Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 115 (2021) 103696

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Drug repurposing for COVID-19 via knowledge graph completion

Rui Zhang [a,*], Dimitar Hristovski [b,1], Dalton Schutte [a,1], Andrej Kastrin [b,1], Marcelo Fiszman [c], Halil Kilicoglu [d]

[a] Institute for Health Informatics and Department of Pharmaceutical Care & Health Systems, University of Minnesota, MN, USA
[b] Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[c] NITES - Núcleo de Inovação e Tecnologia Em Saúde, Pontifical Catholic University of Rio de Janeiro, Brazil
[d] School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA

ARTICLE INFO

Keywords:
COVID-19
Drug repurposing
Knowledge graph completion
Literature-based discovery
Text mining

ABSTRACT

Objective: To discover candidate drugs to repurpose for COVID-19 using literature-derived knowledge and knowledge graph completion methods.

Methods: We propose a novel, integrative, and neural network-based literature-based discovery (LBD) approach to identify drug candidates from PubMed and other COVID-19-focused research literature. Our approach relies on semantic triples extracted using SemRep (via SemMedDB). We identified an informative and accurate subset of semantic triples using filtering rules and an accuracy classifier developed on a BERT variant. We used this subset to construct a knowledge graph, and applied five state-of-the-art, neural knowledge graph completion algorithms (i.e., TransE, RotatE, DistMult, ComplEx, and STELP) to predict drug repurposing candidates. The models were trained and assessed using a time slicing approach and the predicted drugs were compared with a list of drugs reported in the literature and evaluated in clinical trials. These models were complemented by a discovery pattern-based approach.

Results: Accuracy classifier based on PubMedBERT achieved the best performance ($F_1 = 0.854$) in identifying accurate semantic predications. Among five knowledge graph completion models, TransE outperformed others (MR = 0.923, Hits@1 = 0.417). Some known drugs linked to COVID-19 in the literature were identified, as well as others that have not yet been studied. Discovery patterns enabled identification of additional candidate drugs and generation of plausible hypotheses regarding the links between the candidate drugs and COVID-19. Among them, five highly ranked and novel drugs (i.e., paclitaxel, SB 203580, alpha 2-antiplasmin, metoclopramide, and oxymatrine) and the mechanistic explanations for their potential use are further discussed.

Conclusion: We showed that a LBD approach can be feasible not only for discovering drug candidates for COVID-19, but also for generating mechanistic explanations. Our approach can be generalized to other diseases as well as to other clinical questions. Source code and data are available at https://github.com/kilicogluh/lbd-covid.

1. Introduction

Coronavirus disease 2019 (COVID-19), caused by a novel coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly 2019-nCoV), first emerged in China in late 2019, and was declared a global pandemic by the World Health Organization (WHO) on March 11, 2020. Since then, COVID-19 has disrupted human life across the globe, with enormous human, economic, and societal costs. At the time of writing, it shows no sign of abating [1,2], although the final months of 2020 have brought some good news. First, on October 22, 2020, after the initial submission of this manuscript, the Food and Drug Administration (FDA) approved remdesivir for the treatment of COVID-19 requiring hospitalization [3]. Then, on November 9, 2020, Pfizer/BioNTech announced the effectiveness of their coronavirus vaccine BNT162b2 and over a month later after the release of additional data, FDA granted it emergency use authorization [4]. A second vaccine, by Moderna, has also been authorized for emergency use on December 18, 2020 [5].

Rapid development of effective vaccines for COVID-19 was by no means guaranteed, however. Moreover, *de novo* development and

* Corresponding author.
E-mail address: zhan1386@umn.edu (R. Zhang).
[1] Authors contributed equally.

https://doi.org/10.1016/j.jbi.2021.103696
Received 19 October 2020; Received in revised form 23 December 2020; Accepted 1 February 2021
Available online 8 February 2021

approval of an effective antiviral therapy remains a risky, costly, and time-consuming process. In the absence of an effective vaccine or other therapies, there have been significant efforts in repurposing drugs approved for other diseases for COVID-19 treatment, some of which have been tested in clinical trials (e.g., dexamethasone [6], hydroxychloroquine and lopinavir/ritonavir [7]) and one ultimately approved by FDA for treatment of patients hospitalized with COVID-19 (remdesivir [3,8]).

Computational approaches to drug repurposing have also garnered much attention to accelerate discovery of therapies for COVID-19 [9,10]. Common computational drug repurposing methods include drug signature matching, molecular docking, genome-wide association studies, and network analysis [11]. These data-driven approaches involve systematic analysis of various types of biological and clinical data (e.g., gene expression, chemical structure, genome and protein sequences, and electronic health records) to generate hypotheses regarding repurposed use of approved or investigational drugs [11]. The potential of recent advances in artificial intelligence (AI) and machine learning for COVID-19 drug repurposing has also been highlighted [12] and several studies using these techniques have reported promising results [13–16]. In particular, approaches leveraging network medicine [17] principles and biological knowledge graphs have been emphasized [12].

Most of the computational approaches to drug repurposing have focused on biological data, such as gene expression, protein-protein and drug-target interactions, and used SARS-CoV-2-related data. However, COVID-19-specific data is meaningful in the context of the larger body of diverse knowledge underpinning medicine and life sciences, a primary source of which is the biomedical literature. While some COVID-19 drug repurposing studies incorporated literature-based knowledge [13,16], their focus has remained largely COVID-19-specific. We argue that efficiently and safely repurposing drugs to treat COVID-19 requires more effective integration of literature-based knowledge with biological data collected via high-throughput methods.

In this paper, we propose a novel literature-based discovery [18,19] approach for COVID-19 drug repurposing. Similar to related work [16], we cast drug repurposing as a task of knowledge graph completion (or link prediction). We use a large, literature-derived biomedical knowledge graph constructed from SemMedDB [20] as well as COVID-19 research literature [21], as our data source. We use several state-of-the-art, neural network-based algorithms [22–26] for the task, and also complement these approaches with an approach based on discovery patterns [27]. Furthermore, we highlight the role of discovery patterns in search of mechanistic explanations for candidate drugs. Unlike most approaches that focus on COVID-19-specific knowledge [13,16], we consider a larger body of biomedical knowledge, as captured in the PubMed bibliographic database as well as in the COVID-19 research literature. Our results show that our approach can identify known drugs that have been used for COVID-19 and discover other novel drugs that can potentially be repurposed for COVID-19.

2. Related work

2.1. COVID-19 computational drug repurposing

Significant computational work has already been done to prioritize FDA-approved drugs for repurposing to treat COVID-19 [9,10]. For the most part, these studies can be categorized as molecular docking-based drug screening studies and network-based studies, the majority of them belonging to the former category. In molecular docking studies, small molecules in compound libraries are screened for effectiveness against the host proteins in the SARS-CoV-2 host interactome. Many studies of this kind have been reported, and some of the proposed drugs such as ritonavir, ribavirin, remdesivir, and oseltamivir have been used in practice and many are being evaluated in clinical trials [28–35].

While not as common as docking studies, network-based approaches to drug repurposing have also been explored. In one early study, a virus-related knowledge graph which consists of drug-target and protein-protein interactions and similarity networks from publicly available databases (e.g., DrugBank [36], ChEMBL [37], BioGRID [38]) was constructed and network-based machine learning and statistical analysis were used to predict an initial list of COVID-19 drug candidates. This list was narrowed down based on text mining from the literature and gene expression profiles from COVID-19 patients, and a poly-ADP-ribose polymerase 1 (PARP1) inhibitor CVL218, was proposed for therapeutic use against COVID-19 [13]. Cava et al. [39] used gene expression profiles from public datasets to construct a protein-protein interaction network in conjunction with pathway enrichment analysis to identify 36 potential drugs, including nimesulide, thiabendazole, and fluticasone propionate. In another study, network proximity analyses of drug targets and HCoV-host interactions in the human interactome were used to prioritize 16 potential repurposed drugs, including melatonin, mercaptopurine, and sirolimus, which were validated by enrichment analyses of drug-gene signatures and transcriptome data in human cell lines [14]. Potentially useful drug combinations (e.g., melatonin plus mercaptopurine) were also suggested. A follow-up study combined network medicine approaches based on human interactome with clinical patient data from a COVID-19 registry to show that melatonin was associated with reduced likelihood of a positive SARS-CoV-2 laboratory test [15]. The approach was further extended to explore deep learning [16]. A comprehensive knowledge graph of drugs, diseases, and proteins/genes (named CoV-KGE) was constructed by combining molecular interaction information from the literature with knowledge from DrugBank. A knowledge graph embedding model, named RotatE [23] was used to represent the entities and the relationships in the knowledge-based in low-dimensional vector space. Using the ongoing COVID-19 trial data as a validation set, 41 high-confidence repurposed drug candidates (including dexamethasone, indomethacin, niclosamine, and toremifene) were identified, and further validated via an enrichment analysis of gene expression and proteomics data in SARS-CoV-2-infected human cells. Another study used node2vec graph embeddings and variational graph autoencoders for the same purpose [40]. Gysi et al. [41] evaluated three algorithms (i.e., graph neural network, network proximity, and network diffusion) on a network of drug protein targets and disease-associated proteins for COVID-19 drug repurposing. While they obtained low correlations across the three algorithms, an ensembling approach that combined the predictions of all algorithms was shown to outperform the individual methods, ranking ritonavir, chloroquine, and dexamethasone among the most promising candidates. Some limited literature knowledge relevant to COVID-19 has been incorporated to network-based approaches; however, their focus has remained largely on structured molecular interaction information encoded in databases.

2.2. Literature-based discovery

Literature-based discovery (LBD) [18,19] is a method of automatic hypothesis generation pioneered by Swanson [42]. Based on the concept of “undiscovered public knowledge”, LBD seeks to uncover valuable hidden connections between disparate research literatures, and has been proposed as a potential solution for the problem of “research silos” (the view that scientific research areas are largely isolated from one another). The primary LBD paradigm is the so-called ABC model. In the *open discovery* form of this model, a relationship between two concepts A and B is known in one research area and another relationship between concepts B and C is known in another, and a potential relationship between concepts A and C is proposed. Conversely, in *closed discovery*, relationship AC is known, and a concept B is proposed as an explanation for the relationship AC. Extensions to ABC model have also been proposed, such as discovery browsing model that aims to elucidate more complex relationship paths between biomedical concepts [43,44]. Most applications of LBD have been in the biomedical domain, beginning with Swanson’s discovery of fish oil as a treatment for Raynaud disease [42],

a hypothesis supported subsequently by clinical studies. While early LBD systems focused primarily on term co-occurrence [45,46], semantic relations have been widely used in later years for representing scientific content of biomedical publications [27,47–49]. More recently, distributed vector representations based on term or semantic relation co-occurrence have been gaining popularity [50–52].

Drug repurposing has been one of the prominent applications of LBD [27,53–58]. For example, Hristovski et al. [27] used semantic discovery patterns following the ABC model to identify potential therapeutic uses for drugs. Zhang et al. [56] used discovery patterns and SemMedDB relations to identify potential prostate cancer drugs. Cohen et al. [55] used a vector representation approach based on semantic relations to predict a small number of active agents within a large library screened for activity against prostate cancer cells.

2.3. Knowledge graph completion

Knowledge graphs are represented as a collection of head entity-relation-tail entity triples (h, r, t), where entities correspond to nodes and relations to edges between them. Knowledge graph completion is the task of predicting unseen relations between two existing entities or to predict the tail entity given the head entity and the relation (or head entity given the tail entity and the relation). Recent approaches to knowledge graph completion rely on knowledge graph embedding methods [59], which learn a mapping from nodes and edges to continuous vector space that preserve the proximity structure of the knowledge graph and are amenable to application of machine learning methods. Such methods include translational models, which use distance-based scoring functions (e.g., TransE [22], TransH [60], RotatE [23]), and semantic matching models, which use similarity-based scoring functions (e.g., RESCAL [61], DistMult [24], ComplEx [25], and HolE [62]). Graph convolutional networks [63,64] as well as methods that use context-based encoding approach (e.g., KG-BERT [65], STELP [26]) have also been recently proposed. Knowledge graph embedding techniques based on a network of drug, disease, and gene/protein entities have been used to support drug repurposing for rare diseases [66]. Graph convolutional networks were used to model drug side effects resulting from drug-drug interactions [67]. For this purpose, a multimodal graph of protein-protein, drug-protein target, and drug-drug interactions was constructed from publicly available datasets. Sang et al. [68] constructed low-dimensional knowledge graph embeddings from SemMedDB relations and trained a Long Short-Term Memory (LSTM) model using known drug therapies from Therapeutic Target Database [69] to propose potential drugs using the trained model.

