RESEARCH ARTICLE

Word embedding mining for SARS-CoV-2 and COVID-19 drug repurposing [version 1; peer review: 2 approved with reservations]

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

[1]Morgridge Institute for Research, Madison, WI, 53715, USA
[2]Duke University, Durham, NC, 27708, USA

v1 **First published:** 10 Jun 2020, **9**:585
https://doi.org/10.12688/f1000research.24271.1

Latest published: 10 Jun 2020, **9**:585
https://doi.org/10.12688/f1000research.24271.1

Abstract

Background: 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.

Methods: Due to the difficulty of accessing electronic health record 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.

Results: We first validate that our method correctly identifies ground truth treatments for well-known diseases. We then use our method to find several approved 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.

Conclusions: We 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.

Keywords
Word embedding, drug repurposing, SARS-CoV-2, COVID-19

Open Peer Review

Approval Status ? ?

	1	2
version 1 10 Jun 2020	? view	? view

1. **Quentin Vanhaelen**, Insilico Medicine, Hong Kong, Hong Kong
2. **Nansu Zong**, Mayo Clinic, Rochester, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Emerging Diseases and Outbreaks gateway.

This article is included in the Bioinformatics gateway.

This article is included in the Coronavirus collection.

Corresponding author: Finn Kuusisto (fkuusisto@morgridge.org)

Author roles: Kuusisto F: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Page D**: Conceptualization, Methodology, Project Administration, Writing – Review & Editing; **Stewart R**: Conceptualization, Methodology, Project Administration, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The authors acknowledge Marv Conney for a grant to Ron Stewart and Finn Kuusisto.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kuusisto F, Page D and Stewart R. **Word embedding mining for SARS-CoV-2 and COVID-19 drug repurposing [version 1; peer review: 2 approved with reservations]** F1000Research 2020, **9**:585 https://doi.org/10.12688/f1000research.24271.1

First published: 10 Jun 2020, **9**:585 https://doi.org/10.12688/f1000research.24271.1

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.

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.

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[4].

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[5] prebuilt embedding, specifically the intrinsic model. We chose BioWordVec because it is the most recent available biomedical word embedding and it has 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*[6]. 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[7], 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.

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[8], 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 Active-Ingredient 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 to enable 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 Table 1, Table 2, and Table 3). Note that we still used the same seed drug-disease pairs here (metformin-diabetes, benazeprilhypertension, 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 Table 4, Table 5, and Table 6).

Results

Here we present results for validation of our word embedding mining approach along with results from applying our approach

Table 1. 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

Table 2. 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

Table 3. 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 Allergies from each Analogy	
Metformin-Diabetes	lapatinib
	cisplatin
	fulvestrant
	bicalutamide
	docetaxel
	gefitinib
	tamoxifen citrate
	gemcitabine
	erlotinib hydrochloride
	toremifene citrate
Benazepril-Hypertension	bicalutamide
	docetaxel
	cisplatin
	gemcitabine
	exemestane
	lapatinib
	fulvestrant
	erlotinib hydrochloride
	gefitinib
	carboplatin
Albuterol-Asthma	docetaxel
	toremifene citrate
	tamoxifen citrate
	erlotinib hydrochloride
	gemcitabine hydrochloride
	cisplatin
	bicalutamide
	doxorubicin hydrochloride
	gemcitabine
	epirubicin hydrochloride

for COVID-19 drug repurposing. First, 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 Table 1, Table 2, and Table 3). 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.

