# Word Embedding Mining for SARS-CoV-2 and COVID-19 Drug Repurposing

Finn Kuusisto[1], David Page[2], and Ron Stewart[1]

[1]Morgridge Institute for Research, Madison, WI
[2]Duke University, Durham, NC

May 17, 2020

**Abstract**

The rapid spread of illness and death caused by the severe respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated coronavirus disease 2019 (COVID-19) demands a rapid response in treatment development. Limitations of de novo drug development, however, suggest that drug repurposing is best suited to meet this demand. Due to the difficulty of accessing electronic health record (EHR) data in general and in the midst of a global pandemic, and due to the similarity between SARS-CoV-2 and SARS-CoV, we propose mining the extensive biomedical literature for treatments to SARS that may also then be appropriate for COVID-19. In particular, we propose a method of mining a large biomedical word embedding for FDA approved drugs based on drug-disease treatment analogies. We find several drugs that have been suggested or are currently in clinical trials for COVID-19 in our top hits and present the rest as promising leads for further experimental investigation. We thus find our approach promising and present it, along with suggestions for future work, to the computational drug repurposing community at large as another tool to help fight the pandemic. Code and data for our methods can be found at `https://github.com/finnkuusisto/covid19_word_embedding`.

## 1 Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and associated coronavirus disease 2019 (COVID-19) were first identified in December of 2019 and have since spread to become a global pandemic[1]. This rapid spread of illness and death demands a rapid response in treatment development. De novo drug development, however, is slow, expensive, and suffers from low probability of success[2]. In contrast, drug repurposing, identifying new indications for existing drugs, offers the advantages of reduced time and risk to finding treatments. We thus propose that drug repurposing is the most promising approach to treatment development for this pandemic.

There are several strategies we could employ for drug repurposing. Certainly, getting access to the rapidly growing electronic health record (EHR) histories of those afflicted by COVID-19 could be enlightening. We could, for example, track patient recovery times and look for common prescription histories in those who recover sooner. Gaining access to sufficient EHR data would likely prove challenging though due to privacy concerns and limited data at individual institutions, not to mention the added administrative burden that might entail for an already strained health system. Given the similarity of SARS-CoV-2 to its predecessor SARS-CoV[3], we propose leveraging what we have learned about SARS in the intervening years. Specifically, we propose mining a word embedding built on biomedical literature published through early 2019 for candidate FDA approved drugs to treat SARS. Our results show that our proposed approach identifies several promising candidate drugs that have already been suggested or are already in clinical trials for COVID-19. We thus propose other candidate drugs identified by our method as potential leads for further investigation via in vitro and in vivo experimentation.

Table 1: The 20 closest word vectors to the SARS treatment analogy vectors *vector(Seed Drug) - vector(Seed Disease) + vector("SARS")*. All hits related to drugs or targets are highlighted in gray.

| Word Embedding Treatment Analogy Nearest Raw Hits | | |
|---|---|---|
| Metformin-Diabetes | Benazepril-Hypertension | Albuterol-Asthma |
| sars | sars | sars |
| sars-cov | sars-3cl | sars-cov |
| sars-3cl | sars-3clpro | csars |
| sars-3clpro | sars- | sars-covs |
| sars-like | sars-cov | sarspp |
| sars-covs | sars-covs | sars-like |
| sars-cov-induced | p-sars | sars-cov-like |
| sars-cov-mediated | sars-like | peramivir |
| sars-cov-like | sarsp | vero-pipecuronium |
| anti-sars-cov | sars-cov-like | sarsp |
| pcsars-cov | sars-hcov | pancuronium-metocurine |
| hsars-cov | anti-sars-cov | sars-hcov |
| sars-co | sars-s | sarse |
| anticoronaviral | coronavirion | pcsars-cov |
| cantharimide | lycodine | sars-3cl |
| sar405 | sarspp | p-sars |
| peramivir | sarse | sars-3clpro |
| norcantharidin-induced | sars-cov-s | sars- |
| cantharidin-mediated | sars-cov- | sars-coronavirus |
| delaviridine | pcsars-cov | pralidoxime |

In the following sections, we describe our word embedding source, our source and processing method for FDA approved drug names, and our approach to mining the word embedding for drugs to treat SARS. We then present our results and a discussion including manual evaluation of the top candidate drugs proposed by our method, followed by a conclusion and suggestions for future work.

# 2 Materials and Methods

In order to perform our word embedding mining for COVID-19 drug repurposing, we first need a word embedding. Furthermore, we need drug names to look for within the embedding. Here we briefly describe our sources for both the word embedding and drug names, we describe the data processing we perform on these sources, and we describe our methods for analysis. Code and data used for all of this analysis can be found at `https://github.com/finnkuusisto/covid19_word_embedding`.