3. Materials and methods

In this section, we first describe our data sources and the preprocessing steps that were taken to construct a literature knowledge graph from these data sources. Next, we discuss the knowledge graph completion methods that we used to predict candidate drugs for COVID-19 as well as the discovery patterns used for providing mechanistic explanations. Lastly, we detail various evaluation schemes that we used to validate our predictions. A workflow diagram illustrating our approach is provided in Fig. 1. Our source code and data are publicly available at https://github.com/kilicogluh/lbd-covid.

3.1. Data

We constructed our biomedical knowledge graph primarily from SemMedDB [20], a repository of semantic relations automatically extracted from biomedical literature using SemRep natural language processing (NLP) tool [70,71]. SemRep-extracted relations are in the form of subject-predicate-object triples (also called *semantic predications*) and are derived from unstructured text in PubMed citations (i.e., titles and abstracts). For example, the triples `chloroquine-TREATS-Malaria` and `hydroxychloroquine-TREATS-Malaria` are extracted from the fragment *Chloroquine (CQ) and Hydroxychloroquine (HCQ) have been commonly used for the treatment and prevention of malaria* (PMID: 32910933). Subject and object arguments are normalized to concept unique identifiers (CUIs) in the UMLS (Unified Medical Language System) Metathesaurus [72,73]. Concepts are enriched with UMLS semantic type information (Disease or Syndrome, Pharmacologic Substance, etc.) and the relations are linked to the supporting article and sentence. SemMedDB has supported a wide range of computational applications, ranging from gene regulatory network inference [74] to *in silico* screening for drug repurposing [55] and medical reasoning [75], and has also found widespread use for literature-based knowledge discovery and hypothesis generation (e.g., [44,48,76–78]). In its most recent release (version 43, dated 8/28/2020),[2] SemMedDB contains more than 107M relations from more than 31M PubMed citations and 209M sentences. This release also includes COVID-19-related concepts and, thus, can serve as a knowledge graph for COVID-19 drug repurposing.

COVID-19 literature has grown at an unprecedented rate. LitCovid, NCBI's bibliographic database for COVID-19 literature [79] contains over 82K articles (as of 12/21/2020). An even richer dataset is the COVID-19 Open Research Dataset (CORD-19), which contains over 200K articles (including historical research on coronaviruses) [21]. Not all of these articles are included in PubMed. To ensure that our knowledge graph provides adequate coverage of COVID-19 knowledge, we included CORD-19 articles not included in PubMed, as well, and used SemRep to extract relations from titles and abstracts of these articles. We used CORD-19 release dated 09/25/2020.

SemMedDB distribution contains 107,645,218 relations among 339,638 concepts. CORD-19 dataset processed through SemRep contains 505,968 relations among 41,609 concepts.

3.2. Preprocessing

In this work, we focused on a subset of semantic relations derived from the combination of PubMed and CORD-19 datasets, predicted to be accurate and informative for drug repurposing.

First, we eliminated relations involving generic biomedical concepts (i.e., relations in which both subject and object were present in the `GENERIC_CONCEPT` table of SemMedDB such as `Pharmaceutical Preparations`) and relations with identical subject and object arguments. Next, we excluded a subset of predicate types that were not expected to be useful for drug repurposing, such as PART_OF and PROCESS_OF. The predicate types we used are AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, COEXISTS_WITH, COMPLICATES, DISRUPTS, INHIBITS, INTERACTS_WITH, MANIFESTATION_OF, PREDISPOSES, PREVENTS, PRODUCES, STIMULATES, and TREATS. Lastly, we also excluded the relations in which the subject or the object belongs to one of the following semantic groups: Activities & Behaviors, Concepts & Ideas, Objects, Occupations, Organizations, and Phenomena. The combined knowledge graph (SemMedDB + CORD-19) consists of 331,427 unique nodes and 20,017,236 relations.

In the second step, we eliminated (i) high-degree concepts using network degree centrality and (ii) uninformative semantic relations using log-likelihood ratio. The adjacency matrix A of a knowledge graph (i.e., directed network with multiedges) with n nodes (i.e., concepts) has entries $A_{ij} = 1$ if there is a relation from concept i to concept j. The in- and out-degrees of concept i can then be expressed as [80]:

$$k_i^{\text{in}} = \sum_{j=1}^{n} A_{ji} \quad \text{and} \quad k_i^{\text{out}} = \sum_{j=1}^{n} A_{ij}$$

[2] https://ii.nlm.nih.gov/SemRep_SemMedDB_SKR/SemMedDB/SemMedDB_download.shtml.

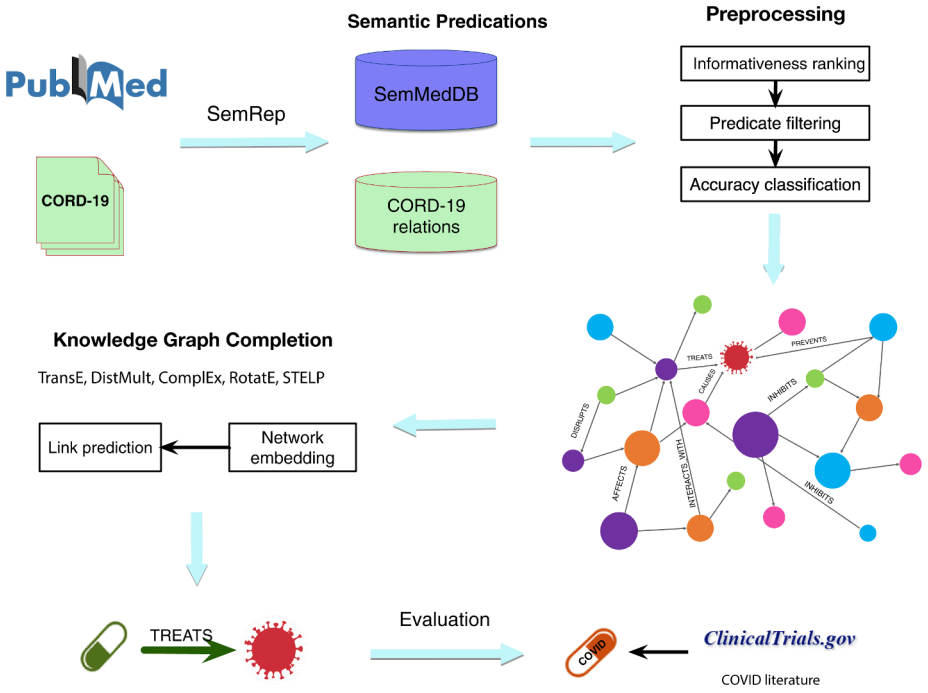

Fig. 1. Diagram illustrating the workflow of our approach.

To filter out uninformative links, we assigned each semantic relation a G^2 score indicating how strongly the terms within a triple are associated with each other [81]. A high G^2 score means that the observed and expected frequencies are significantly different, indicating that the triple is less likely to occur by chance. For computational purposes, we created two three-dimensional contingency tables with indices i, j, and k. The first table (OT) holds observed frequencies of a triple from the knowledge graph and the second table (ET) contains the expected values assuming independence of terms in each triple. G^2 was then calculated using the equation

$$G^2 = 2 \times \sum_{i,j,k} n_{ijk} \times \log\left(\frac{n_{ijk}}{m_{ijk}}\right), \quad m_{ijk} = \frac{\sum_i n_{jk} \times \sum_j n_{ik} \times \sum_k n_{ij}}{T^2},$$

where n_{ijk} is the cell i,j,k in OT, m_{ijk} is the cell i,j,k in ET, and $T = \sum n_{ijk}$.

Next, we normalized all three measures (G^2,k_i^{in}, and k_i^{out}) to the range $[0,1]$ and summed them up into a final score. The lower the score, the more specific and informative the relation is. For example, the relation `Operative Surgical Procedures-TREATS-Woman` which has a high score is more general than relation `interleukin-6-AFFECTS-Autoimmune Diseases`. We also kept all relations with biomedical concepts that refer to COVID-19 terms in the UMLS[3]:

- `C5203670:COVID19 (disease)`
- `C5203671:Suspected COVID-19`
- `C5203672:SARS-CoV-2 vaccination`
- `C5203673:Antigen of SARS-CoV-2`
- `C5203674:Antibody to SARS-CoV-2`
- `C5203675:Exposure to SARS-CoV-2`
- `C5203676:SARS-CoV-2`

We estimated that approximately 2.5M relations could be processed in reasonable amount of time with our GPU and eliminated relations with high final scores. At the end of the preprocessing stage, the knowledge graph consists of 131,355 nodes and 2,558,935 relations.

3.2.1. Accuracy classification

The precision of semantic predications generated by SemRep vary by domain (e.g., clinical relationships are more precise than molecular interactions). To improve the precision of the relations used for drug repurposing, we extended the SemRep accuracy classifiers previously proposed [82,83]. We fine-tuned a collection of transformer-based pretrained language models to classify semantic predications as correct vs. incorrect. These models include vanilla BERT (base size, cased and uncased) [84], BioBERT [85], BioClinicalBERT [86], BlueBERT [87], and PubMedBERT [88].

To extend the coverage of our existing classifiers, we used 6,492 predications annotated as correct vs. incorrect with respect to their source sentences. We leveraged 6,000 annotations from a previous study [83] (Cohen's $\kappa = 0.80$) and annotated 492 additional semantic predications. Annotation guidelines generated in the previous study was used. Two of the authors (HK and MF) and two health informatics graduate students annotated predications containing predicates of interest absent in the prior study (Fleiss' $\kappa = 0.41$, indicating moderate agreement). Fleiss' κ was used in this case, as more than two annotators were involved in annotation [89].

The resulting annotated set was split into 80/10/10 as training/validation/test sets. Hyperparameters were determined empirically and the learning rate was set to 1×10^{-5}, the batch size was 16, the maximum number of epochs was set to 10 but early stopping was employed. Optimization was done using the Adam optimizer [90] with decoupled weight decay regularization using betas (0.9, 0.999) and decay 0.01. The pooled output from the model was fed through a linear layer to produce logits that then underwent a softmax transformation to return class probabilities. A single Tesla V100 GPU was used to train the models. We compared the performance of various above-mentioned transformers. The best classifier was then used to filter incorrect semantic predications. This resulted in 1,016,124 relations being kept for the knowledge graph completion methods.

[3] https://metamap.nlm.nih.gov/Covid19Terms.shtml.

3.3. Knowledge graph completion

Consider a knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{E})$, where $\mathcal{E}$ refers to a set of entities, $\mathcal{R}$ denotes a set of possible relations, and $\mathcal{T}$ stands for a set of triples in the form (h)ead-(r)elation-(t)ail, formally denoted as $\{(h,r,t)\} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. The aim of knowledge graph completion is to infer new triples (h', r', t') such that $h', t' \in \mathcal{E}$ and $r' \in \mathcal{R}$. In this setting, the knowledge graph completion problem could be represented as a ranking task in which a prediction function $\psi(h,r,t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \mapsto \mathbb{R}$ which generates higher scores for true triples and lower scores for false triples is learned.

We explored three classes of knowledge graph completion methods: TransE [22] and RotatE [23] for translational models, DistMult [24] and ComplEx [25] for semantic matching models, and STELP [26] for context-based encoding. These methods differ in the way that they encode entities and relations in a knowledge graph into a low-dimensional vector space (i.e., knowledge graph embedding). Such distributed vector representations can be used for downstream reasoning and machine learning tasks.

3.3.1. Translational models (TransE and RotatE)

TransE [22] describes a triplet (h,r,t) as a translation between head entity h and tail entity t through relation r in a continuous vector space, i.e., $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$ is the embedding of h, r, and t, respectively. To measure plausibility of relations, TransE employs a distance-based score function $s(h,r,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$. Either L_1 or L_2 norm can be employed. Fig. 2 illustrates TransE model with two-dimensional embedding.