Next, 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 Table 4, Table 5, and Table 6, 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 as a partial evaluation of the repurposing via positive controls, suggesting that other hits may be good candidates for further investigation. We find 22 positive control hits out of 50 for the metformin-diabetes analogy, 12 of 50 for the benazepril-hypertension analogy, and eight 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 7. Seven drugs are shared by all three analogies

Table 4. 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[9]
erdafitinib
atovaquone and proguanil hydrochloride[10]
rimantadine hydrochloride[11,12]
delavirdine mesylate
atazanavir sulfate and ritonavir[13]
cobimetinib fumarate
niclosamide[14]
lopinavir and ritonavir[13]
temsirolimus[15]
rilpivirine hydrochloride
alectinib hnydrochloride
lefamulin acetate
perphenazine and amitriptyline hydrochloride[16]
alogliptin and metformin hydrochloride
tamiflu[17]
selinexor[18]
amprenavir
ibuprofen and diphenhydramine citrate[19]
olanzapine and fluoxetine hydrochloride
probenecid and colchicine[20]
erlotinib hydrochloride
bicalutamide[21]
alomide
amantadine hydrochloride[11,12]
azelastine hydrochloride and fluticasone propionate[22]
revefenacin
imipramine pamoate
doravirine
rosiglitazone maleate and metformin
hydrochloride nefazodone hydrochloride
mefloquine hydrochloride[23,24]
abacavir sulfate and lamivudine
carisoprodol compound
triprolidine and pseudoephedrine hydrochlorides codeine
soma compound codeine
chloroquine hydrochloride[25]
saquinavir mesylate[26]
linagliptin and metformin hydrochloride[27]
nilutamide
donepezil hydrochloride and memantine hydrochloride[11,12]
nelfinavir mesylate[28]
ceritinib
virazole[29]
vorinostat
triprolidine and pseudoephedrine hydrochlorides
fulvestrant
gefitinib

Table 5. 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[17]
zanamivir[9]
gilteritinib fumarate
rimantadine hydrochloride[11,12]
benazepril hydrochloride
doravirine
galantamine hydrobromide
cetirizine hydrochloride hives
lanadelumab
aliskiren hemifumarate[30]
desloratadine
entacapone
invirase
daclatasvir dihydrochloride
indacaterol maleate
loratadine
peganone
nitazoxanide[31]
denavir
triprolidine and pseudoephedrine hydrochlorides codeine
rivastigmine
telavancin hydrochloride
donepezil hydrochloride
triprolidine and pseudoephedrine hydrochlorides
tazemetostat hydrobromide
relenza[9]
benazepril hydrochloride and hydrochlorothiazide
nulojix
ecallantide
alectinib hydrochloride
virazole[29]
levocetirizine hydrochloride
donepezil hydrochloride and memantine hydrochloride[11,12]
amantadine hydrochloride[11,12]
cetirizine hydrochloride
comtan
fluvoxamine maleate[32]
amlodipine besylate and benazepril hydrochloride[33]
delafloxacin meglumine
acrivastine
dalbavancin hydrochloride
fexofenadine hydrochloride hives[26]
rilpivirine hydrochloride
aricept
bendamustine hydrochloride
viramune xr
revefenacin
olodaterol hydrochloride
meloxicam

Table 6. 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[9]
rimantadine hydrochloride[11,12]
pralidoxime chloride
meperidine and atropine sulfate
amantadine hydrochloride[11,12]
doxacurium chloride
biperiden lactate
atropine sulfate syringe
gallamine triethiodide
atropine and demerol
colistin sulfate
oseltamivir phosphate[17]
revefenacin
dextromethorphan hydrobromide and quinidine sulfate
conivaptan hydrochloride
glycopyrronium tosylate
cefiderocol sulfate tosylate
fentanyl citrate and droperidol
pancuronium bromide
relenza[9]
telavancin hydrochloride
guaifenesin and dextromethorphan hydrobromide
diphenoxylate hydrochloride and atropine sulfate
esketamine hydrochloride[34]
galantamine hydrobromide
naloxone hydrochloride and pentazocine hydrochloride
glycopyrrolate[35]
levalbuterol hydrochloride
calfactant
rilpivirine hydrochloride
pipecuronium bromide
tamiflu[17]
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

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.

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, recall that we have used our drug ranking approach with the same seed analogy vectors for three 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 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 seed drug bias. For the cancer drugs, we see that every drug is indicated for some form of cancer. All of this reassures us that our approach does, in fact, find drugs appropriate for the query disease even if the query disease has no relationship with the seed drug-disease pair.

