PATTER, Volume *1*

Supplemental Information

Machine Learning Maps Research Needs in COVID-19 Literature

Anhvinh Doanvo, Xiaolu Qian, Divya Ramjee, Helen Piontkivska, Angel Desai, and Maimuna Majumder

Supplemental Information 1. Singular Value Decomposition Enables The Use of Sparse Matrices and Randomized Algorithm Dramatically Speeds Computation

Principal components analysis (PCA) is typically completed in three key steps:

1. We center the data. In other words, we subtract the mean of each column from the original data, yielding a matrix where the mean of each column is zero.
2. We calculate the covariance matrix of this centered data. This represents all of the correlations between every column of data.
3. We perform an eigendecomposition of this covariance matrix. This yields what we consider to be the final products of dimensionality reduction: the principal components (eigenvectors) that represent key patterns in the data, along with information on how important they are (eigenvalues, which represent how much variance they capture).

The second and third steps of this process primarily rely on matrix multiplication which has been highly optimized in most scientific computing packages, including Python's scipy and numpy, as well as sklearn's implementation of covariance-based PCA. But the first step - centering the data - relies on matrix subtraction, which has not had nearly as much technical development aimed at its optimization. This process thus typically requires a dense matrix, where every value, even if they are zero, is explicitly delineated.

However, this is not computationally feasible with our DTMs, where we had tens of thousands of rows and columns, resulting in billions of elements that we could not store entirely in memory. We instead stored DTMs as sparse matrices, which are efficient because most elements are zero and only the values of nonzero elements are stored, but are incompatible column-wise addition and subtraction operations. Therefore, we calculated PCA instead by performing the singular value decomposition (SVD) on the original sparse and uncentered DTM, which is possible with the sklearn Python package (Supplemental Information 8). While the SVD operation is equivalent to PCA when SVD is performed on centered data, it is worth noting that when data is uncentered, the first principal component (PC) outputted may capture *less* variance than the second PC because the first PC captures the mean of the data.[1]

The complexity of PCA through SVD scales with an order of $O(max(m,n) * min(m,n)^2)$, or $O(m * n^2)$ when m is large, and where m and n are the number of observations (documents) and features (unique words) respectively. With nearly 10,000 articles mentioning coronavirus-related terms in their abstracts and tens of thousands of unique words, SVD computations can take some time. We accelerated this step by using randomized SVD, which has an order of complexity of just $O(mn\, log(k))$, where k is the number of PCs computed. Indeed, this enabled our SVD calculations to proceed almost instantaneously. And while there is some randomness associated with results from randomized SVD, existing literature indicates that its output converges super-exponentially to the true output of SVD with additional iterations.[2]

Supplemental Information 2. Distribution of the Top 50 Key Terms Separating COVID-19 Abstracts from non-COVID-19 Abstracts along Principal Component 2 (shown in Figure 3)

Lemmatized Word	**Component Value**	**Percentage of COVID-19 Abstracts**	**Percentage of non-COVID-19 Abstracts**
patient	-0.20184	51.3	26.7
covid	-0.19725	99.8	6.1
case	-0.13793	34.9	20.1
hospit	-0.09987	18.9	10.8
pandem	-0.08412	44.4	9.7
risk	-0.07939	21.2	9.1
care	-0.0786	16.6	5.8
epidem	-0.07234	16.1	10.4
countri	-0.07022	16.7	7.6
sever	-0.06779	22.9	12.8
manag	-0.06461	15.5	5.6
estim	-0.06186	8.9	4.5
death	-0.06158	13.8	7
number	-0.06069	18.8	10.8
function	0.060604	4.8	10.1
specif	0.06181	12.8	18.9
mice	0.062183	0.4	4.5
tgev	0.06324	0	2.6
amino_acid	0.063691	0.6	5.1
inhibit	0.063901	2	6.6
activ	0.064333	10.1	14.2