We choose TransE because of its simplicity and good prediction performance. However, TransE is able to model only one-to-one relations and fails to embed one-to-many, many-to-one, and many-to-many relations. To solve this problem, several other solutions have been proposed including RotatE [23]. RotatE treats each relation in a complex vector space as a rotation from the head entity to the tail entity, i.e., $s(h,r,t) = \left|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\right|_{l_1}$, where $\circ$ is a Hadamard product. We selected RotatE as a counterpart to TransE, as TransE reportedly does not perform well on some data sets (e.g., FB15k benchmark data set [22], commonly used in knowledge graph completion), which require symmetric pattern modeling.

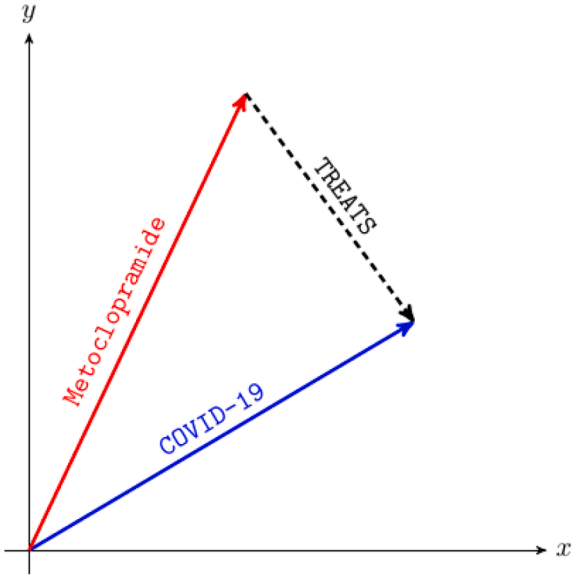

Fig. 2. TransE models relations as translations on a low-dimensional embedding of the entities. If (h,r,t) is true, the embedding of the tail entity t (i.e., `COVID-19`) should be close to the embedding of the head entity h (i.e., `Metoclopramide`) plus the vector that depends on the relationship r (i.e., TREATS).

3.3.2. Semantic matching models (DistMult and ComplEx)

DistMult [24] is the simplest approach among semantic matching models. Its scoring function is defined as $s(h,r,t) = \langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$. However, DistMult is limited only to symmetric relations, generating same scores for triples (h,r,t) and (t,r,h). ComplEx [25] extends DistMult to the complex domain. Head and tail embeddings for the same entity are complex conjugates, enabling ComplEx to model asymmetric relations. Its score function is defined as $s(h,r,t) = \mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \bar{\mathbf{t}} \rangle)$, where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^k$, $\mathrm{Re}(\cdot)$ is a real part of a complex vector, and k is dimension of an embedding.

Hyperparameters for both sets of models were tuned using the grid search on the validation set for each prediction model. We tuned the learning rate $\eta \in \{0.001, 0.01, 0.1\}$, number of hidden dimensions $k \in \{50, 100, 250, 400\}$, regularization coefficient $\lambda \in \{2 \times 10^{-6}, 2 \times 10^{-8}\}$, negative adversarial sampling $\in \{\texttt{True}, \texttt{False}\}$, fixed margin $\gamma \in \{1, 5, 10, 20\}$ for RotatE and norm $d \in \{L_1, L_2\}$ for TransE model.

3.3.3. Context-encoding models (STELP)

Semantic Triple Encoder for Link Prediction (STELP) [26], is a context-based encoding approach to knowledge graph completion. At its core is a Siamese BERT model that leverages sharing one set of weights across two models to produce encoded, contextual representations of the relations that are then fed to either a multi-layer perceptron (MLP) for classification or a similarity function for contrasting. The STELP architecture uses two learning objectives for training: triple classification and triple contrasting. The final learning objective is a linear combination of the two. During training, STELP takes a single positive relation (h,r,t) and produces five negative relations (h,r,t') by corrupting the tail. The head context, (h,r) term, is sent into one BERT model while each tail, (t) or (t'), is sent to the other BERT model that shares weights with the other. The classification objective seeks to classify (h,r,t) as 1 and each (h,r,t') as 0 while the contrastive objective seeks to measure the distance between the contextual embedding of the head and tail portions in a learned semantic space (see Fig. 3). Formally, the classification loss and constrastive loss functions are as follows:

$$\mathcal{L}^c = \frac{-1}{|\mathcal{D}|} \sum_{tp \in \mathcal{D}} \frac{1}{1 + |\mathcal{N}(tp)|} \left(\log s^c + \sum_{tp' \in \mathcal{N}(tp)} \log \left(1 - s^{c'} \right) \right)$$

$$\mathcal{L}^d = \frac{1}{|\mathcal{D}|} \sum_{tp \in \mathcal{D}} \frac{1}{|\mathcal{N}(tp)|} \sum_{tp' \in \mathcal{N}(tp)} \max \left(0, \lambda - s^d + s^{d'} \right)$$

where $\mathcal{D}$ is the set of correct triples, $\mathcal{N}(tp)$ is the set of corrupted triples

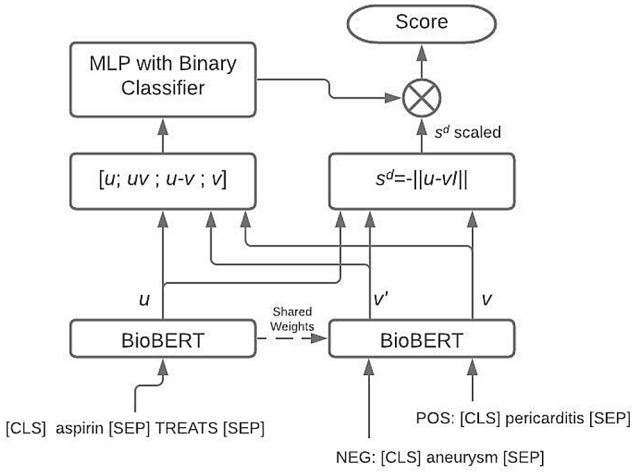

Fig. 3. Diagram for the high-level architecture of STELP.

for given positive triple tp, s^c and $(1 - s^{c'})$ are the positive class probability for tp and negative class probability for tp', respectively, λ is the margin size, s^d and $s^{d'}$ are the negative Euclidean distances between the contextual embeddings for the head and tail portions of the triple. The complete multi-objective loss function then is:

$$\mathcal{L} = \mathcal{L}^c + \gamma \mathcal{L}^d$$

where γ is a scaling factor for the contribution of the contrastive loss.

At inference, STELP considers every entity-context combination for a given partial relation, (h, r) to find (t) or (r, t) to find (h), and ranks every pair using the sum of the positive class probability and the scaled negative Euclidean distance.

We replaced the vanilla base BERT model proposed in the STELP paper with BioBERT, trained on biomedical literature corpora. The 1,016,124 unique relations remaining after preprocessing were each corrupted to produce five negative relations for a total of 5,080,620 negative relations and a grand total of 6,096,744 relations. The hyperparameters were set to the same values as in the original STELP paper and the learning rate was set to 1×10^{-5}, the batch size was 16, the contrastive loss scaling factor was 1.0. Optimization was done using Adam with decoupled weight decay with betas (0.9, 0.999) and decay 0.01. Training was run for 190,523 training iterations. Ranking was done by adding the scaled contrast score to the positive class probability and entities ordered in descending rank order.

3.3.4. Implementation of neural network models

All preprocessing was done using custom Bash and Python scripts. TransE, RotatE, DistMult, and ComplEx link prediction models were implemented in PyTorch using the DGL-KE package [91] for learning large-scale knowledge graph embeddings. The BERT models were based on HuggingFace BERT implementations using PyTorch. Pre-trained weights for BioBERT (BioBERT-Base v1.1 (+ PubMed 1 M)),[4] BioClinicalBERT,[5] PubMedBERT[6] and BlueBERT (BlueBERT-Base, Uncased, PubMed + MIMIC-III)[7] came from various sources associated with each paper. We implemented STELP ourselves using a combination of a HuggingFace BERT model and PyTorch.

3.4. Discovery patterns

Discovery patterns are defined as a set of constraints that need to be satisfied for the discovery of new relations between concepts [27]. Herein, we used discovery patterns for two purposes. First, we explored an open discovery pattern to identify drugs that can be repurposed for COVID-19. Second, we used the same pattern in closed discovery to propose plausible mechanisms for drugs identified via knowledge graph completion methods described above. Discovery patterns are expressed in terms of predication pairs (or predication chains). In particular, we focused on the following discovery pattern:

`DrugA`-INHIBITS|INTERACTS_WITH-`ConceptB` AND

`ConceptB`-AFFECTS|CAUSES|PREDISPOSES|ASSOCIATED_WITH-`COVIDConcept` AND NOT (`DrugA`-TREATS-`COVIDConcept`)

where `DrugA` is a drug concept with the semantic type Pharmacologic Substance and `COVIDConcept` refers to one of the following UMLS concepts (C5203670: `COVID-19`, C5203676: `2019 novel coronavirus`, C5203671: `suspected covid 19`). `ConceptB` can be any concept, and | indicates logical OR. When `DrugA` is unknown, this corresponds to an open discovery pattern. We used a Neo4j graph database of semantic relations and browser front-end for our exploration.

3.5. Evaluation

3.5.1. Ground truth generation

We semi-automatically generated a ground truth drug list, similar to the approach in other computational drug repurposing studies for COVID-19 [16]. We downloaded the interventions used in COVID-19 drug trials from clinicaltrials.gov using the following query: https://clinicaltrials.gov/ct2/results?cond=COVID-19&term=EXPAND[Term]+COVER[FullMatch]+AREA[InterventionType]+%22Drug%22.

This search yielded a set of 1167 clinical trials. We extracted all the interventions used in these studies and mapped the intervention terms to UMLS CUIs using MetaMap (v2016) [92] and filtered the resulting concepts by their semantic groups [93], keeping only those concepts with the semantic group Chemicals & Drugs. Additionally, we considered the semantic types Therapeutic Procedure and Gene or Genome, which also appeared for some concepts in intervention lists. We removed the duplicates from the resulting concept list and some general concepts (e.g., `Therapeutic procedure`, `Placebo`) as well as incorrect mappings. Drug concepts that only differ in their dosage or mode of administration were grouped together and considered a single element in the ground truth. For example, concepts `ruxolitinib`, `ruxolitinib Oral Tablet`, and `ruxolitinib 5 MG` were clustered together. This pruning and clustering process resulted in a final list of 283 concept clusters. The automatic evaluation described below was performed against this set.

3.5.2. Time slicing

Time slicing is an evaluation technique often used in LBD and link prediction tasks [18]. The idea is to train models on data prior to a specific date and test them on data after that date and evaluate whether links that formed only after the cutoff date can be predicted from the trained model. In this study, we trained our models on semantic relations extracted from publications dated 03/11/2020 or earlier and tested whether they can predict the drugs that have been proposed for COVID-19 since then or have been evaluated in clinical trials. This date was selected as cutoff, as it is the date on which WHO declared COVID-19 a pandemic. It is also a date by which enough biological knowledge about SARS-CoV-2 had accumulated, although COVID-19 therapies were still in their infancy, making it a suitable cutoff for time slicing experiments.