Next, 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 22 of 50 drugs either suggested or under investigation for treatment against SARS-CoV-2 and COVID-19. With the benazepril-hypertension analogy, we find 12 of 50 hits, and from the albuterol-asthma analogy, we find eight 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, four 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[11,12]. Zanamavir is an antiviral that has been suggested based on *in silico* molecular docking models of the 3C-like proteinase[9], which is a major protease thought essential to viral replication of coronaviruses, including SARS-CoV and SARS-CoV-2[36,37]. Oseltamivir (Tamiflu) is another antiviral that is under investigation via clinical trial[17].

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 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*[29].

We also note that 13 of all the proposed treatments are in clinical trials: atovaquone, lopinavir and ritonavir, sirolimus (suggested here as the prodrug temsirolimus), oseltamivir, selinexor,

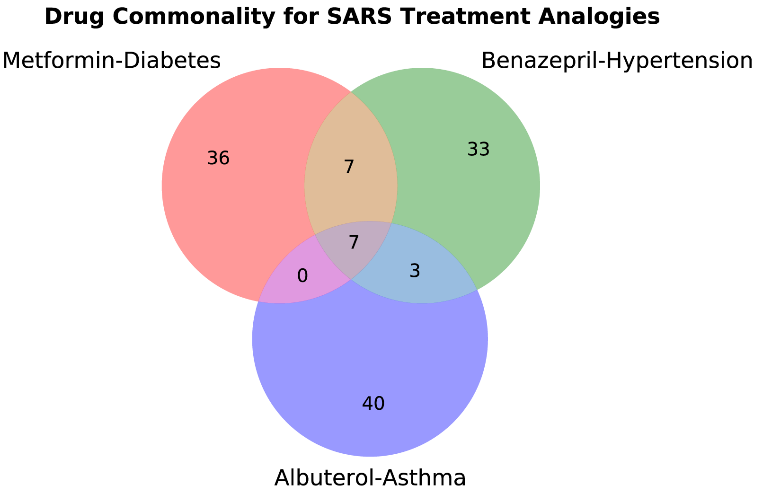

Figure 1. **Venn diagram of the top 50 drug candidates identified by each SARS treatment analogy vector.**

Table 7. **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[11,12]
	peramivir revefenacin rilpivirine hydrochloride
	rimantadine hydrochloride[11,12] tamiflu[17] zanamivir[9]
Common to two	alectinib hydrochloride
	donepezil hydrochloride and memantine hydrochloride[11,12]
	doravirine galantamine hydrobromide gilteritinib fumarate
	relenza[9]
	telavancin hydrochloride triprolidine and pseudoephedrine hydrochlorides triprolidine and pseudoephedrine hydrochlorides codeine
	virazole[29]

ibuprofen, colchicine, bicalutamide, mefloquine, chloroquine, linagliptin, fluvoxamine, and ketamine (suggested here as the enantiomer esketamine). Interestingly, these drugs come from a wide range of primary indications including antiparasitic, antiviral, anti-inflammatory, anticancer, anesthetic, 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 clinical, *in silico*, *in vitro*, and *in vivo* experimentation, all done here with literature published before SARS-CoV-2 was discovered.

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 outside of the embedding vocabulary. 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 is a neuraminidase inhibitor used to treat influenza. While it is thus an antiviral, coronaviruses do not use neuraminidase, so it would seem less likely to be effective against SARS-CoV-2[22]. On the other hand, zanamivir and oseltamivir, two of our common positive controls[9,17], are also neuraminidase inhibitors and should thus be less likely candidates. Given that the potential mechanism of action for zanamivir at least 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 any 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.

Conclusions

In this work, we present a word embedding mining approach to identifying candidate treatments for SARS-CoV-2 and COVID-19. We first use seed drug-disease pairs to produce treatment analogy vectors for a query disease using a prebuilt biomedical word embedding. We then use a simple word vector averaging approach to get vectors for a list of FDA approved drugs and sort them by their distance to our treatment analogy vectors. We validate that this approach identifies ground truth treatments for well-known diseases. Next, we use the same approach to produce a list of candidate drugs for the query disease SARS, 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.