## 2.1 Word Embedding

Rather than spend the time building our own word embedding on biomedical text, we instead searched the literature where there are several prebuilt biomedical word embeddings available. For this work, we chose the BioWordVec[4] prebuilt embedding, specifically the intrinsic model. We chose BioWordVec because it is the most recent available biomedical word embedding and has it performed well on several benchmark tasks.

In order to find a vector representation for COVID-19 treatments, we use a simple analogy approach. The original Word2vec publication demonstrated that the structure of a word embedding space could carry semantic meaning by showing that *vector("King") - vector("Man") + vector("Woman")* resulted in a vector closest to the word vector for *Queen*[5]. Effectively, this vector math asks the analogy *King* is to *Man* as what is to *Woman*? We use the same approach here, but instead use common drug-disease pairs as the seed analogy and SARS as the query disease. For example, one analogy we use is: *vector("Metformin") - vector("Diabetes") + vector("SARS")*. Effectively, we get the word vector analogy of *Metformin* is to *Diabetes* as what is to *SARS*? Note that the BioWordVec embedding we are using was published before SARS-CoV-2 was discovered and thus contains no reference to SARS-CoV-2 or COVID-19 in the vocabulary. Given, that SARS-CoV-2 is a strain of SARS-CoV[6], we use SARS as an approximation. To get a sense of analogy consistency, we use three separate drug-disease pairs as our seed treatment analogies: metformin and diabetes, benazepril and hypertension, and albuterol and asthma.

As a preliminary validation that our analogy vectors were close to reasonable results, we manually inspected the 20 closest vectors in the embedding vocab (see Table 1) for each of our seed drug-disease pairs. For this preliminary validation, we wanted to find potential drugs and drug targets near the analogy vector as we use these analogy vectors as our starting point for filtering to FDA approved drugs to treat COVID-19.

## 2.2 FDA Approved Drug Filtering

Given the urgency of the situation, we consider drug repurposing the most appropriate approach to finding treatments for COVID-19. We thus chose to tailor our treatment mining toward finding FDA approved drugs, allowing for the potential of off-label prescription in the short term. To get a list of approved drugs for our embedding analysis, we downloaded the FDA's approved drug database[7], extracted the drug names, and processed them for use in the word embedding.

To extract raw drug names from the FDA database, we first pulled all entries from the DrugName and ActiveIngredient fields of the Products table. We next manually inspected all raw entries that ended with parentheticals (e.g. "prempro (premarin;cycrin)") to identify entries that contain aliases or combinations versus those that contain tokens related to branding or packaging (e.g. "rogaine (for men)"). From these parentheticals, we manually collected additional drug names and then removed all parentheticals from the drug entries. These manually collected additional names included Ampicillin, Cycrin, Hydrocortisone, Premarin, Sulfabenzamide, Sulfacetamide, Sulfathiazole, Sulfadiazine, Sulfamerazine, and Sulfamethazine. We then split all of the entries by the semicolon character to separate drug names and ingredients entered as lists. Finally, we manually added back in those drugs and ingredients that were manually extracted from the deleted parentheticals. This gave us a list of 8,561 candidate approved drug names.

We next converted our candidate drug names into word vectors for ranking by their similarity with our treatment analogy vector. Here we simply split each candidate drug by white space and averaged the individual token vectors to get a final vector for the drug overall. When a token was not present in the embedding vocabulary, we simply dropped that token from the average and from the initial drug name. We used this approach rather than dropping a drug entirely to allow greater flexibility, for example if the embedding vocabulary is missing an ingredient from a combination drug. Finally, we removed duplicate drug names with the same tokens to account for exact duplicates and those with combinations stated in multiple orders. As a result, we successfully derived 5,833 distinct drug vectors from our initial 8,561 candidate drugs. We then sort these drug vectors by cosine similarity with our treatment analogy vectors and evaluate the closest hits.

As a preliminary validation that our approach can work to find useful drugs for diseases from treatment analogy vectors, we first considered major diseases and disease families with well-known treatments. Specifically, we used our treatment analogy vector approach to rank drugs for the query diseases Alzheimer's, allergies, and cancer (see Tables 2, 3, and 4). Note that we still used the same seed drug-disease pairs here (metformin-diabetes, benazepril-hypertension, and albuterol-asthma) but searched for analogous treatments for Alzheimer's, allergies, and cancer instead of SARS. For example, one analogy we used for initial validation is: *vector("Metformin") - vector("Diabetes") + vector("Alzheimer's")*. For this preliminary validation, we wanted to find drugs whose main indication is to treat the query disease in the top candidates. We chose these query diseases because they are fairly broad and have minimal treatment overlap with the seed drug-disease pairs that we used for the analogy. After initial validation of our method, we manually reviewed the top 50 drug candidates for SARS using the same method (see Tables 5, 6, and 7).