All five knowledge graph completion models were automatically assessed using an evaluation protocol proposed by Bordes et al. [22]. Suppose that $\mathcal{X}$ is a set of triples, Θ_E be the embeddings of entities $\mathcal{E}$, and Θ_R be the embeddings of relations $\mathcal{R}$. In the first, corruption step, we go through a set of triples and for each triple $\mathbf{x} = (h, r, t) \in \mathcal{X}$ replace its head and tail with all other entities in $\mathcal{E}$. Each triple is corrupted exactly $2|\mathcal{E}| - 1$ times. Formally, the corrupted triple is defined as:

$$\tilde{\mathbf{x}} = \bigcup_{h' \in \mathcal{E}} \left(h', r, t \right) \cup \bigcup_{t' \in \mathcal{E}} \left(h, r, t' \right),$$

where $h' \neq h$ and $t' \neq t$. We employ the filtered setting protocol not taking into account any corrupted triple that already appears in the knowledge graph. In the second, scoring phase, original and corrupted triples are tested using the constructed classifier ψ. Intuition behind this is that the model will assign a higher score to the original triple and a lower score to the corrupted triple. In the third, evaluation phase, the proposed models are assessed using three measures: mean rank (MR), mean reciprocal rank (MRR), and Hits@k measure. MR is an average rank assigned to the true relation, over all relations in a test set:

$$\mathrm{MR} = \frac{1}{2|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \left(\mathrm{rank}_i^h + \mathrm{rank}_i^t \right)$$

where rank_i^h and rank_i^t denote the rank position:

[4] https://github.com/naver/biobert-pretrained.
[5] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT.
[6] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract.
[7] https://github.com/ncbi-nlp/bluebert.

$$\text{rank}_i^h = 1 + \sum_{\widetilde{\mathbf{x}}_i \in \mathcal{C}^h(\mathbf{x}_i) \setminus \mathcal{G}} I[\psi(\mathbf{x}_i) < \psi(\widetilde{\mathbf{x}}_i)]$$

$$\text{rank}_i^t = 1 + \sum_{\widetilde{\mathbf{x}}_i \in \mathcal{C}^h(\mathbf{x}_i) \setminus \mathcal{G}} I[\psi(\mathbf{x}_i) < \psi(\widetilde{\mathbf{x}}_i)],$$

where the indicator function $I[P]$ is 1 iff P is true, and 0 otherwise.

MRR is the average inverse rank for all test triples and is formally computed as:

$$\text{MRR} = \frac{1}{2|\mathcal{T}|} \sum_{\mathbf{x}_i \in \mathcal{T}} \frac{1}{\text{rank}_i^h} + \frac{1}{\text{rank}_i^t}$$

Hits@k measures the percentage of relations in which the true triple appears in the top k ranked triples, where $k \in \{1, 3, 10\}$; formally:

$$\text{Hits@}k = \frac{100}{2|\mathcal{T}|} \sum_{\mathbf{x}_i \in \mathcal{T}} I\left[\text{rank}_i^h \leqslant k\right] + I\left[\text{rank}_i^t \leqslant k\right]$$

Our aim was to achieve low MR and high MRR and Hits@k.

3.5.3. Qualitative evaluation

We also performed a qualitative evaluation. One of the authors (MF, MD with a PhD in medical informatics) used Neo4j browser to assess the plausibility of some of the drugs highly ranked by the knowledge completion models, guided by literature search and review, using closed discovery. We also evaluated discovery patterns directly using open discovery. For this purpose, we issued a query for fifty drugs ranked on the number of intermediate `ConceptB` concepts between the drug and `COVIDConcept`. Then, MF assessed a subset of the candidate drugs for plausibility.

3.5.4. Comparison of candidate drug lists

We compared the drug lists proposed by our methods to each other, as well as to drug lists reported in three prior studies [14,16,94]. For TransE, which performed best, we identified a subset of plausible drugs from its top 150 candidate drug predictions. We used top 50 predictions from other knowledge graph completion methods as well as the top 50 drugs generated using the discovery pattern in open discovery mode.

4. Results

We report the performance of the semantic relation accuracy classifier as well as the knowledge graph completion methods in this section. We also provide a comparison of the drug lists proposed in previous studies and identified by our methods.

Table 1
Results of SemMedDB semantic relation classification using biomedical BERT variants.

	Vanilla BERT		BioBERT[a]	BioClinical BERT	PubMed BERT[b]	BlueBERT[c]
	Uncased	Cased	Cased	Cased	Uncased	Uncased
Validation set						
Rec	0.815	0.767	0.861	0.822	**0.896**	0.822
Pre	0.695	0.723	**0.762**	0.685	0.693	0.700
F_1	0.743	0.744	**0.808**	0.748	0.781	0.756
Test set						
Rec	0.815	0.782	0.842	0.832	**0.895**	0.845
Pre	0.795	0.815	**0.838**	0.804	0.816	0.782
F_1	0.805	0.798	0.840	0.818	**0.854**	0.812

Note: Rec = recall, Pre = precision. Results highlighted in bold are the best for each method.

[a] Trained on PubMed 1 M

[b] Trained on Abstracts + Full text

[c] Trained on PubMed + MIMIC

4.1. Accuracy classifier

The full table of results for the comparison of various BERT models for the accuracy classifier is included below (Table 1). The chosen model, PubMedBERT, obtained an F_1 score of 0.854 (recall = 0.895; precision = 0.816).

The best model (i.e., PubMedBERT) was then applied to the 2,558,935 predications. Of those, 1,907,717 (74.9%) were classified as correct predications and retained for use in the training of the downstream models.

This preprocessing yielded 115,451 unique biomedical concepts and 1,907,717 relations among them. The distribution of these predications are listed in Table 2.

4.2. Knowledge graph completion

The knowledge graph completion results for all employed models are presented in Table 2. For MR, a lower score is considered better; for all others, a higher score is considered better. The score for each method is the mean value over all triplets in the testing set.

On average, TransE outperforms all counterparts on all performance measures. Optimal TransE configuration was achieved with $k = 400$ hidden dimensions, L_1 norm, learning rate $\eta = 0.01$ and regularization coefficient $\lambda = 2 \times 10^{-8}$. Model training was limited to 20,000 epochs. Relatively small number of relations (15) ensure that all entities and relations can be smoothly embedded into the same vector space.

4.3. Embedding representation of knowledge graph

Next, we used t-SNE (t-distributed stochastic neighbor embedding) [95] algorithm to graphically represent embeddings of computed concepts in a two-dimensional space (Fig. 4). t-SNE algorithm enables reduction of high-dimensional data into a low-dimensional space such that similar concepts are presented by nearby points. The plot demonstrates relatively good co-localization of selected concepts, especially for `Suspected COVID-19` and `paclitaxel`.

4.4. Comparison of proposed drug lists

Thirty-three drugs (out of top 150) identified by TransE were deemed plausible after manual analysis (Table 3). Comparing this set to the repurposing proposals from three recently published papers [14,16,94], we find that there is one drug in common (estradiol) with the list in Zeng et al. [16]. On the other hand, Singh et al. [94] and Zeng et al. [16] have eight drugs in common and Zhou et al. [14] and Zeng et al. [16] have three. TransE predictions tended to contain more general drug classes (e.g., anthelmintics, antiplatelet agents), which were not specifically excluded, in contrast to previous methods. On the other hand, it is worth noting that specific drugs in some of these classes have been proposed in other studies. For example, TransE predicted anthelmintics as a candidate, while some of the drugs in this class (e.g., ivermectin, levamisole, nitazoxanide) have been proposed by others and tested in clinical studies. The same can be said about other drug classes, such as mTOR inhibitors and neuraminidase inhibitors.

Table 2
Distribution of semantic predications after filtering.

Predicate	Count (%)	Predicate	Count (%)
TREATS	518,267 (27.2%)	PRODUCES	38,602 (2.0%)
COEXISTS_WITH	420,633 (22.1%)	AUGMENTS	37,887 (2.0%)
INTERACTS_WITH	224,809 (11.8%)	PREVENTS	25,103 (1.3%)
CAUSES	205,441 (10.8%)	STIMULATES	24,734 (1.3%)
AFFECTS	192,092 (10.1%)	PREDISPOSES	18,613 (1.0%)
ASSOCIATED_WITH	106,418 (5.6%)	COMPLICATES	1,479 (0.1%)
INHIBITS	52,518 (2.8%)	MANIFESTATION_OF	1,156 (0.1%)
DISRUPTS	39,960 (2.1%)		

	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	**9.223**	**0.525**	**0.417**	**0.585**	**0.699**
DistMult	11.639	0.325	0.216	0.340	0.515
ComplEx	11.045	0.332	0.216	0.352	0.553
RotatE	10.864	0.377	0.246	0.428	0.633
STELP	22.960	0.073	0.000	0.027	0.234

Note: MR = mean rank, MRR = mean reciprocal rank. Results highlighted in bold are the best for each method.

Comparison of the 33 plausible drugs from TransE with the top 50 predictions from the other four knowledge graph completion methods revealed one common drug class between TransE and STELP (5-alpha reductase inhibitors) and five drugs between RotatE and STELP. DistMult and ComplEx did not share any predictions with the other methods.

Interestingly, using the discovery pattern in open discovery mode, we identified several drugs common with other methods: estradiol with TransE, paclitaxel with RotatE, as well as hydrocortisone and indomethacin with Zeng et al. [16]. Table 4 lists the overlapping candidate drugs from different methods and other studies.

5. Discussion

5.1. Knowledge graph completion models

Thus far, the following classes of drugs have been used for the management of COVID-19: antivirals, monoclonal antibodies, anti-inflammatory agents, immunomodulators, anticoagulants, and adjuvants [96,97]. In addition, several trials have studied antimalarials and antiparasites.

The knowledge graph completion models predicted drugs in all these classes, although they did not always rank them highly. For example, TransE predicted ribavirin (antiviral), trastuzumab (monoclonal antibody), indomethacin (anti-inflammatory), interferon beta-1b (immunomodulator), heparin (anticoagulant), vitamin D (adjuvant), metronidazole (antiparasite), and artemisone (antimalarial). Dexamethasone, one of the drugs considered most effective for reducing mortality in hospitalized patients, was the highest ranking drug from the RotatE model. Results from TransE and RotatE were a mix of individual drugs and drug classes (with little overlap), whereas STELP predictions were largely limited to very specific drugs and also included natural substances such as bioflavonoid quercetin and riboflavin (vitamin B2). While the quantitative evaluation against clinical trial data suggests

Table 3
Thirty-three candidate drugs highly ranked by TransE and deemed plausible in manual analysis.

Metoclopramide	Trilostane
Oxymatrine	Cyproterone Acetate
Mitogen-Activated-Protein Kinase Inhibitor	Nucleoside Reverse-Transcriptase Inhibitors
Oxophenylarsine	Methyltrienolone
5-Alpha reductase inhibitor	Bosentan
Folic acid	Estramustine
Anthelmintics	Allicin
Sildenafil	Proteasome inhibitors
Furosemide	Antiplatelet Agents
Beclomethasone	Fibrinolytic Agents
Cangrelor	Contraceptive Agents
Gymnemic acid	Neuraminidase inhibitor
Estradiol	Vitamin D Analogue
mTOR Inhibitor	Tyrosine kinase inhibitor
Clobetasol propionate	Mometasone furoate
Carbenoxolone	Vasopressin Antagonist
Anti-Retroviral Agents	

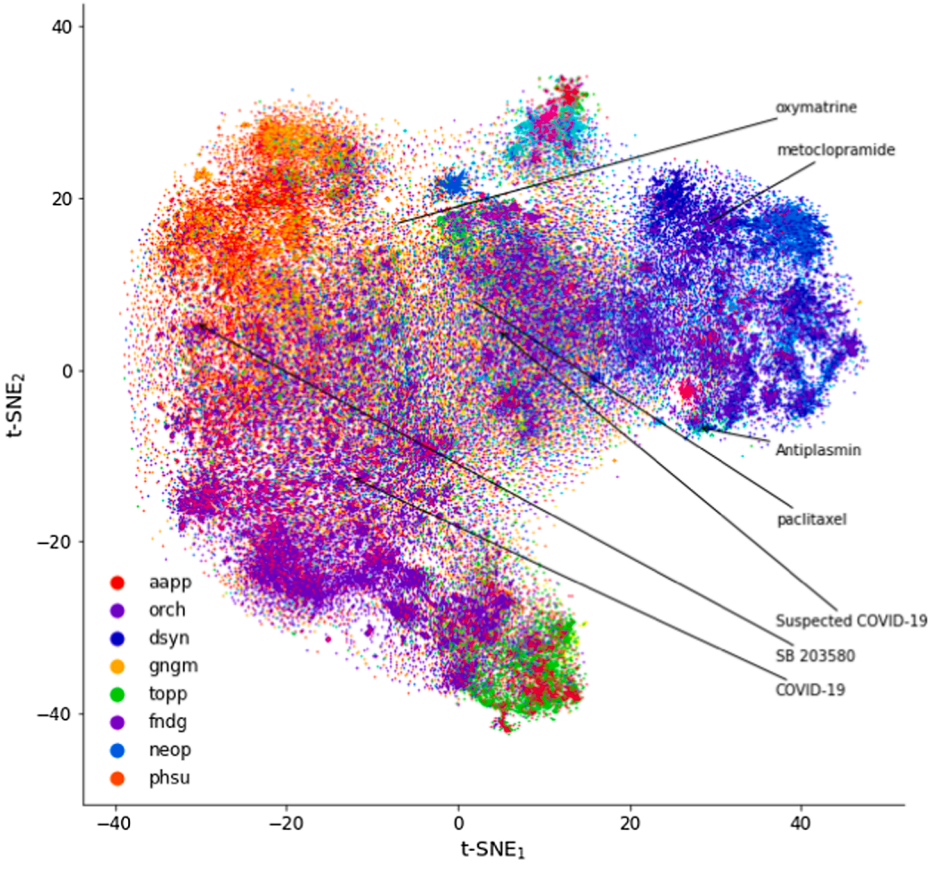

Fig. 4. Visualization of biomedical concepts learned by t-SNE (t-distributed stochastic neighbor embedding) algorithm and embedded in a two-dimensional space. We highlighted five drugs identified as potential new drugs to treat COVID-19. Color refers to semantic type of a particular concept; note that only the eight most frequent semantic types are presented. aapp: Amino Acid, Peptide, or Protein; dsyn: Disease or Syndrome; fndg: Finding, gngm: Gene or Genome; neop: Neoplastic Process; orch: Organic Chemical; phsu: Pharmacologic Substance; topp: Therapeutic or Preventive Procedure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of drug overlap between methods and studies.