Data availability

The FDA database of approved drugs is available at: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files.

All code and processed data used to produce these results are available at: https://github.com/finnkuusisto/covid19_word_embedding.

Archived code and data as at time of publication: http://doi.org/10.5281/zenodo.3860057[4].

License: CC0

The code is provided in Python (v 3.8) as Jupyter Notebooks (v 6.0.3), and additionally requires Gensim (v 3.8.1), Matplotlib (v 3.2.1), and Matplotlib-Venn (v 0.11.5).

Software availability

The BioWordVec prebuilt embedding is available via the official GitHub repository: https://github.com/ncbi-nlp/BioWordVec.

References

1. World Health Organization: **WHO director-general's opening remarks at the media briefing on COVID-19**. Geneva, Switzerland. 2020.
Reference Source

2. Ashburn TT, Thor KB: **Drug repositioning: identifying and developing new uses for existing drugs.** *Nat Rev Drug Discov.* 2004; **3**(8): 673–683.
PubMed Abstract | Publisher Full Text

3. Wu A, Peng Y, Huang B, *et al.*: **Genome composition and divergence of the novel coronavirus (2019-ncov) originating in china.** *Cell Host Microbe.* 2020; **27**(3): 325–328.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Kuusisto F: **finnkuusisto/covid19_word_embedding: First release for publication. (Versionv1.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3860057

5. Zhang Y, Chen Q, Yang Z, *et al.*: **BioWordVec, improving biomedical word embeddings with subword information and MeSH.** *Sci Data.* 2019; **6**(1): 52.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Mikolov T, Chen K, Corrado G, *et al.*: **Efficient estimation of word representations in vector space.** *arXiv preprint arXiv: 1301.3781.* 2013.
Reference Source

7. 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.** *Nat Microbiol.* 2020; **5**(4): 536–544.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Federal Drug Administration: **Drugs@FDA Data Files**. 2020.
Reference Source

9. Hall DC Jr, Ji HF: **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 Med Infect Dis.* 2020; 101646.
PubMed Abstract | Publisher Full Text | Free Full Text

10. HonorHealth Research Institute: **Atovaquone and azithromycin combination for confirmed COVID-19 infection**. 2020.
Reference Source

11. Rejdak K, Grieb P: **Adamantanes might be protective from covid-19 in patients with neurological diseases: multiple sclerosis, parkinsonism and cognitive impairment.** *Mult Scler Relat Disord.* 2020; **42**: 102163.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Cimolai N: **Potentially repurposing adamantanes for covid-19.** *J Med Virol.* 2020; **92**(6): 531–532.
PubMed Abstract | Publisher Full Text

13. Cao B, Wang Y, Wen D, *et al.*: **A trial of lopinavir-ritonavir in adults hospitalized with severe COVID-19.** *N Engl J Med.* 2020; **382**(19): 1787–1799.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Xu J, Shi PY, Li H, *et al.*: **Broad spectrum antiviral agent niclosamide and its therapeutic potential.** *ACS Infect Dis.* 2020; **6**(5): 909–915.
PubMed Abstract | Publisher Full Text | Free Full Text

15. University of Cincinnati: **Sirolimus treatment in hospitalized patients with COVID-19 pneumonia (SCOPE)**. 2020.
Reference Source

16. Liu X, Wang XJ: **Potential inhibitors against 2019-ncov coronavirus M protease from clinically approved medicines.** *J Genet Genomics.* 2020; **47**(2): 119–121.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Tongji Hospital: **A prospective/retrospective,randomized controlled clinical study of antiviral therapy in the 2019-nCoV pneumonia**. 2020.
Reference Source

18. Karyopharm Therapeutics Inc.: **Coronavirus disease 2019 treatment: a review of early and emerging options**. 2020.
Reference Source