# 3 Results

Here we present results for validation of our word embedding mining approach along with results from applying our approach for COVID-19 drug repurposing. First, we present the 20 unfiltered closest embedding vocab vectors to our SARS treatment analogy vectors in Table 1 with all hits related to drugs or potential

Table 2: The top 10 candidate drugs for Alzheimer's from each of the three seed drug-disease analogies. Drugs with a primary indication for Alzheimer's are highlighted in gray.

| Top 10 Candidate Drugs for Alzheimer's from each Analogy | |
|---|---|
| Metformin-Diabetes | rivastigmine |
| | donepezil hydrochloride |
| | galantamine hydrobromide |
| | donepezil hydrochloride and memantine hydrochloride |
| | memantine hydrochloride |
| | selegiline |
| | rivastigmine tartrate |
| | rasagiline mesylate |
| | sulindac |
| | selegiline hydrochloride |
| Benazepril-Hypertension | rivastigmine |
| | aricept |
| | rivastigmine tartrate |
| | donepezil hydrochloride |
| | selegiline |
| | entacapone |
| | galantamine hydrobromide |
| | aricept odt |
| | memantine hydrochloride |
| | rasagiline mesylate |
| Albuterol-Asthma | galantamine hydrobromide |
| | rivastigmine |
| | donepezil hydrochloride |
| | rivastigmine tartrate |
| | memantine hydrochloride |
| | donepezil hydrochloride and memantine hydrochloride |
| | biperiden lactate |
| | exelon |
| | tacrine hydrochloride |
| | selegiline |

drug targets highlighted in gray. This highlighting is simply to verify that there are reasonable vocab vectors close to the analogy vectors. Five of the top 20 unfiltered hits are drug or target-related for the metformin-diabetes analogy, two of the top 20 hits are highlighted for the benazepril-hypertension analogy, and six of 20 hits are highlighted for the albuterol-asthma analogy.

Next, we present validation results for our approach to ranking FDA approved drugs for three diseases or disease families with well-established treatments. Specifically, we use the same three seed drug-disease pairs as analogies to find drugs for Alzheimer's, allergies, and cancer (see Tables 2, 3, and 4). All drugs with a primary indication for the query disease are highlighted in gray. This is to verify that our complete approach (drug vectors ranked by cosine similarity to treatment analogy vector) can identify effective ground-truth drugs for diseases that are not closely related to the seed disease-drug pair. In nearly every example, a vast majority (if not all) of the top 10 hits have a primary indication for the query disease.

Finally, we present the 50 closest FDA approved drugs to the treatment analogy vectors for SARS, thereby filtering to what may be the most promising drugs for repurposing. The top repurposing hits are presented in Tables 5, 6, and 7, and all drugs that have been suggested for or are currently under investigation for treatment of COVID-19 are highlighted in gray. This highlighting serves a partial evaluation of the repurposing via positive controls, suggesting that other hits may be good candidates for further investigation. We find 18 positive control hits out of 50 for the metformin-diabetes analogy, 11 of 50 for the benazepril-hypertension analogy, and five of 50 for the albuterol-asthma analogy. We present a Venn diagram of the overlap between the three analogies in Figure 1, and a table containing the drugs shared by all three and by at least two of the analogies in Table 8. Seven drugs are shared by all three analogies in their top 50 hits, and another 10 are shared by at least two of the analogies for a total of 17 higher confidence hits.

Table 3: The top 10 candidate drugs for allergies from each of the three seed drug-disease analogies. Drugs with a primary indication for allergies are highlighted in gray.

| Top 10 Candidate Drugs for Allergies from each Analogy | |
|---|---|
| Metformin-Diabetes | cetirizine hydrochloride allergy |
| | fexofenadine hydrochloride allergy |
| | zyrtec allergy |
| | rhinocort allergy |
| | xyzal allergy 24hr |
| | azelastine hydrochloride and fluticasone propionate |
| | loratadine |
| | cetirizine hydrochloride hives |
| | ketotifen fumarate |
| | fexofenadine hydrochloride hives |
| Benazepril-Hypertension | cetirizine hydrochloride allergy |
| | zyrtec allergy |
| | fexofenadine hydrochloride allergy |
| | rhinocort allergy |
| | cetirizine hydrochloride hives |
| | desloratadine |
| | loratadine |
| | fexofenadine hydrochloride hives |
| | acrivastine |
| | xyzal allergy 24hr |
| Albuterol-Asthma | albuterol |
| | cetirizine hydrochloride allergy |
| | fexofenadine hydrochloride allergy |
| | albuterol sulfate |
| | levalbuterol hydrochloride |
| | albuterol sulfate and ipratropium bromide |
| | diphenhydramine citrate |
| | diphenhydramine hydrochloride preservative free |
| | levalbuterol tartrate |
| | triprolidine pseudoephedrine hydrochloride and codeine phosphate |

# 4 Discussion

Here we review the validation results to demonstrate that our approach can find useful drugs for various diseases, followed by manual review of the FDA approved drug repurposing candidates for SARS. First, again note that several of the closest embedding vocabulary vectors to our treatment analogy vectors are related to drugs or drug targets. We find this result reassuring as it tells us that several of the top hits are at least within the category of results we want to find from these vectors. Looking deeper into these hits is further reassuring as they appear to not be any drugs or targets, but are in fact related to viral treatments, or SARS coronavirus treatments more specifically.