Methods	Common Drugs
Zeng et al. [16]	Estradiol
TransE	
Discovery Patterns	
Zeng et al. [16]	Dexamethasone
Singh et al. [94]	
RotatE	
Zeng et al. [16]	Hydrocortisone
Discovery Patterns	Indomethacin
Zeng et al. [16]	Zidovudine
STELP	
TransE	5-alpha Reductase Inhibitors
STELP	
RotatE	Pibrentasvir
STELP	Anti-ILDR2 Monoclonal-Antibody BAY 1905254
	Mood Stabilizer
	Opium alkaloids and-derivative combination-cough suppressants
	Valoctocogene roxaparvovec
RotatE	Paclitaxel
Discovery Patterns	

Note: Drugs are from the top 50 ranked drugs from RotatE, STELP, the 33 drugs from TransE identified by MF as plausible, and the drugs specified in Zeng et al. [16], Zhou et al. [14], and Singh et al. [94]. We also use top 50 drugs identified using the discovery pattern in open discovery mode.

TransE as the best-performing model, it is worth noting that this only measures how well a method predicts drugs that are currently being trialed. It is difficult to assess the ultimate clinical effectiveness of the proposed drugs, and it is possible that models that do not perform as well quantitatively yield results that prove more promising (as in the case of RotatE and dexamethasone). Despite these issues, qualitative assessment of knowledge graph completion models showed that all methods could identify useful repurposing candidates.

The results indicate that more complex knowledge graph completion models might not be very efficient in drug repurposing tasks. Due to its relative simplicity, it might be expected that TransE be outperformed by its successors [23–25]. However, it showed efficiency in embedding a large-scale complex biomedical knowledge graph, such as the extended SemMedDB used here. On the other hand, differences in performances among DistMult, ComplEx, and RotatE were relatively small. All three models achieved low performance on MRR, Hits@1, and Hits@3 measures, and moderate score on Hits@10. Empirical evidence shows that DistMult and ComplEx usually perform well for high-degree entities, but fails with low-degree entities [98]. Because we eliminated highly frequent concepts due to their lack of informativeness, it is possible that this is reflected in lower performance of both models.

The context-encoding model, STELP, showed rather poor performance in evaluation. One possibility is that the model was only able to learn high-level groupings for the predicates. This is likely the case as it was observed the model produced much higher scores ($\text{MR} = 3.740$, $\text{MRR} = 0.867$, $\text{Hits@1} = 0.792$, $\text{Hits@10} = 0.969$) when evaluating a mix of corrupted triples containing other predicates in addition to TREATS. Thus, it may be the case that while the model can discriminate between what subjects are feasible for TREATS-COVID-19 versus AFFECTS-COVID-19 etc., it did not learn more granular features that allow it to differentiate between subjects within the context of TREATS-COVID-19. However, analysis of the t-SNE embedding and the qualitative evaluation show that the model mostly clustered the ground truth drugs into a couple of large clusters.

To further compare the drug rankings between TransE and STELP, we performed the Wilcoxon signed-rank test, which shows no correlation between how the two models were ranking novel relations ($p = 0.846$). Spearman's rank correlation between the novel relation rankings for both models was found to be -0.004, which further supports the results of the Wilcoxon test. Table 5 and Table 6 show that there is very little agreement between TransE and STELP, particularly in the top 1000 rankings for each model. It is worth noting that there were 47 items in common in the top 1000 rankings for both models.

5.1.1. Computational efficiency

The TransE and RotatE are much faster to train than the STELP model (approximately 15 min vs. 5 days on our dataset). Due to the size of BERT, which lies at the core of the architecture, STELP is a computationally expensive model which makes hyperparameter tuning difficult. This difficulty is compounded on the link prediction task which requires STELP to perform, just for inference, $\mathcal{O}((L/2)^2|\mathcal{V}|(1+|\mathcal{E}|))$ steps, where L is the sequence length, $\mathcal{V}$ the number of vertices, and $\mathcal{E}$ the number of edges. As the base BERT model contains 110 million parameters, adding in the scale of the link prediction may make the STELP and similar context-encoding based models infeasible for limited resource settings. TransE and RotatE demonstrate good results at a small fraction of the required computing power and time compared to the STELP model. Due to their reduced required computation time, it can be possible to explore larger graphs than that explored in this work. On the other hand, with adequately large computational resources, it may be possible to optimize STELP hyperparameters and train over multiple random seeds to generate a model that obtains better results than TransE or RotatE, which are limited by their smaller representational capacity.

5.2. Discovery patterns

Discovery patterns based on semantic relations provide an intuitive way of exploring potential mechanistic links between biological phenomena. Neo4j and its query language, Cypher, are powerful tools that complement semantic relations nicely in quickly pinpointing promising research directions, although massive graphs present some challenges for effective query and retrieval. In addition, a human expert is needed to sort out the noise in semantic relations (some of it obvious) due to NLP errors. However, given that predictions made by the knowledge completion models above are largely opaque, a human-in-the-loop discovery browsing approach based on patterns [43,44] remains an effective alternative to these more complex approaches, and also complements them by providing potential explanations. Given the size of the graph and time constraints, we limited ourselves to a single discovery pattern in this study and were able to both identify promising drugs (*open discovery*) and generate potential explanations for drugs predicted by the knowledge graph completion methods (*closed discovery*).

Using the open discovery pattern approach, we identified five promising drugs that were ranked highly and were not, to our knowledge, discussed in the literature (paclitaxel, SB 203580, alpha 2-antiplasmin, pyrrolidine dithiocarbamate, and butylated hydroxytoluene). The same approach also ranked highly some drugs and substances evaluated in clinical trials (e.g., quercetin, melatonin, vitamin D, estradiol, and simvastatin). We discuss below in more detail three that

Table 5
Statistics for absolute differences of TransE and STELP rankings.

	Median	Mean	Standard Deviation
Top 1000 TransE Rankings	10789.0	10567.140	6128.881
Top 1000 STELP Rankings	10224.0	10420.0	6002.522
All Rankings	6342.0	7207.910	5070.927

Note: The values for the first two rows are calculated by taking the top 1000 ranked triples for the specified model, calculating the absolute difference between the rankings from the two models for each of those triples, and calculating the statistics. For example, the triples that TransE ranked as the top 1000 triples were gathered, the absolute differences of rankings between TransE and STELP for those 1000 triples were calculated, and the statistics were calculated from those differences.

Table 6
Summary of absolute differences for TransE and STELP rankings. Semantic types are aapp: Amino Acid, Peptide, or Protein; gngm: Gene or Genome; orch: Organic Chemical; sosy: Sign or Symptom; topp: Therapeutic or Preventive Procedure.

Max Absolute Difference	Count (%)	Top 3 Most Common Semantic Types
0	1 (0.005%)	aapp
1	1 (0.005%)	aapp
3	5 (0.023%)	orch, topp, aapp
10	15 (0.070%)	gngm, aapp, orch
100	189 (0.877%)	gngm, aapp, orch
500	973 (4.516%)	gngm, aapp, orch
1000	1937 (8.990%)	gngm, aapp, orch

Note: Count column represents the number of triples where the two models rankings differed by at most the corresponding value in the Max Absolute Difference column. For example, there were 4 triples where both models rankings for those triples differed by at most 3.

have not been proposed for COVID-19 (i.e., paclitaxel, SB 203580, and alpha 2-antiplasmin). Notably, paclitaxel as well as quercetin, melatonin, vitamin D, estradiol, and simvastatin were predicted by the knowledge graph completion models. Fig. 5 shows a network resulting from the aforementioned discovery pattern generated by Neo4j browser.

We also used discovery patterns to generate mechanistic explanations for two other drugs ranked highly by TransE and deemed plausible. These drugs are metoclopramide and oxymatrine, which are also discussed in more detail below.

5.2.1. *Paclitaxel*

Paclitaxel is used to treat several cancer types, including ovarian cancer, breast cancer, lung cancer, cervical cancer, and pancreatic cancer. It stabilizes the microtubule polymer and protects it from disassembly, rendering chromosomes unable to achieve a metaphase spindle configuration. This blocks the progression of mitosis and prolonged activation of the mitotic checkpoint triggers apoptosis or reversion to the G0-phase of the cell cycle without cell division [99]. The following patterns support the paclitaxel discovery:

1. `paclitaxel-INHIBITS-interleukin-6-CAUSES-COVID-19`
2. `paclitaxel-INHIBITS-NF-kappa B-ASSOCIATED_WITH-COVID-19`
3. `paclitaxel-INHIBITS-interleukin-1, beta-ASSOCIATED_WITH-COVID-19`
4. `paclitaxel-INHIBITS-Granulocyte Colony-Stimulating Factor-ASSOCIATED_WITH-COVID-19`
5. `paclitaxel-INHIBITS-interleukin-10-PREDISPOSES-COVID-19`
6. `paclitaxel-INHIBITS-interleukin-8-PREDISPOSES-COVID-19`
7. `paclitaxel-INHIBITS-Thromboplastin-ASSOCIATED_WITH-COVID-19`
8. `paclitaxel-INTERACTS_WITH-TLR4-CAUSES-COVID-19`