19. King's College London: **LIBERATE Trial in COVID-19 (LIBERATE)**. 2020.
Reference Source

20. Montreal Heart Institute: **Colchicine coronavirus SARS-CoV2 trial (COLCORONA)**. 2020.
Reference Source

21. Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins: **Trial to promote recovery from COVID-19 with ivermectin or endocrine therapy (RECOVER)**. 2020.
Reference Source

22. McCreary EK, Pogue JM: **Coronavirus disease 2019 treatment: a review of early and emerging options**. In: *Open Forum Infectious Diseases.* Oxford University Press US, 2020; **7**(4): ofaa105.
Publisher Full Text

23. Weston S, Coleman CM, Sisk JM, *et al.*: **Broad anti-coronaviral activity of FDA approved drugs against SARS-CoV-2 *in vitro* and SARS-CoV *in vivo*.** *bioRxiv.* 2020.
Publisher Full Text

24. Burnasyan Federal Medical Biophysical Center: **An open randomized study of the effectiveness of mefloquine for the treatment of patients with COVID19**. 2020.
Reference Source

25. Wang M, Cao R, Zhang L, *et al.*: **Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*.** *Cell Res.* 2020; **30**(3): 269–271.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Farag A, Wang P, Ahmed M, *et al.*: **Identification of FDA Approved Drugs Targeting COVID-19 Virus by Structure-Based Drug Repositioning.** *ChemRxiv.* 2020; **4**.
Publisher Full Text

27. University of Miami: **Effects of DPP4 inhibition on COVID-19**. 2020.
Reference Source

28. Xu Z, Peng C, Shi Y, *et al.*: **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.
Publisher Full Text

29. Khalili JS, Zhu H, Mak A, *et al.*: **Novel coronavirus treatment with ribavirin: Groundwork for evaluation concerning covid-19.** *J Med Virol.* 2020.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Mourad JJ, Levy BI: **Interaction between raas inhibitors and ace2 in the context of covid-19.** *Nat Rev Cardiol.* 2020; **17**(5): 313.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Liu C, Zhou Q, Li Y, *et al.*: **Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases.** *ACS Cent Sci.* 2020; **6**(3): 315–331.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Washington University School of Medicine: **Doubleblind, placebo-controlled clinical trial of fluvoxamine for symptomatic individuals with COVID-19 infection.** 2020.
Reference Source

33. Zhang L, Sun Y, Zeng HL: **Calcium channel blocker amlodipine besylate is associated with reduced case fatal- ity rate of COVID-19 patients with hypertension.** *medRxiv.* 2020.
Publisher Full Text

34. William Beaumont Hospitals: **Study of immunomodulation using naltrexone and ketamine for COVID-19 (SINK COVID-19).** 2020.
Reference Source

35. Garg H: **Can glycopyrrolate come to the airway rescue in COVID-19 patients?** *J Clin Anesth.* 2020; **64**: 109843.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Goetz DH, Choe Y, Hansell E, *et al.*: **Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the SARS coron- avirus.** *Biochemistry.* 2007; **46**(30): 8744–8752.
PubMed Abstract | Publisher Full Text

37. Zhang L, Lin D, Sun X, *et al.*: **Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors.** *Science.* 2020; **368**(6489): 409–412.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 15 December 2020

https://doi.org/10.5256/f1000research.26777.r75542

?

Nansu Zong
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

The authors proposed a novel method for drug repurposing to mine a large biomedical word embedding for FDA approved drugs based on drug-disease treatment analogies. The paper is well presented, and the results seem promising. The writing is clear and easy to understand with the help of abundant tables and figures. While the contribution described in the manuscript is worthwhile, there are some limitations needed to be improved:

1. The biggest question is why the three drug pairs are selected? Any metrics used to select those pairs? The authors have other candidate pairs, and what are those results?