For example, “sars-3cl” and “sars-3clpro”, which show up for every analogy vector, likely refer to 3C-like proteinase, which is a major protease thought essential to viral replication of coronaviruses, including SARS-CoV and SARS-CoV-2[29, 30]. Peramivir, which shows up in two analogies, is an antiviral for influenza, and Delaviridine is a non-nucleoside reverse transcriptase inhibitor (NNRTI) for treatment of human immunodeficiency virus (HIV). Cantharimides are cantharidin derivatives, and cantharidin has been shown as potentially useful in therapy for hepatitis B[31]. Pancuronium and pipecuronium are both muscle relaxants that competitively inhibit the nicotine acetylcholine receptor, and both have use to aid in intubation[32, 33]. Pralidoxime is used to treat organophosphate poisoning via reactivation of acetylcholinesterase inhibited by the oranophosphorus agent[34]. Other interesting details to note are that delaviridine is a CYP3A4 inhibitor[35] like ritonavir, SAR405 has been studied in combination with hydroxychloroquine for cancer treatment[36], and lycodine-type alkaloids have demonstrated acetylcholinesterase inhibition which is of potential interest for treating Alzheimer’s[37]. The treatment analogy vectors are apparently in reasonable word vector neighborhoods.

Next, recall that we have used our drug ranking approach with the same seed analogy vectors for three

Table 4: The top 10 candidate drugs for cancer from each of the three seed drug-disease analogies. Drugs with a primary indication for cancer are highlighted in gray.

| Top 10 Candidate Drugs for Cancer from each Analogy | |
|---|---|
| Metformin-Diabetes | lapatinib<br>cisplatin<br>fulvestrant<br>bicalutamide<br>docetaxel<br>gefitinib<br>tamoxifen citrate<br>gemcitabine<br>erlotinib hydrochloride<br>toremifene citrate |
| Benazepril-Hypertension | bicalutamide<br>docetaxel<br>cisplatin<br>gemcitabine<br>exemestane<br>lapatinib<br>fulvestrant<br>erlotinib hydrochloride<br>gefitinib<br>carboplatin |
| Albuterol-Asthma | docetaxel<br>toremifene citrate<br>tamoxifen citrate<br>erlotinib hydrochloride<br>gemcitabine hydrochloride<br>cisplatin<br>bicalutamide<br>doxorubicin hydrochloride<br>gemcitabine<br>epirubicin hydrochloride |

major diseases with well-established ground-truth treatments. For the validation of our approach on drugs for Alzheimer's, nearly all of the drugs suggested from each analogy were drugs with primary indications for Alzheimer's, and several of the drugs seemingly incorrect drugs have a primary indication for Parkinson's, another neurodegenerative disease. We see a similar result for allergies where only the albuterol-asthma analogy suggests drugs not indicated for allergies in the top 10. Specifically, we see albuterol and levalbuterol show up several times, perhaps as a result of self seed drug bias. For the cancer drugs, we see that every drug is indicated for some form of cancer. This reassures us that our approach does, in fact, find drugs appropriate for the query disease even if the query disease have very little relationship with the seed drug-disease pair.

Finally, we manually reviewed every one of our top 50 FDA approved drugs suggested for repurposing with SARS as the query disease, and marked every one that has either been suggested for or is currently under investigation for treatment of SARS-CoV-2 and COVID-19. From the metformin-diabetes analogy, we find 18 of 50 drugs either suggested or under investigation for treatment against SARS-CoV-2 and COVID-19. With the benazepril-hypertension analogy, we find 11 of 50 hits, and from the albuterol-asthma analogy, we find five of 50. Across the analogies, seven hits are common to all three, and 10 are common to two of the three.

In the seven hits common to all, three have been suggested for treatment of SARS-CoV-2 and COVID-19. Amantadine and rimantadine are both adamantanes, which have been shown to have antiviral properties in vitro and have demonstrated possible protective effects in a clinical study of patients with neurological diseases[10, 11]. Zanamavir has been suggested based on in silico molecular docking models of the 3C-like proteinase previously mentioned[8].

