems use for restricting the allowed search terms. COVID-19 is currently classified as MeSH Supplementary Concept Data, which BITOLA and SemBT might not recognize. Once COVID-19 is classified as MeSH Descriptor Data, this limitation may disappear. If the restriction is removed, SemBT is the best choice, in our opinion.

SemBT uses a semantic model that incorporates SemRep for semantic knowledge extraction. This reduces the number of irrelevant terms generated by the system. SemBTs potential has been shown using the example of chloroquine and malaria. In open and closed discovery, the system identified relevant mechanisms of action involved in chloroquine treating malaria. Developing a new LBD system seems unnecessary. The existing systems viability has been proven by discoveries made and validated in the past, and the systems make good use of the knowledge resources available in the biomedical domain. Also, developing a new LBD system would take considerable time, which is sparse in the middle of a pandemic. Therefore, future work should instead focus on searching for new treatments for COVID-19 using existing systems.

References

[1] Scavone C, Brusco S, Bertini M, Sportiello L, Rafaniello C, Zoccoli A, et al. Current pharmacological treatments for COVID-19: What's next? Br J Pharmacol. 2020 Apr.

[2] Agostini ML, Andres EL, Sims AC, Graham RL, Sheahan TP, Lu X, et al. Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. mBio. 2018 Mar;9(2).

[3] Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million dollars? Health Aff (Millwood). 2006;25(2):420-8.

[4] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003 Mar;22(2):151-85.

[5] Boguski MS, Mandl KD, Sukhatme VP. Drug discovery. Repurposing with a difference. Science. 2009 Jun;324(5933):1394-5.

[6] Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc. 2006:349-53.

[7] Henry S, McInnes BT. Literature Based Discovery: Models, methods, and trends. Journal of Biomedical Informatics. 2017;74:20-32.

[8] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7-18.

[9] Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988;31(4):526-57.

[10] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 1997;91(2):183-203.

[11] Swanson DR, Smalheiser NR, Torvik VI. Ranking indirect connections in literature-based discovery: the role of medical subject headings. J Am Soc Inf Sci Technol. 2006 Sep;57(11):1427–1439.

[12] Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics. 2007 Jul;23(13):1658-65.

[13] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Improving literature based discovery support by genetic knowledge integration. Stud Health Technol Inform. 2003;95:68-73.

[14] Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the literature-based discovery paradigm to investigate drug mechanisms. AMIA Annu Symp Proc. 2007 Oct:6-10.

[15] Thilakaratne M, Falkner K, Atapattu T. A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. ACM Comput Surv. 2019 Dec;52(6).

[16] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.

[17] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan;32:267-70.

[18] Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012 Dec;28(23):3158-60.

[19] Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput. 2006:64-75.

[20] Symonds M, Bruza P, Sitbon L. The efficiency of corpus-based distributional models for literature-based discovery on large data sets. In: Proceedings of the Second Australasian Web Conference - Volume 155. AWC '14. Australian Computer Society, Inc.; 2014. p. 49-57.

[21] Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Methods Programs Biomed. 2009 May;94(2):190-7.

[22] Gordon CJ, Tchesnokov EP, Feng JY, Porter DP, Götte M. The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from middle east respiratory syndrome coronavirus. J Biol Chem. 2020 Apr;295(15):4773-9.

[23] Konkolova E, Dejmek M, Hřebabecký H, Šála M, Böserle J, Nencka R, et al. Remdesivir triphosphate can efficiently inhibit the RNA-dependent RNA polymerase from various flaviviruses. Antiviral Res. 2020 Aug;182:104899.

[24] Tchesnokov EP, Feng JY, Porter DP, Götte M. Mechanism of inhibition of ebola virus RNA-dependent RNA polymerase by remdesivir. Viruses. 2019 Apr;11(4).

[25] Brown AJ, Won JJ, Graham RL, Dinnon KH, Sims AC, Feng JY, et al. Broad spectrum antiviral remdesivir inhibits human endemic and zoonotic deltacoronaviruses with a highly divergent RNA dependent RNA polymerase. Antiviral Res. 2019 Sep;169:104541.

[26] McKee DL, Sternberg A, Stange U, Laufer S, Naujokat C. Candidate drugs against SARS-CoV-2 and COVID-19. Pharmacol Res. 2020 Jul;157:104859.

[27] Chen YW, Yiu CB, Wong KY. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res. 2020;9:129.

[28] Ang MJ, Lau QY, Ng FM, Then SW, Poulsen A, Cheong YK, et al. Peptidomimetic ethyl propenoate covalent inhibitors of the enterovirus 71 3C protease: a P2-P4 study. J Enzyme Inhib Med Chem. 2016;31(2):332-9.

[29] Zhai Y, Ma Y, Ma F, Nie Q, Ren X, Wang Y, et al. Structure-activity relationship study of peptidomimetic aldehydes as enterovirus 71 3C protease inhibitors. Eur J Med Chem. 2016 Nov;124:559-73.

[30] St John SE, Therkelsen MD, Nyalapatla PR, Osswald HL, Ghosh AK, Mesecar AD. X-ray structure and inhibition of the feline infectious peritonitis virus 3C-like protease: structural implications for drug design. Bioorg Med Chem Lett. 2015 Nov;25(22):5072-7.

[31] Chuck CP, Chen C, Ke Z, Wan DC, Chow HF, Wong KB. Design, synthesis and crystallographic analysis of nitrile-based broad-spectrum peptidomimetic inhibitors for coronavirus 3C-like proteases. Eur J Med Chem. 2013 Jan;59:1-6.