2. Evaluation metrics are missing, e.g., AUCROC, Precision, Recall.

3. Comparison to the state-of-the-art methods is missing

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Deep Learning, computational drug repurposing, translational bioinformatics, EHR

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 30 June 2020

https://doi.org/10.5256/f1000research.26777.r64552

? **Quentin Vanhaelen**
Insilico Medicine, Hong Kong, Hong Kong

The authors propose a new computational drug repurposing method based on word embedding for FDA approved drugs based on drug-disease treatment analogies. Acknowledging that the onset of the COVID-19 outbreak requires the quick identification of already known drugs which could be repurposed against COVID-19, the authors propose a relatively straightforward method where candidate drug names are converted into word vectors to enable ranking by their similarity with a treatment analogy vector. The paper is well written and organized. The authors have done a good job in describing the rationale behind the development of their method and give an overview of its design process as well as the testing steps. Nevertheless, several points should be addressed regarding the validation of the method and its performance, benchmarking with respect to other similar methods, and applicability of the method for other diseases.
The field of computational repurposing methods is very large and in contrast, the introduction of this article is relatively short and does not contain any comprehensive referenced overview of prior works in this field. My recommendation is that the authors should at least shortly cover the existing methods based on word mining and semantic and discuss how their own method positions itself in this context. Discussing those methods could serve as a starting point for a proper assessment of the performance of the proposed algorithm (see comment below).

Below I summarize my comments and suggestions which I hope could contribute to improving the manuscript.

1. The proposed algorithm is based on drug-disease relationships. It might be interesting to take into account the drug-target relationship as well. Indeed, one drug can often interact with several targets and usually, the drug-disease association is established through one specific target. Including the drug- target and disease-target relationships offer additional possibilities to repurpose the drug against a condition for which the drug-target relationship is relevant.

2. The algorithm mainly uses association by keywords. As the authors has noted, this approach has several limitations. It would be difficult to apply the algorithm for finding repurposing candidates for conditions without prior known drugs (This restriction being encountered by many repurposing algorithms). Furthermore, the method might be limited by the vocabulary used. Indeed, it is not uncommon for a condition or a drug to be known by different names in the scientific literature. This issue can be addressed when dealing with formatted databases but it is more difficult to handle with natural language in the literature.

3. By its nature, this method does not take into account all the relevant properties at the genomic, transcriptomic, chemical, structural levels which are of paramount importance to correctly establish the compatibility of a drug-target-disease relationship. For instance, the authors have based their study on the assumption that SARS-CoV and SARS-CoV-2 share common features, which is true but it was also pointed out that some structural and non-structural proteins of SARS-CoV-2 have low sequence similarities with the corresponding ones from SARS-CoV and this may have strong implication on the use of SARS-CoV drug to treat SARS-CoV-2 associated disease.

4. This method is in my opinion would be interesting as an initial approach to reduce the number of potential repurposing candidates which should be included as input data for a subsequent computational repurposing method capable of taking into account the information about chemical/structural and genomic features of the target- drug-disease relationships. As many methods can be relatively costly from a computational perspective, being able to pre-select the initial set of data to be integrated is of interest.

5. It would be interesting to have a more detailed description of how the treatment analogy vectors are built and how the similarities are calculated. The authors have initially a list of 8,561 candidate drugs (after some manual curation) but end up using a reduced list of 5,833 distinct drug vectors. This is a significant reduction in the initial amount of information the algorithm can be applied to. Maybe the authors could suggest alternative methods to be able to keep a higher proportion of the initial list of drugs? Also the table could include the similarity scores of the drugs. The authors decided to list the top 50 drugs, they do not provide a reason why 50 should be chosen as a cut off. Is there any systematic way to define an appropriate cut off for the most relevant results? This should be discussed in more details in the paper.

6. As the assembly of the list of drug vectors requires several steps of preparation, it would be interesting for the readers to have those steps summarized in a diagram. This diagram could describe the key steps for the preparation of the drug vectors from a general perspective (so not only for the SARS-CoV-2 case), with an emphasis on what the user should take care of when deciding to include or not a drug. The article already contains many tables summarising the results of the different experiments. Maybe some of them could be moved in the supplementary materials to give space to this diagram.