In the 10 hits common to two of the analogies, three have been suggested for treatment of SARS-CoV-2 and COVID-19. Memantine is another adamantane similar to amantadine and rimantadine suggested by all

Table 5: Top 50 FDA approved drugs identified by word embedding mining with the Metformin-Diabetes analogy. Hits containing drugs suggested or under investigation for COVID-19 are highlighted in gray.

| Metformin-Diabetes as ?-SARS |
|---|
| gilteritinib fumarate |
| peramivir |
| zanamivir[8] |
| erdafitinib |
| atovaquone and proguanil hydrochloride[9] |
| rimantadine hydrochloride[10, 11] |
| delavirdine mesylate |
| atazanavir sulfate and ritonavir[12] |
| cobimetinib fumarate |
| niclosamide[13] |
| lopinavir and ritonavir[12] |
| temsirolimus[14] |
| rilpivirine hydrochloride |
| alectinib hnydrochloride |
| lefamulin acetate |
| perphenazine and amitriptyline hydrochloride[15] |
| alogliptin and metformin hydrochloride |
| tamiflu |
| selinexor[16] |
| amprenavir |
| ibuprofen and diphenhydramine citrate |
| olanzapine and fluoxetine hydrochloride |
| probenecid and colchicine[17] |
| erlotinib hydrochloride |
| bicalutamide |
| alomide |
| amantadine hydrochloride[10, 11] |
| azelastine hydrochloride and fluticasone propionate[18] |
| revefenacin |
| imipramine pamoate |
| doravirine |
| rosiglitazone maleate and metformin hydrochloride |
| nefazodone hydrochloride |
| mefloquine hydrochloride[19] |
| abacavir sulfate and lamivudine |
| carisoprodol compound |
| triprolidine and pseudoephedrine hydrochlorides codeine |
| soma compound codeine |
| chloroquine hydrochloride[20] |
| saquinavir mesylate[21] |
| linagliptin and metformin hydrochloride |
| nilutamide |
| donepezil hydrochloride and memantine hydrochloride[10, 11] |
| nelfinavir mesylate[22] |
| ceritinib |
| virazole[23] |
| vorinostat |
| triprolidine and pseudoephedrine hydrochlorides |
| fulvestrant |
| gefitinib |

Table 6: Top 50 FDA approved drugs identified by word embedding mining with the Benazepril-Hypertension analogy. Hits containing drugs suggested or under investigation for COVID-19 are highlighted in gray.

| Benazepril-Hypertension as ?-SARS |
|---|
| peramivir |
| tamiflu |
| zanamivir[8] |
| gilteritinib fumarate |
| rimantadine hydrochloride[10, 11] |
| benazepril hydrochloride |
| doravirine |
| galantamine hydrobromide |
| cetirizine hydrochloride hives |
| lanadelumab |
| aliskiren hemifumarate[24] |
| desloratadine |
| entacapone |
| invirase |
| daclatasvir dihydrochloride |
| indacaterol maleate |
| loratadine |
| peganone |
| nitazoxanide[25] |
| denavir |
| triprolidine and pseudoephedrine hydrochlorides codeine |
| rivastigmine |
| telavancin hydrochloride |
| donepezil hydrochloride |
| triprolidine and pseudoephedrine hydrochlorides |
| tazemetostat hydrobromide |
| relenza[8] |
| benazepril hydrochloride and hydrochlorothiazide |
| nulojix |
| ecallantide |
| alectinib hydrochloride |
| virazole[23] |
| levocetirizine hydrochloride |
| donepezil hydrochloride and memantine hydrochloride[10, 11] |
| amantadine hydrochloride[10, 11] |
| cetirizine hydrochloride |
| comtan |
| fluvoxamine maleate[26] |
| amlodipine besylate and benazepril hydrochloride[27] |
| delafloxacin meglumine |
| acrivastine |
| dalbavancin hydrochloride |
| fexofenadine hydrochloride hives[21] |
| rilpivirine hydrochloride |
| aricept |
| bendamustine hydrochloride |
| viramune xr |
| revefenacin |
| olodaterol hydrochloride |
| meloxicam |

Table 7: Top 50 FDA approved drugs identified by word embedding mining with the Albuterol-Asthma analogy. Hits containing drugs suggested or under investigation for COVID-19 are highlighted in gray.