The first six patterns support a role for paclitaxel in alleviating the

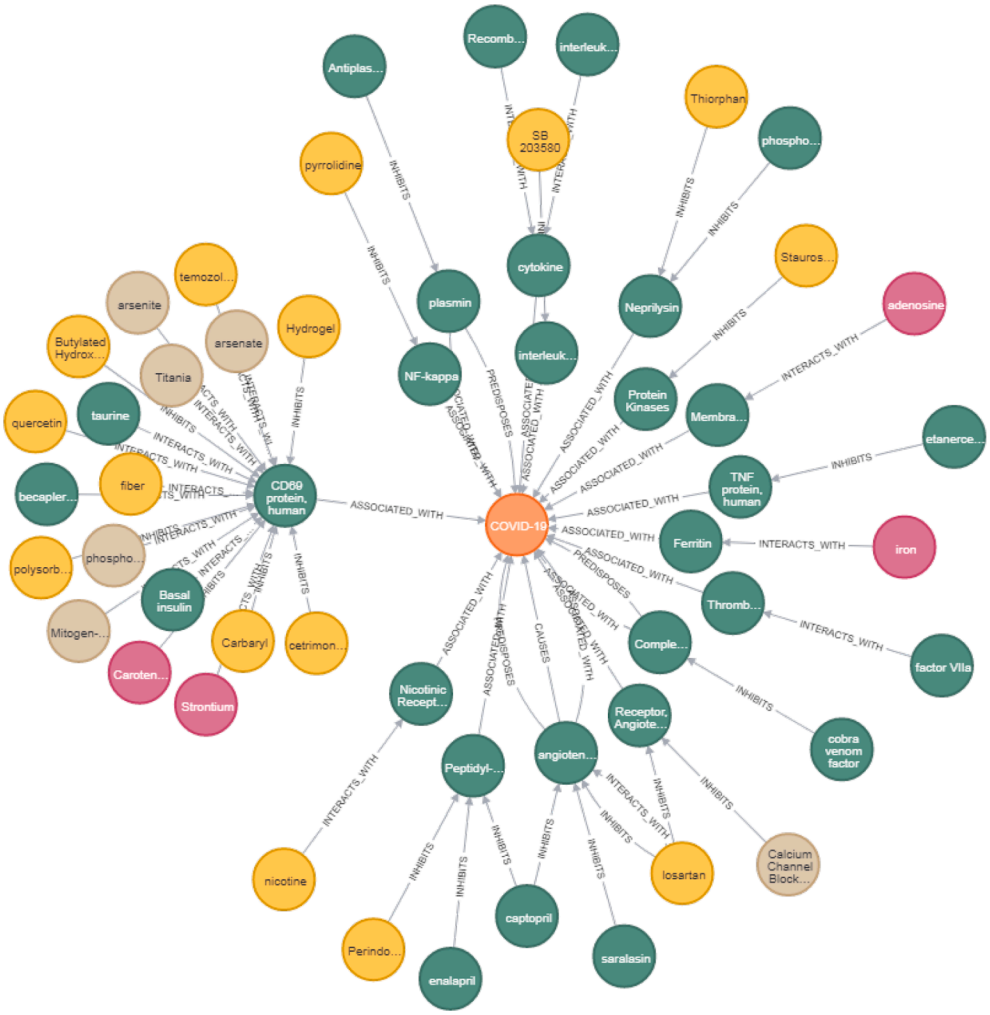

Fig. 5. Drug repurposing for COVID-19 with the open discovery pattern `DrugA-INHIBITS|INTERACTS_WITH-ConceptB` AND `ConceptB-AFFECTS|ASSOCIATED_WITH|CAUSES|PREDISPOSES-COVID-19`. The directionality is from the periphery (the predicted drugs) through the intermediate concepts to COVID-19 in the center.

cytokine storm of COVID-19, triggered by dysfunctional immune response and mediating widespread lung inflammation. Paclitaxel may plausibly help as an immunosuppressive therapy to immunomediated damage in COVID-19 [100]. Thromboplastin (pattern 7) is a complex enzyme found in brain, lung, and other tissues and especially in blood platelets and functions in the conversion of prothrombin to thrombin in the clotting of blood and may be elevated in patients with COVID-19. As pulmonary microvascular thrombosis plays an important role in progressive lung failure in COVID-19 patients, paclitaxel may reduce the state of hypercoagulability by acting as an inhibitor of thromboplastin [101]. The final pattern involves the interaction of paclitaxel with TLR4. Paclitaxel is known to have high affinity for TLR4 receptors. SARS-CoV-2 Spike protein binds with human innate immune receptors, mainly TLR4, increasing secretion of IL-6 and TNF-α and neuroimmune response. This suggests that paclitaxel may dislocate SARS-CoV-2 Spike proteins [102,103].

We note that paclitaxel, as a chemotherapy drug, is associated with adverse effects, some serious, such as neutropenia, leukopenia, alopecia, arthralgia, myalgia, and peripheral neuropathy [104].

5.2.2. SB 203580

SB 203580 is a specific inhibitor of p38α, which suppresses downstream activation of MAPKAP kinase-2, involved in many cellular processes including stress and inflammatory responses and cell proliferation. The following patterns support the SB 203580 discovery:

1. SB 203580-INHIBITS-interleukin-6 -CAUSES-COVID-19
2. SB 203580-INHIBITS-TNF protein, human-ASSOCIATED_WITH-COVID-19
3. SB 203580-INHIBITS-interleukin-1, beta-ASSOCIATED_WITH-COVID-19
4. SB 203580-INHIBITS-interleukin-8-PREDISPOSES-COVID-19
5. SB 203580-INHIBITS-NF-kappa B-ASSOCIATED_WITH-COVID-19
6. SB 203580-INHIBITS-Interleukin-1-CAUSES-COVID-19
7. SB 203580-INHIBITS-Granulocyte-Macrophage Colony-Stimulating Factor -ASSOCIATED_WITH-COVID-19
8. SB 203580-INHIBITS-Interleukin-17-ASSOCIATED_WITH-COVID-19
9. SB 203580-INHIBITS-Macrophage Colony-Stimulating Factor-ASSOCIATED_WITH-COVID-19

Similarly to paclixatel, all patterns involving SB 203580 point to a potential inhibition of the hyperinflammatory response in COVID-19. According to Gaestel [105], "the role of the protein kinases p38α in inflammation and innate immunity was found when the compound SB 203580 suppressed tumor necrosis factor (TNF) production in monocytes, and this resulted in inhibition of septic (inflammatory) shock."

5.2.3. Alpha 2-antiplasmin

Alpha 2-antiplasmin is a serine protease inhibitor responsible for inactivating plasmin. Elevated plasmin is a common risk factor for COVID-19 susceptibility, especially in patients with comorbidities such as hypertension, diabetes, and coronary heart disease [106]. The following patterns support the alpha 2-antiplasmin discovery:

1. Alpha 2-antiplasmin-INHIBITS-plasmin-PREDISPOSES-COVID-19
2. Alpha 2-antiplasmin-INHIBITS-fibrinogen-ASSOCIATED_WITH-COVID-19
3. Alpha 2-antiplasmin-INTERACTS_WITH-IgY-ASSOCIATED_WITH-COVID-19

More specifically, plasmin may cleave a newly inserted furin site in the S protein of SARS-CoV-2, which increases its infectivity and virulence in COVID-19. In addition, fibrinogen levels are higher in COVID-19 patients and may contribute to hypercoagulability [106]. By inhibiting plasmin and fibrinogen (first two patterns), alpha 2-antiplasmin may confer protection to COVID-19. In addition, pattern 3 suggests a mechanism of protection via immunoglobulin Y (IgY). In the immunology field, IgY against acute respiratory tract infection has been developed for more than 20 years. Several IgY applications have been effectively confirmed in both human and animal health. IgY antibodies extracted from chicken eggs have been used in bacterial and viral infection therapy. IgY production has been proposed as immunization as an adjuvant therapy in viral respiratory infection caused by COVID-19 infection [107]. Chicken immunized with alpha 2-antiplasmin and the peptide-specific antibody (IgY) was isolated from the egg yolks of hens that could be used as potential protections for COVID-19 patients [108].

5.2.4. Metoclopramide

Metoclopramide is used to relieve symptoms such as nausea, vomiting, and heartburn, caused by gastroesophageal reflux disease or diabetic gastroparesis. Metoclopramide is, mostly, a dopamine D2 antagonist but acts on many other neurotransmitters and proteins. Using our discovery pattern, we identified two pathways through which metoclopramide may protect against COVID-19.

1. metoclopramide-INTERACTS_WITH-cholinergic system-ASSOCIATED_WITH-COVID-19
2. metoclopramide-INHIBITS-TNF protein, human-ASSOCIATED_WITH-COVID-19

The first pattern suggests a cholinergic pathway for the protective effect of metoclopramide. The first relation of this pattern is extracted from a study which suggests that metoclopramide activates the sympathetic nervous system by mediating the central cholinergic system in humans [109]. The second piece of the evidence is based on a paper that explains how a cholinergic anti-inflammatory pathway acting through acetylcholine receptors can inhibit the production of pro-inflammatory cytokines [110]. Therefore, by activating the cholinergic pathway, metaclopropamide may prevent the inflammatory cytokine storm associated with COVID-19.

The second potential link is via tumor necrosis factor-α (TNF-α), a cytokine used by the immune system for cell signaling. The inhibitory effect of metoclopramide on TNF-α is suggested by a study on anti-inflammatory properties of benzamides, a class of drugs to which metoclopramide belongs ("Our data have shown that metoclopramide …gave dose dependent inhibition of TNFα" [111]). The second piece of the link comes from a paper that studies the role of TNF-α as a key driver of inflammatory macrophage response in severe COVID-19 and proposes anti-cytokine (especially, anti-TNF) treatment for COVID-19 [112].

5.2.5. Oxymatrine

Oxymatrine is a quinazine alkaloid with organ- and tissue-protective effects, primarily related to its anti-inflammatory, anti-oxidative stress, anti- or pro-apoptotic, anti-fibrotic, metabolism-regulating, and anti-nociceptive functions [113]. In addition, a variety of signal pathways, cells, and molecules are influenced by oxymatrine.

The following patterns support the repurposing of oxymatrine:

1. oxymatrine-INHIBITS-interleukin-6-PREDISPOSES-COVID-19
2. oxymatrine-INHIBITS-TNF protein, human-ASSOCIATED_WITH-COVID-19
3. oxymatrine-INHIBITS-interleukin-1, beta-ASSOCIATED_WITH-COVID-19
4. oxymatrine-INHIBITS-NF-kappa B-ASSOCIATED_WITH-COVID-19
5. oxymatrine-INTERACTS_WITH-interleukin-10-PREDISPOSES-COVID-19
6. oxymatrine-INHIBITS-TLR4 gene-ASSOCIATED_WITH-COVID-19

The first five patterns illustrate the effect of oxymatrine on proinflammatory cytokine and chemokine production induced by SARS-CoV-2. The first piece of the evidence is often an inhibitory relationship, as stated in Huang et al. [114]: "Oxymatrine at 120 mg/kg significantly

suppressed gene expressions of TLR-4 and NF-κB, decreased levels of TNF-α, interleukin-1beta and interleukin-6". The relationship between cytokine response and COVID-19 is well-established, for example as stated in Chi et al. [115]: "IL-6, IL-7, IL-10, …were found to be associated with the severity of COVID-19". Furthermore, the authors of the latter article propose that immunomodulatory treatment to regulate the cytokine responses could be an effective therapeutic strategy for SARS-CoV-2 infection. Oxymatrine could be one such candidate.

The relevance of TLR4 (Toll-like receptor 4) (pattern 6) for COVID-19, on the other hand, can be gleaned from Choudhury et al. [116], which states that "TLR4 may have a crucial role in the virus-induced inflammatory consequences associated with COVID-19." The authors further make the point that TLR4 antagonists (such as oxymatrine) could pave the way for COVID-19 treatment.

5.3. Error analysis

As error analysis, we manually examined the top 150 predictions by the best-performing model, TransE, for plausibility. 99 of these were deemed implausible, as they were drug classes that were considered too general. Examples of such classes include `C0003205: Anti-Infective Agents`, `C0003367: Antilipemic Agents`, and `C0010858: Cytostatic Agents`. As noted above, some members of these classes may indeed be plausible; however, the classes themselves were considered errors. A more systematic approach to exclude drug classes (e.g., by using MeSH concept hierarchy) could help reduce such errors. A more fine-grained evaluation could also consider such cases as partially correct, although this is unlikely to be useful for drug repurposing.

The other 18 candidates in the list that were deemed implausible are those that were classified as pharmacologic substances in UMLS, but were not drugs. These include `C0279328: hyperbaric oxygen`, `C1618233:husk` and `C1443923: Oral rehydration`, among others. It may be possible to leverage drug knowledge resources, such as DrugBank, to exclude such concepts and reduce errors.

5.4. Limitations and future work

Our approach relies on accuracy of the predications extracted by SemRep. SemRep precision is about 0.70 and its recall around 0.42 [71]. While the accuracy classifier helped us improve the accuracy of the predications used, the remaining errors were still significant, impacting the knowledge graph completion task.

In addition, despite aggressive filtering, the graph formed by the relations in extended SemMedDB is very large, making it difficult to apply computationally intensive models like STELP. In this study, we examined a sub-graph which, inevitably, results in a loss of information available to knowledge graph completion techniques. While we were still able to apply modeling techniques to a fairly large sub-graph focusing on drug repurposing, there exists a larger, complementary sub-graph that may provide further drug candidates.