7. The main criteria in this paper to assess the method is to look for predicted drugs that are either already known as being effective against SARS-CoV or already taken into clinical trials for SARS-CoV-2. But those criteria are restrictive and not necessarily generalizable to other

type of disease. One can assume that many potential repurposing candidates for a given condition are not undergoing clinical trial for a similar condition. It should be also noted that having a drug entering clinical trial is not a guarantee of success considering the relatively high failure rate in this area. It is true that for assessing the results provided by this type of computational method, a literature review is a first step that needs to be followed by in vitro/in vivo validation of the most promising candidates. Before those more expensive validations could take place, what would be the computational or data-based criteria to assess the accuracy of the predictions? The authors could discuss how this method performs compared to other methods based on word embedding and text mining for instance.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology, system biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 01 Sep 2020

Finn Kuusisto, Morgridge Institute for Research, USA

Dr. Vanhaelen,
We greatly appreciate your thoughtful response on this work and apologize for the long delay. This is all really helpful advice. We do not intend for this to be a formal and complete response to your review, but while we wait for a second reviewer, and before resubmitting, we hope to understand if our proposals here might be sufficient to address your concerns.

The overarching themes seem to be centered around providing background and context, clarity of methods and evaluation, and overall messaging on the goal of this work. Briefly, rather than being an automated method for validating potential drug candidates, we see this work more as a helpful starting point for researchers to begin literature search. In this way, we hope to accelerate the discovery process by helping researchers prioritize what are the most promising candidates for more in-depth in vitro and in vivo analysis.

Regarding a lack of background, that was an oversight. In the resubmission we will add a background section to review methods in computational drug repurposing, including some discussion of drug-drug similarity, matching genetic signatures of phenotypes to drugs, in silico measures of molecular docking, mining electronic health record data for drug side effects, and literature-based discovery. We hope this provides better context for our work.

We also have plans for clarifying the methods and intent of the work. For improvement of the evaluation, we intend to include some comparison to the BEST text mining tool and likely Polysearch2 as well. These two are the most similar approaches we could find that would allow for collecting a large number of proposed candidate drugs from the literature quickly.

-Specific Points-
1. Drug-target relationships
We really like this idea and alluded to something similar in the future work section. It may also help alleviate the limitation of only finding drugs that commonly exist in the literature. We are pursuing the idea, but given our initial promising results just using drug names, we decided that angle was out of scope and might even confuse the message of what we have accomplished so far. We will add to our discussion of drug-target relationships as a promising area for future work.

2. Keywords/vocabulary
You are absolutely correct that token matching approaches have major limitations in what they can find. They simply cannot directly find something that is not present in the vocabulary. This is the reason for the drug count reducing from 8,561 to 5,833 as explained below. Complicated terms and those with many synonyms are another limitation that would probably be best addressed at the word embedding training phase. Furthermore, some of our other work on more traditional text mining has shown synonyms to be particularly challenging in that they can lead to increases of false positives, and confusion when collisions occur (e.g. drug A and drug B have the same synonym X). It will take more effort to address properly.

We tried to limit this work to readily available tools for the sake of expedience and have found that, despite these limitations, our method presents interesting findings. We plan to add your concerns to the limitations/future work discussion.

3. Relevant genomic/viral/drug properties not considered
It is true that our current approach misses several relevant properties related to the drug-target-disease relationship. Some of that may be alleviated by future work investigating word embedding mining at the drug-gene/protein/pathway level, but pursuing those lines

of inquiry here would greatly lengthen and complicate this paper. We are aiming here to show simply that analogy vectors are likely useful. We will explain that further analogy-based analysis will be useful for establishing links between drugs and targets and targets to disease which will further filter candidates prior to expensive wet lab validation. We see this method as a potentially helpful tool for quickly prioritizing drug candidates for further investigation. Rather than providing automated in-depth support for candidate drugs, we would expect researchers to use our method to conduct initial literature search, thereby accelerating the process of choosing promising candidates for more in-depth in vitro and in vivo testing. We will attempt to clarify this intent in the resubmission.