| Albuterol-Asthma as ?-SARS |
|---|
| peramivir |
| albuterol |
| albuterol sulfate |
| albuterol sulfate and ipratropium bromide |
| zanamivir[8] |
| rimantadine hydrochloride[10, 11] |
| pralidoxime chloride |
| meperidine and atropine sulfate |
| amantadine hydrochloride[10, 11] |
| doxacurium chloride |
| biperiden lactate |
| atropine sulfate syringe |
| gallamine triethiodide |
| atropine and demerol |
| colistin sulfate |
| oseltamivir phosphate |
| revefenacin |
| dextromethorphan hydrobromide and quinidine sulfate |
| conivaptan hydrochloride |
| glycopyrronium tosylate |
| cefiderocol sulfate tosylate |
| fentanyl citrate and droperidol |
| pancuronium bromide |
| relenza[8] |
| telavancin hydrochloride |
| guaifenesin and dextromethorphan hydrobromide |
| diphenoxylate hydrochloride and atropine sulfate |
| esketamine hydrochloride |
| galantamine hydrobromide |
| naloxone hydrochloride and pentazocine hydrochloride |
| glycopyrrolate[28] |
| levalbuterol hydrochloride |
| calfactant |
| rilpivirine hydrochloride |
| pipecuronium bromide |
| tamiflu |
| biperiden hydrochloride |
| mivacurium chloride |
| metocurine iodide |
| ceftolozane sulfate |
| atropine sulfate |
| terbutaline sulfate |
| nesiritide recombinant |
| diphenoxylate hydrochloride atropine sulfate |
| tubocurarine chloride |
| benzonatate |
| rapacuronium bromide |
| naloxone hydrochloride |
| propoxyphene hydrochloride and acetaminophen |
| acetaminophen and pentazocine hydrochloride |

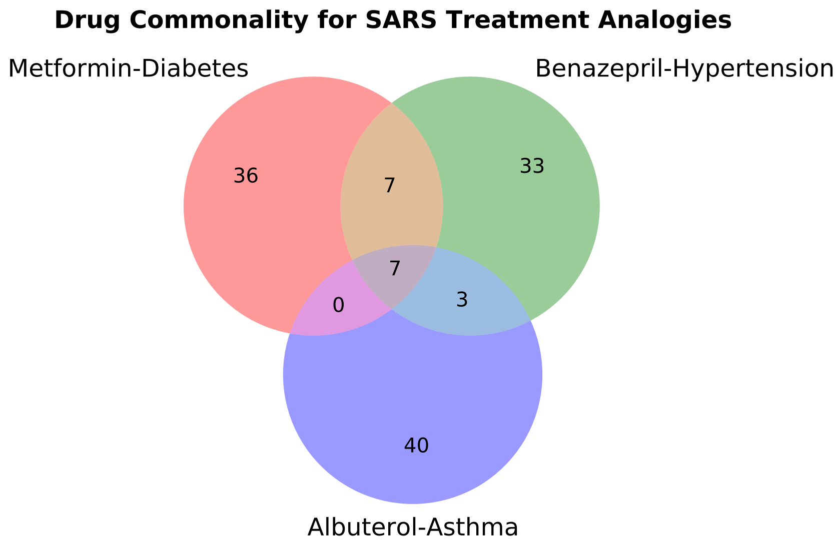


Figure 1: Venn diagram of the drug candidates identified by each SARS treatment analogy vector.

Table 8: The SARS drug repurposing candidates that are common to all three analogies, and those common to two analogies.

| Drug Repurposing Candidate Commonality for SARS | |
|---|---|
| Common to all | amantadine hydrochloride[10, 11] |
| | peramivir |
| | revefenacin |
| | rilpivirine hydrochloride |
| | rimantadine hydrochloride[10, 11] |
| | tamiflu |
| | zanamivir[8] |
| Common to two | alectinib hydrochloride |
| | donepezil hydrochloride and memantine hydrochloride[10, 11] |
| | doravirine |
| | galantamine hydrobromide |
| | gilteritinib fumarate |
| | relenza[8] |
| | telavancin hydrochloride |
| | triprolidine and pseudoephedrine hydrochlorides |
| | triprolidine and pseudoephedrine hydrochlorides codeine |
| | virazole[23] |

three analogies. Relenza is a trade name for zanamivir, so is essentially a duplicate, though it does perhaps suggest even more confidence in the drug. Virazole is a trade name for ribavirin, an antiviral which has shown antiviral activity against SARS-CoV-2 in vitro.

We also note that six of the proposed treatments are in clinical trials: atovaquone, lopinavir and ritonavir, sirolimus (suggested here as the prodrug temsirolimus), selinexor, colchicine, and fluvoxamine. Interestingly, these six drugs come from a wide range of primary indications including antiparasitic, antiviral, anti-inflammatory, anticancer, and antidepressant effects. Furthermore, the proposed drugs that are not currently in trials show a similar breadth of primary indication. Overall, we find that our approach shows a great deal of promise as it is able to discover a wide range of drugs that have elsewhere been proposed for COVID-19 from in silico, in vitro, and in vivo experimentation, all with literature published before SARS-CoV-2 was discovered.