As noted above, the TransE model benefited from hyperparameter tuning using a grid search method to find an optimal configuration. Similarly, STELP would likely benefit from a similar tuning to find an optimal configuration. For example, a single linear layer was used on the pooled output from the BioBERT model to produce the logits while increasing the representational capacity of the linear layer, by depth or width, might allow for STELP to develop a richer model of the underlying space formed by the BioBERT contextualized embeddings.

Our methods were limited to knowledge from the literature. Other types of biological data (e.g., protein-protein interactions, drug-target interactions, gene/protein sequences, pharmacogenomic and pharmacokinetic data) are likely to benefit identification of drug candidates, as shown to some extent by other studies [12], as well as our prior work [53]. However, the computational resources needed for training models based on such massive data can be prohibitive. TransE and similar methods seem more promising in that respect.

Lastly, with our *in silico* approach, we can of course only propose drug candidates for repurposing. To evaluate whether these drugs could indeed act as effective treatments for COVID-19, wet lab experiments and clinical studies are needed. However, the fact that we were able to identify drugs known to have some benefit for COVID-19 (e.g., dexamethasone) via purely computational methods that rely only on automatically extracted literature knowledge is encouraging. Moreover, the use of discovery patterns to explain why a particular drug or substance can be repurposed may be beneficial in prioritizing the most promising candidates for clinical studies.

6. Conclusion

In this study, we proposed an approach that combines literature-based discovery and knowledge graph completion for COVID-19 drug repurposing. Unlike similar efforts that largely focused on COVID-19-specific knowledge, we incorporated knowledge from a wider range of biomedical literature. We used state-of-the-art knowledge graph completion models as well as simple but effective discovery patterns to identify candidate drugs. We also demonstrated the use of these patterns for generating plausible mechanistic explanations, showing the complementary nature of both methods.

The approach proposed here is not specific to COVID-19 and can be used to repurpose drugs for other diseases. It can also be generalized to answer other clinical questions, such as discovering drug-drug interactions or identifying drug adverse effects.

As COVID-19 pandemic continues its spread and disruption around the globe, we are reminded how the spread of infectious diseases is increasingly common and future pandemics are ever more likely. Innovative computational methods leveraging existing biomedical knowledge and infrastructure could help us plan for, respond to, and mitigate the effects of such global health crises. Drug repurposing is a key piece of this response, and our approach provides an efficient computational method to facilitate this goal.

Funding

RZ and DS were supported by the U.S. National Institutes of Health's National Center for Complementary and Integrative Health (Grant No. R01AT009457). DH was supported by Slovenian Research Agency (Grant No. J5-1780, J5-2552, and P3-0154). AK was supported by the Slovenian Research Agency (Grant No. P3-0154, Z5-9352, and J5-2552).

CRediT authorship contribution statement

Rui Zhang: Conceptualization, Methodology, Writing - original draft. **Dimitar Hristovski:** Conceptualization, Methodology, Writing - original draft. **Dalton Schutte:** Methodology, Software, Writing - original draft, Formal analysis. **Andrej Kastrin:** Methodology, Software, Writing - original draft, Formal analysis. **Marcelo Fiszman:** Conceptualization, Methodology, Data curation, Validation, Writing - original draft. **Halil Kilicoglu:** Conceptualization, Methodology, Data curation, Validation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank François-Michel Lang, Leif Neve, and Jim Mork for their assistance with processing the CORD-19 dataset with SemRep and providing updates to SemMedDB. We acknowledge Tom Rindflesch for

his encouragement with the project.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2021.103696.

References

[1] Coronavirus disease (COVID-19), 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (Online; accessed 12/13/2020).

[2] Home - Johns Hopkins Coronavirus Resource Center, 2020. https://coronavirus.jhu.edu/ (Online; accessed 12/13/2020).

[3] FDA Approves First Treatment for COVID-19, 2020. https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-covid-19 (Online; accessed 12/21/2020).

[4] FDA Takes Key Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for First COVID-19 Vaccine, 2020. https://www.fda.gov/news-events/press-announcements/fda-takes-key-action-fight-against-covid-19-issuing-emergency-use-authorization-first-covid-19 (Online; accessed 12/21/2020).

[5] FFDA Takes Additional Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for Second COVID-19 Vaccine, 2020. https://www.fda.gov/news-events/press-announcements/fda-takes-additional-action-fight-against-covid-19-issuing-emergency-use-authorization-second-covid (Online; accessed 12/21/2020).

[6] R.C. Group, Dexamethasone in hospitalized patients with covid-19–preliminary report, N. Engl. J. Med. (2020).

[7] P. Horby, M. Mafham, L. Linsell, J.L. Bell, N. Staplin, J.R. Emberson, M. Wiselka, A. Ustianowski, E. Elmahi, B. Prudon, et al., Effect of Hydroxychloroquine in Hospitalized Patients with COVID-19: Preliminary results from a multi-centre, randomized, controlled trial, MedRxiv (2020), https://doi.org/10.1101/2020.07.15.20151852.

[8] J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, et al., Remdesivir for the treatment of Covid-19–preliminary report, New Engl. J. Med. (2020).

[9] O. Altay, E. Mohammadi, S. Lam, H. Turkez, J. Boren, J. Nielsen, M. Uhlen, A. Mardinoglu, Current status of COVID-19 therapies and drug repositioning applications, Iscience (2020) 101303.

[10] X. Wang, Y. Guan, COVID-19 drug repurposing: A review of computational screening methods, clinical trials, and protein interaction assays, Med. Res. Rev. (2020).

[11] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, et al., Drug repurposing: progress, challenges and recommendations, Nat. Rev. Drug Discov. 18 (1) (2019) 41–58.

[12] Y. Zhou, F. Wang, J. Tang, R. Nussinov, F. Cheng, Artificial intelligence in COVID-19 drug repurposing, Lancet Digital Health (2020).

[13] Y. Ge, T. Tian, S. Huang, F. Wan, J. Li, S. Li, H. Yang, L. Hong, N. Wu, E. Yuan, L. Cheng, Y. Lei, H. Shu, X. Feng, Z. Jiang, Y. Chi, X. Guo, L. Cui, L. Xiao, Z. Li, C. Yang, Z. Miao, H. Tang, L. Chen, H. Zeng, D. Zhao, F. Zhu, X. Shen, J. Zeng, A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19, bioRxiv (2020). doi:10.1101/2020.03.11.986836.

[14] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2, Cell Discov. 6 (1) (2020) 1–18.

[15] Y. Zhou, Y. Hou, J. Shen, A. Kallianpur, J. Zein, D.A. Culver, S. Farha, S. Comhair, C. Fiocchi, M.U. Gack, et al., A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for covid-19, ChemRxiv (2020), https://doi.org/10.26434/chemrxiv.12579137.v1.

[16] X. Zeng, X. Song, T. Ma, X. Pan, Y. Zhou, Y. Hou, Z. Zhang, K. Li, G. Karypis, F. Cheng, Repurpose open data to discover therapeutics for covid-19 using deep learning, J. Proteome Res. (2020).

[17] A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (1) (2011) 56–68.

[18] S. Henry, B.T. McInnes, Literature based discovery: models, methods, and trends, J. Biomed. Informat. 74 (2017) 20–32.

[19] Y. Sebastian, E.-G. Siew, S.O. Orimaye, Emerging approaches in literature-based discovery: techniques and performance review, Knowl. Eng. Rev. 32 (2017).

[20] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T.C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (23) (2012) 3158–3160.

[21] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R.M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D.A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A.D. Wade, K. Wang, N.X.R. Wang, C. Wilhelm, B. Xie, D.M. Raymond, D.S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, 2020.

[22] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 2787–2795.

[23] Z. Sun, Z. Deng, J. Nie, J. Tang, RotatE: Knowledge Graph Embedding by Relational Rotation in Complex sSpace, arXiv abs/1902.10197 (2019). http://arxiv.org/abs/1902.10197.

[24] B. Yang, W.-T. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575 (2014).

[25] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: International Conference on Machine Learning (ICML), 2016.

[26] B. Wang, T. Shen, G. Long, T. Zhou, Y. Chang, Semantic triple encoder for fast open-set link prediction, arXiv preprint arXiv:2004.14781 (2020).

[27] D. Hristovski, C. Friedman, T.C. Rindflesch, B. Peterlin, Exploiting semantic relations for literature-based discovery, in: AMIA Annual Symposium proceedings, 2006, pp. 349–353.

[28] D.E. Gordon, G.M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K.M. White, M. J. O'Meara, V.V. Rezelj, J.Z. Guo, D.L. Swaney, et al., A sars-cov-2 protein interaction map reveals targets for drug repurposing, Nature (2020) 1–13.

[29] L. Riva, S. Yuan, X. Yin, L. Martin-Sancho, N. Matsunaga, L. Pache, S. Burgstaller-Muehlbacher, P.D. De Jesus, P. Teriete, M.V. Hull, et al., Discovery of sars-cov-2 antiviral drugs through large-scale compound repurposing, Nature (2020) 1–11.

[30] C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, Q. Wang, Y. Xu, M. Li, X. Li, et al., Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods, Acta Pharmaceutica Sinica B (2020).

[31] A.A. Elfiky, Anti-hcv, nucleotide inhibitors, repurposing against covid-19, Life Sci. (2020), 117477.

[32] M. Kandeel, M. Al-Nazawi, Virtual screening and repurposing of fda approved drugs against covid-19 main protease, Life Sci. (2020), 117627.

[33] K. Al-Khafaji, D. AL-Duhaidahawi, T. Taskin Tok, Using integrated computational approaches to identify safe and rapid treatment for sars-cov-2, J. Biomol. Struct. Dyn. (2020) 1–11.

[34] J. Wang, Fast identification of possible drug treatment of coronavirus disease-19 (covid-19) through computational drug repurposing study, J. Chem. Inf. Model. (2020).

[35] A.A. Elfiky, Ribavirin, remdesivir, sofosbuvir, galidesivir, and tenofovir against sars-cov-2 rna dependent rna polymerase (rdrp): A molecular docking study, Life Sci. (2020), 117592.

[36] 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 Res. 36(suppl_1) (2008) D901–D906.

[37] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al., Chembl: a large-scale bioactivity database for drug discovery, Nucleic acids research 40 (D1) (2012) D1100–D1107.

[38] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, Biogrid: a general repository for interaction datasets, Nucleic acids research 34 (suppl_1) (2006) D535–D539.

[39] C. Cava, G. Bertoli, I. Castiglioni, In silico discovery of candidate drugs against covid-19, Viruses 12 (4) (2020) 404.

[40] S. Ray, S. Lall, A. Mukhopadhyay, S. Bandyopadhyay, A. Schönhuth, Predicting potential drug targets and repurposable drugs for covid-19 via a deep generative model for graphs, arXiv preprint arXiv:2007.02338 (2020).

[41] D.M. Gysi, Í. D. Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, H. Sanchez, R.M. Baron, D. Ghiassian, J. Loscalzo, et al., Network medicine framework for identifying drug repurposing opportunities for covid-19, arXiv preprint arXiv:2004.07229 (2020).

[42] D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1) (1986) 7–18.

[43] B. Wilkowski, M. Fiszman, C.M. Miller, D. Hristovski, S. Arabandi, G. Rosemblat, T.C. Rindflesch, Graph-based methods for discovery browsing with semantic predications, in: AMIA annual symposium proceedings, vol. 2011, American Medical Informatics Association, 2011, p. 1514.

[44] M.J. Cairelli, C.M. Miller, M. Fiszman, T.E. Workman, T.C. Rindflesch, Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox., in: AMIA Annual Symposium Proceedings, 2013, pp. 164–173.

[45] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery, Artif. Intell. 91 (2) (1997) 183–203.

[46] M. Weeber, H. Klein, L.T. de Jong-van den Berg, R. Vos, Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries, J. Am. Soc. Inform. Sci. Technol. 52 (7) (2001) 548–557.

[47] C.B. Ahlers, D. Hristovski, H. Kilicoglu, T.C. Rindflesch, Using the literature-based discovery paradigm to investigate drug mechanisms, in: AMIA Annual Symposium Proceedings, vol. 2007, American Medical Informatics Association, 2007, p. 6.