4. Preselection comment
We wholeheartedly agree that our work is best viewed as an initial preselection step. Our intent is for this method to serve as an early step, providing suggestions for further investigation. As you have pointed out, this approach ignores several relevant properties that define the drug-target-disease relationship. More than anything, we see this work as providing a quick and powerful summarization of leads within the literature and helping to prioritize research in promising directions. We will clarify this message in our resubmission.

5. Vector creation and cosine similarity cutoff
You are correct that we do lose a large portion of our initial drug set (8.5k down to 5.8k) when converting them into drug vectors. We will be sure to clarify why this is in our resubmission, but I can briefly explain here. The process of converting from drug names to drug vectors is automated rather than manual and consists of 1) splitting drug names into individual tokens by whitespace, 2) getting word vectors from the embedding for those tokens, and 3) averaging those constituent token word vectors for each drug. When an individual token is not present in the BioWordVec embedding vocabulary, we drop it from the drug name. If none of the tokens from a drug name exist in the BioWordVec vocabulary, we drop the drug entirely. Ultimately, what that means is that there are ~3k drugs in our initial set that simply have no tokens present in the embedding vocabulary and cannot be salvaged. This speaks to some of the limitations previously discussed. We will clarify this in the resubmission.
For the cosine-similarity cutoff versus top 50 hits, we are unaware of any particularly principled methods for picking a similarity cutoff. We considered inspecting the cosine-similarity distributions for change points, and they do follow something like an inverted logit, but we found that the change points selected hit counts on the order of hundreds of drugs. Ultimately, we chose to simply pick what seemed a manageable number of drugs to review by hand. We can clarify in the text how this is somewhat an arbitrary decision.

6. Visual diagram
A generalized visual diagram of the whole pipeline would be helpful. We will add one to the resubmission.

7. Evaluation process
Evaluation is certainly tough in drug repurposing. As you say, comparing to drugs suggested in other publications and ongoing clinical trials may be quite restrictive and not particularly generalizable to other diseases. At the same time, it may even be too loose of a restriction in this case as the sheer volume of publications on COVID-19 at this point may

suggest a vast number of false positive drugs. Again, evaluation in drug repurposing is tough and eventually necessitates experimentation in vitro and in vivo. Thus, we likely still won't know which hits were true positives for quite some time. That is the main reason we first demonstrated that the treatment analogy vectors could identify gold-standard drugs for well-known diseases in tables one through three. Ultimately though, given our goal of augmenting the initial search process, we simply want to determine if our method sorts drug candidates in a reasonable fashion for this new disease. If so, we can be more confident that other not-yet-considered drugs on the list, and those just a bit farther down the list, may be promising candidates for researchers to consider. We will elaborate on the limitations of this evaluation further.

As for other comparable approaches, we plan to include the top drug hits provided by the BEST and Polysearch2 text mining tools. BEST allows perhaps the closest comparison because it allows for limiting the search to literature through 2019, as ours is based on a word embedding built in 2019. Both provide many positive hits according to our manual evaluation process, but we argue that our approach still offers some advantages for two reasons, both related to flexibility. First, our approach here allows for the input of any list of drug names or query entities, whereas BEST and Polysearch2 are limited to prebuilt term lists. Those prebuilt lists can almost certainly be expanded, but our approach is flexible right out of the box. Second, because our method is based on treatment analogy vectors, it can provide a greater number and greater diversity of possible candidates. BEST and Polysearch2 provide a single ranking of drugs to disease based on literature co-occurrence, whereas our approach can be reseeded with a different treatment analogy vector (e.g. metformin/diabetes vs albuterol/asthma) and immediately provide a related but different ranking of drugs. We see this as an interesting advantage to explore in future work.

Overall, we mainly hope to demonstrate that our very quick and simple method could be useful in the early stages of research, rather than demonstrating that our method is better than current state of the art methods for deeper evaluation of drug effects. Bear in mind again that these sorted drug suggestions are essentially time-censored to literature from 2019. We were pleasantly surprised to see the quality of suggestions our method provided and how early it may have been used to quickly gather promising candidates that are now under investigation.

Competing Interests: No competing interests were disclosed.