## 5 Limitations

Of course, while our method appears promising, it is not without limitations. First, our method is limited to what has already been published in the scientific literature and cannot propose new drugs or treatments. We also caution readers that, in most cases, these drugs have not been tested for COVID-19 efficacy, and we make no claims other than that some of these drugs deserve further exploration. We can say with confidence that at least a few proposed drugs seem less promising. Peramivir, and Tamiflu (oseltamivir), are neuraminidase inhibitors used to treat influenza. While they are thus antivirals, coronaviruses do not use neuraminidase, so these particular drugs are perhaps less likely to be effective against SARS-CoV-2[18]. On the other hand, zanamivir, one of our common positive controls[8], is also a neuraminidase inhibitor and should thus be a less likely candidate. Given that the potential mechanism of action for zanamivir is based on computed binding to the 3C-like proteinase, perhaps some drugs may demonstrate efficacy outside of their traditional mechanism. Nevertheless, the lesson is that we should expect to find false positives in our top hits along with true positives. Finally, our embedding approach does not take into account the potential of drug-drug interactions to increase or decrease efficacy in any fashion. All of this is to say that further in vitro and in vivo experimentation, and observational EHR or claims data would all be useful additional sources of evidence for or against repurposing candidates listed here.

## 6 Conclusion

We present a word embedding mining approach to identifying candidate treatments for SARS-CoV-2 and COVID-19. We first use common drug-disease pairs to produce treatment analogy vectors for SARS using a prebuilt biomedical word embedding. We then use a simple word vector averaging approach to get word vectors for a list of FDA approved drugs and sort them by their distance to our treatment analogy vector. Finally, we manually evaluate the top candidate drugs and find several positive controls that have been suggested in the literature or are currently under investigation for SARS-CoV-2 or COVID-19 treatment. While there are certain to be several false positives amongst our top hits as well, we find the presence of positive controls reassuring, and propose the remainder as potential candidates for further investigation. We furthermore propose this word vector embedding approach in general as a useful tool for COVID-19 drug repurposing. These results only scratch the surface of what is possible and we present this work as a suggestion to the community to investigate further. Immediate avenues for future investigation include exploring even more drug-disease analogy vectors, ranking drugs directly by their cosine similarity to proven treatments as they arise, and investigating drug-gene target analogy vectors rather than the disease treatment analogy we demonstrate here.

## References


[1] World Health Organization. WHO director-general's opening remarks at the media briefing on COVID-19. *Geneva, Switzerland*, March 11 2020.

[2] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683, 2004.

[3] Aiping Wu, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu, Jing Meng, Zhaozhong Zhu, Zheng Zhang, Jiangyuan Wang, et al. Genome composition and divergence of the novel coronavirus (2019-ncov) originating in china. *Cell host & Microbe*, 2020.

[4] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):1–9, 2019.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, page 1, 2020.

[7] Federal Drug Administration. Drugs@FDA Data Files. `https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files`, March 2020.

[8] Donald C Hall Jr and Hai-Feng Ji. A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. *Travel Medicine and Infectious Disease*, page 101646, 2020.

[9] HonorHealth Research Institute. Atovaquone and azithromycin combination for confirmed COVID-19 infection, April 2020. `https://clinicaltrials.gov/ct2/show/study/NCT04339426`.

[10] Konrad Rejdak and Paweł Grieb. Adamantanes might be protective from covid-19 in patients with neurological diseases: multiple sclerosis, parkinsonism and cognitive impairment. *Multiple Sclerosis and Related Disorders*, page 102163, 2020.

[11] Nevio Cimolai. Potentially repurposing adamantanes for covid-19. *Journal of Medical Virology*, 92(6):531–532, 2020.

[12] Bin Cao, Yeming Wang, Danning Wen, Wen Liu, Jingli Wang, Guohui Fan, Lianguo Ruan, Bin Song, Yanping Cai, Ming Wei, et al. A trial of lopinavir–ritonavir in adults hospitalized with severe COVID-19. *New England Journal of Medicine*, 2020.

[13] Jimin Xu, Pei-Yong Shi, Hongmin Li, and Jia Zhou. Broad spectrum antiviral agent niclosamide and its therapeutic potential. *ACS infectious diseases*, 2020.

[14] University of Cincinnati. Sirolimus treatment in hospitalized patients with COVID-19 pneumonia (SCOPE), April 2020. `https://clinicaltrials.gov/ct2/show/study/NCT04341675`.

[15] Xin Liu and Xiu-Jie Wang. Potential inhibitors against 2019-ncov coronavirus m protease from clinically approved medicines. *Journal of Genetics and Genomics*, 2020.

[16] Karyopharm Therapeutics Inc. Coronavirus disease 2019 treatment: a review of early and emerging options, April 2020. `https://clinicaltrials.gov/ct2/show/study/NCT04349098`.

[17] Montreal Heart Institute. Colchicine coronavirus SARS-CoV2 trial (COLCORONA), March 2020. `https://clinicaltrials.gov/ct2/show/study/NCT04322682`.

[18] Erin K McCreary and Jason M Pogue. Coronavirus disease 2019 treatment: a review of early and emerging options. In *Open Forum Infectious Diseases*. Oxford University Press US, 2020.