[48] J. Preiss, M. Stevenson, R. Gaizauskas, Exploring relation types for literature-based discovery, J. Am. Med. Inform. Assoc. 22 (5) (2015) 987–992.

[49] D. Cameron, R. Kavuluru, T.C. Rindflesch, A.P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, J. Biomed. Informat. 54 (2015) 141–157.

[50] T. Cohen, R. Schvaneveldt, D. Widdows, Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections, J. Biomedi. Informat. 43 (2) (2010) 240–256.

[51] T. Cohen, D. Widdows, R. Schvaneveldt, T.C. Rindflesch, Finding schizophrenia's prozac emergent relational similarity in predication space, in: International Symposium on Quantum Interaction, Springer, 2011, pp. 48–59.

[52] T. Cohen, D. Widdows, Embedding of semantic predications, J. Biomed. Informat. 68 (2017) 150–166.

[53] D. Hristovski, A. Kastrin, B. Peterlin, T.C. Rindflesch, Combining semantic relations and dna microarray data for novel hypotheses generation, in: Linking literature, information, and knowledge for biology, Springer, Berlin, Heidelberg, 2010, pp. 53–61.

[54] D. Hristovski, T. Rindflesch, B. Peterlin, Using literature-based discovery to identify novel therapeutic approaches, Cardiovasc. Hematol. Agents Medicinal Chem. (Formerly Curr. Medicinal Chem. Cardiovasc. Hematol. Agents) 11 (1) (2013) 14–24.

[55] T. Cohen, D. Widdows, C. Stephan, R. Zinner, J. Kim, T. Rindflesch, P. Davies, Predicting high-throughput screening results with scalable literature-based discovery methods, CPT: Pharmacometrics Syst. Pharmacol. 3 (10) (2014) 1–9.

[56] R. Zhang, M.J. Cairelli, M. Fiszman, H. Kilicoglu, T.C. Rindflesch, S.V. Pakhomov, G.B. Melton, Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs, Cancer Informat. 13 (2014). CIN-S13889.

[57] M. Rastegar-Mojarad, K.E. Ravikumar, D. Li, R. Prasad, H. Liu, A new method for prioritizing drug repositioning candidates extracted by literature-based discovery, in: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 669–674.

[58] H.-T. Yang, J.-H. Ju, Y.-T. Wong, I. Shmulevich, J.-H. Chiang, Literature-based discovery of new candidates for drug repurposing, Briefings Bioinformat. 18 (3) (2017) 488–497.

[59] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Trans. Knowl. Data Eng. 29 (12) (2017) 2724–2743.

[60] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes., in: AAAI, vol. 14, Citeseer, 2014, pp. 1112–1119.

[61] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data., in: ICML, vol. 11, 2011, pp. 809–816.

[62] M. Nickel, L. Rosasco, T. Poggio, Holographic embeddings of knowledge graphs, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 1955–1961.

[63] T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, arXiv preprint arXiv:1707.01476 (2017).

[64] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: European Semantic Web Conference, Springer, 2018, pp. 593–607.

[65] L. Yao, C. Mao, Y. Luo, Kg-bert: Bert for knowledge graph completion, arXiv preprint arXiv:1909.03193 (2019).

[66] D. Sosa, A. Derry, M. Guo, E. Wei, C. Brinton, R. Altman, A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 25 (2020) 463–474.

[67] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 34 (13) (2018) i457–i466.

[68] 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.

[69] X. Chen, Z.L. Ji, Y.Z. Chen, Ttd: therapeutic target database, Nucl. Acids Res. 30 (1) (2002) 412–415.

[70] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (6) (2003) 462–477.

[71] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with semrep, BMC Bioinformat. 21 (2020) 1–28.

[72] D.A.B. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods Inf. Med. 32 (1993) 281–291.

[73] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucl. Acids Res. 32 (Database issue) (2004) 267–270.

[74] G. Chen, M.J. Cairelli, H. Kilicoglu, D. Shin, T.C. Rindflesch, Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference, PLOS Comput. Biol. 10 (6) (2014) 1–16, https://doi.org/10.1371/journal.pcbi.1003666.

[75] S.R. Sukumar, L.W. Roberts, J.A. Graves, A Reasoning And Hypothesis-Generation Framework Based On Scalable Graph Analytics Enabling Discoveries In Medicine Using Cray Urika-XA And Urika-GD, 2016.

[76] A. Kastrin, T.C. Rindflesch, D. Hristovski, Link prediction on the semantic medline network, in: International Conference on Discovery Science, Springer, 2014, pp. 135–143.

[77] J. Sybrandt, A. Carrabba, A. Herzog, I. Safro, Are abstracts enough for hypothesis generation?, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 1504–1513.

[78] T.C. Rindflesch, C.L. Blake, M.J. Cairelli, M. Fiszman, C.J. Zeiss, H. Kilicoglu, Investigating the role of interleukin-1 beta and glutamate in inflammatory bowel disease and epilepsy using discovery browsing, J. Biomed. Semant. 9 (1) (2018) 25.

[79] Q. Chen, A. Allot, Z. Lu, Keep up with the latest coronavirus research, Nature 579 (7798) (2020) 193.

[80] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (4–5) (2006) 175–308.

[81] B.T. McInnes, Extending the log-likelihood measure to improve collocation identification, Master's thesis, Univerity of Minnesota, Minneapolis, MN, Dec 2004.

[82] R. Zhang, T.J. Adam, G. Simon, M.J. Cairelli, T. Rindflesch, S. Pakhomov, G. B. Melton, Mining biomedical literature to explore interactions between cancer drugs and dietary supplements, AMIA Summits Translat. Sci. Proc. 2015 (2015) 69.

[83] J. Vasilakes, R. Rizvi, G.B. Melton, S. Pakhomov, R. Zhang, Evaluating active learning methods for annotating semantic predications, JAMIA Open 1 (2) (2018) 275–282.

[84] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), 2019.

[85] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[86] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, in, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[87] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, in, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 58–65.

[88] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, arXiv preprint arXiv:2007.15779 (2020).

[89] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378–382.

[90] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[91] D. Zheng, X. Song, C. Ma, Z. Tan, Z. Ye, J. Dong, H. Xiong, Z. Zhang, G. Karypis, DGL-KE: Training knowledge graph embeddings at scale, arXiv preprint arXiv: 2004.08532 (2020).

[92] A.R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Informat. Assoc. (JAMIA) 17 (3) (2010) 229–236.

[93] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating UMLS semantic types for reducing conceptual complexity. Proc. Medinfo 10 (pt 1) (2001) 216–220.

[94] T.U. Singh, S. Parida, M.C. Lingaraju, M. Kesavan, D. Kumar, R.K. Singh, Drug repurposing approach to fight COVID-19, Pharmacol. Rep. 72 (6) (2020) 1479–1508, https://doi.org/10.1007/s43440-020-00155-6, http://link.springer.com/10.1007/s43440-020-00155-6.

[95] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[96] J.M. Sanders, M.L. Monogue, T.Z. Jodlowski, J.B. Cutrell, Pharmacologic treatments for coronavirus disease 2019 (covid-19): a review, Jama 323 (18) (2020) 1824–1836.

[97] W.J. Wiersinga, A. Rhodes, A.C. Cheng, S.J. Peacock, H.C. Prescott, Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review, JAMA 324 (8) (2020) 782–793.

[98] D.Q. Nguyen, T. Vu, T.D. Nguyen, D.Q. Nguyen, D. Phung, A capsule network-based embedding model for knowledge graph completion and search personalization, arXiv preprint arXiv:1808.04122 (2018).

[99] B.A. Weaver, How taxol/paclitaxel kills cancer cells, Mol. Biol. Cell 25 (18) (2014) 2677–2681.

[100] M.Z. Tay, C.M. Poh, L. Rénia, P.A. MacAry, L.F. Ng, The trinity of COVID-19: immunity, inflammation and intervention, Nat. Rev. Immunol. (2020) 1–12.

[101] W. Miesbach, M. Makris, COVID-19: coagulopathy, risk of thrombosis, and the rationale for anticoagulation, Clin. Appl. Thromb. Hemost. 26 (2020), 1076029620938149.

[102] S. Ran, The role of TLR4 in chemotherapy-driven metastasis, Cancer Res. 75 (12) (2015) 2405–2410.

[103] S.C.S. Brandão, J. d. O.X. Ramos, L.T. Dompieri, E.T.A.M. Godoi, J.L. Figueiredo, E.S.C. Sarinho, S. Chelvanambi, M. Aikawa, Is Toll-like receptor 4 involved in the severity of COVID-19 pathology in patients with cardiometabolic comorbidities?, Cytokine & Growth Factor Reviews (2020).

[104] DailyMed: Paclitaxel injection, 2020. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=9ffd3e34-537f-4f65-b00e-57c25bab3b01 (Online; accessed 12/21/2020).

[105] M. Gaestel, What goes up must come down: molecular basis of MAPKAP kinase 2/3-dependent regulation of the inflammatory response and its inhibition, Biol. Chem. 394 (10) (2013) 1301–1315.

[106] H.-L. Ji, R. Zhao, S. Matalon, M.A. Matthay, Elevated plasmin (ogen) as a common risk factor for COVID-19 susceptibility, Physiol. Rev. (2020).

[107] C. Constantin, M. Neagu, T.D. Supeanu, V. Chiurciu, D.A. Spandidos, IgY-turning the page toward passive immunization in COVID-19 infection, Exp. Therapeutic Med. 20 (1) (2020) 151–158.

[108] S.C. Lee, K.N. Lee, D.G. Schwartzott, K. Jackson, W.-C. Tae, P. McKee, Purification of human 2-antiplasmin with chicken IgY specific to its carboxy-terminal peptide, Preparative Biochem. Biotechnol. 27 (4) (1997) 227–237.

[109] Y. Takeuchi, T. Ikeda, S. Takeuchi, H. Ito, Y. Sugiyama, T. Matsukawa, S. Iwase, T. Mano, Effect of metoclopramide on muscle sympathetic nerve activity in humans, in: Environmental medicine: annual report of the Research Institute of Environmental Medicine, 37, Nagoya University, 1993, p. 95.

[110] Y. Tizabi, B. Getachew, R.L. Copeland, M. Aschner, Nicotine and the nicotinic cholinergic system in COVID-19, FEBS J. 287 (17) (2020) 3656–3663.

[111] R.W. Pero, B. Axelsson, D. Siemann, D. Chaplin, G. Dougherty, Newly discovered anti-inflammatory properties of the benzamides and nicotinamides, in: ADP-Ribosylation Reactions: From Bacterial Pathogenesis to Cancer, Springer, 1999, pp. 119–125.

[112] F. Zhang, J.R. Mears, L. Shakib, J.I. Beynor, S. Shanaj, I. Korsunsky, A. Nathan, A. M.P.R. Arthritis, et al., IFN- and TNF- drive a CXCL10+ CCL2+ macrophage phenotype expanded in severe COVID-19 and other diseases with tissue inflammation, bioRxiv.

[113] X. Lan, J. Zhao, Y. Zhang, Y. Chen, Y. Liu, F. Xu, Oxymatrine exerts organ-and tissue-protective effects by regulating inflammation, oxidative stress, apoptosis, and fibrosis: From bench to bedside, Pharmacol. Res. 151 (2020) 104541.
[114] M. Huang, Y.-Y. Hu, X.-Q. Dong, Q.-P. Xu, W.-H. Yu, Z.-Y. Zhang, The protective role of oxymatrine on neuronal cell apoptosis in the hemorrhagic rat brain, J. Ethnopharmacol. 143 (1) (2012) 228–235.
[115] Y. Chi, Y. Ge, B. Wu, W. Zhang, T. Wu, T. Wen, J. Liu, X. Guo, C. Huang, Y. Jiao, et al., Serum cytokine and chemokine profile in relation to the severity of coronavirus disease 2019 in China, J. Infectious Dis. 222 (5) (2020) 746–754.
[116] A. Choudhury, S. Mukherjee, In silico studies on the comparative characterization of the interactions of SARS-CoV-2 spike glycoprotein with ACE-2 receptor homologs and human TLRs, J. Med. Virol. (2020).