[19] Stuart Weston, Christopher M. Coleman, Jeanne M. Sisk, Rob Haupt, James Logue, Krystal Matthews, and Matthew Frieman. Broad anti-coronaviral activity of fda approved drugs against SARS-CoV-2 in vitro and SARS-CoV in vivo. *bioRxiv*, 2020.

[20] Manli Wang, Ruiyuan Cao, Leike Zhang, Xinglou Yang, Jia Liu, Mingyue Xu, Zhengli Shi, Zhihong Hu, Wu Zhong, and Gengfu Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell research*, 30(3):269–271, 2020.

[21] Ayman Farag, Ping Wang, Mahmoud Ahmed, and Hesham Sadek. Identification of FDA Approved Drugs Targeting COVID-19 Virus by Structure-Based Drug Repositioning. *ChemRxiv*, 4 2020.

[22] Zhijian Xu, Cheng Peng, Yulong Shi, Zhengdan Zhu, Kaijie Mu, Xiaoyu Wang, and Weiliang Zhu. Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *bioRxiv*, 2020.

[23] Jahan S Khalili, Hai Zhu, Amanda Mak, Yongqi Yan, and Yi Zhu. Novel coronavirus treatment with ribavirin: Groundwork for evaluation concerning covid-19. *Journal of medical virology*, 2020.

[24] Jean-Jacques Mourad and Bernard I Levy. Interaction between raas inhibitors and ace2 in the context of covid-19. *Nature Reviews Cardiology*, pages 1–1, 2020.

[25] Cynthia Liu, Qiongqiong Zhou, Yingzhu Li, Linda V Garner, Steve P Watkins, Linda J Carter, Jeffrey Smoot, Anne C Gregg, Angela D Daniels, Susan Jervey, et al. Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases, 2020.

[26] Washington University School of Medicine. Double-blind, placebo-controlled clinical trial of fluvoxamine for symptomatic individuals with COVID-19 infection, April 2020. `https://clinicaltrials.gov/ct2/show/study/NCT04342663`.

[27] Leike Zhang, Yuan Sun, Hao-Long Zeng, Yudong Peng, Xiaming Jiang, Wei-Juan Shang, Yan Wu, Shufen Li, Yu-Lan Zhang, Liu Yang, Hongbo Chen, Runming Jin, Wei Liu, Hao Li, Ke Peng, and Gengfu Xiao. Calcium channel blocker amlodipine besylate is associated with reduced case fatality rate of COVID-19 patients with hypertension. *medRxiv*, 2020.

[28] Heena Garg. Can glycopyrrolate come to the airway rescue in COVID-19 patients? *Journal of Clinical Anesthesia*, 2020.

[29] DH Goetz, Y Choe, E Hansell, YT Chen, M McDowell, CB Jonsson, WR Roush, J McKerrow, and CS Craik. Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the SARS coronavirus. *Biochemistry*, 46(30):8744–8752, 2007.

[30] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved $\alpha$-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[31] Marta R Romero, Maria A Serrano, Thomas Efferth, Marcelino Alvarez, and Jose JG Marin. Effect of cantharidin, cephalotaxine and homoharringtonine on in vitro models of hepatitis b virus (HBV) and bovine viral diarrhoea virus (BVDV) replication. *Planta medica*, 73(06):552–558, 2007.

[32] TM Speight and GS Avery. Pancuronium bromide: A review of its pharmacological properties and clinical application. *Drugs*, 4(3-4):163–226, 1972.

[33] AA Buniatian, MA Vyzhigina, VM Mizikov, IuV Deshko, VA Kozhevnikov, SG Zhukova, and ShS Batchaev. New russian myorelaxant vero-pipecuronium (pipecuronium bromide) used for the anesthetic management of operations on the thorax and abdomen organs. *Anesteziologiia i reanimatologiia*, (5):49–52, 2004.

[34] Milan Jokanovic and Milica Prostran. Pyridinium oximes as cholinesterase reactivators. structure-activity relationship and efficacy in the treatment of poisoning with organophosphorus compounds. *Current medicinal chemistry*, 16(17):2177–2188, 2009.

[35] Kaiser Famil. Guidelines for the use of antiretroviral agents in hiv-1-infected adults and adolescents, 2006.

[36] Ting-Ting Shi, Xiao-Xu Yu, Li-Jun Yan, and Hong-Tao Xiao. Research progress of hydroxychloroquine and autophagy inhibitors on cancer. *Cancer chemotherapy and pharmacology*, 79(2):287–294, 2017.

[37] Dong-Bo Zhang, Jian-Jun Chen, Qiu-Yan Song, Li Zhang, and Kun Gao. Lycodine-type alkaloids from lycopodiastrum casuarinoides and their acetylcholinesterase inhibitory activity. *Molecules*, 19(7):9999–10010, 2014.