coronavirus	0.065021	43.3	67.1
assay	0.069902	2.3	10
induc	0.070871	2.7	9.1
host	0.071237	2.7	8.2
target	0.072913	5.9	10.5
receptor	0.07325	2.7	5.4
interact	0.07327	3.9	7.3
domain	0.073451	1.6	6
antigen	0.075788	1	8.4
hcov	0.076117	0.4	3.5
spike_protein	0.078466	1.5	6.1
recombin	0.079413	0.6	6
replic	0.082887	1.5	8.5
strain	0.085891	2.5	11.3
genom	0.094031	2.2	10.3
vaccin	0.097755	6.2	10.5
bind	0.099378	2.3	7.3
antibodi	0.100884	2.9	10.5
structur	0.103286	4.3	11.2
sequenc	0.104208	2.4	13
gene	0.108125	1.6	10.8
viral	0.10938	12.2	27.9
human	0.119375	11.5	24.6
express	0.119808	3.2	11.3
mer	0.134674	3.6	10.6

virus	0.164846	22.2	51.8
sar	0.165963	37.4	41.7
cell	0.228506	6.3	21
protein	0.378095	4.1	21.3

Supplemental Information 3. Examples of Abstracts Identified as Either COVID-19 or non-COVID related

Relationship with COVID-19	**Abstract Title**
Related to COVID-19 (Bottom 1% of SecondPrincipal Component)	Recommendations for standardized management of CML patients in the core epidemic area of COVID-19[3]
	Transmission risk of patients with COVID-19 meeting discharge criteria should be interpreted with caution[4]
	COVID-19 in a Designated Infectious Diseases Hospital Outside Hubei Province, China[5]
Unrelated to COVID-19 (Top 1% of Second Principal Component)	Characterization of the expression and immunogenicity of the ns4b protein of human coronavirus 229E[6]
	Severe acute respiratory syndrome coronavirus nucleocapsid protein expressed by an adenovirus vector is phosphorylated and immunogenic in mice[7]
	Molecular cloning and expression of a spike protein of neurovirulent murine coronavirus JHMV variant cl-2[8]

Supplemental Information 4. August 2020 Analysis Update

In our analysis of CORD-19 data up through July 31, 2020, we found results very similar to our analysis conducted on data up through May 28, 2020: COVID-19 research has continued to focus heavily on CMF-based study, much more so than laboratory-based study, especially when compared to research done on other coronaviruses.

PCA Analysis

We found that PC2 strongly differentiates between COVID-19 and non-COVID-19 coronavirus abstracts. COVID-19 abstracts tend to have lower projection values on PC2.

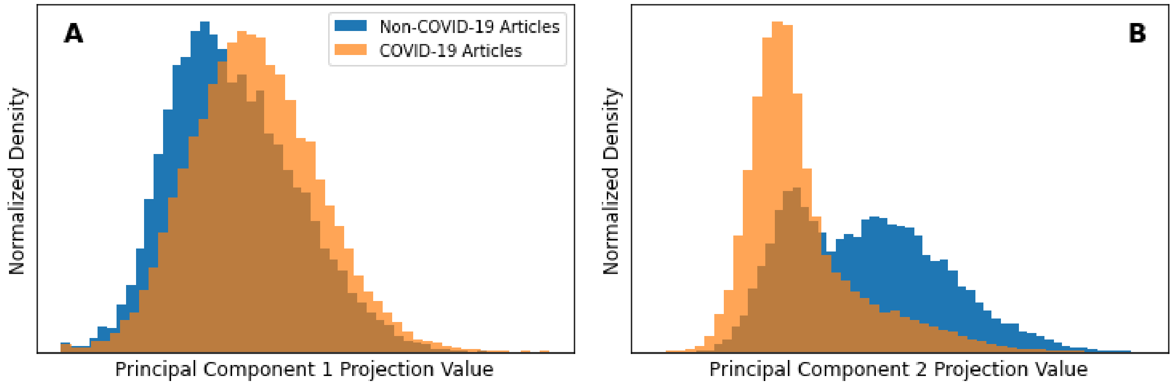


Figure SI.10.1. The distributions (y-axis) of non-COVID-19 abstracts (blue) and COVID-19 abstracts (orange) are plotted against the PC projection values (x-axis) in each panel. PC2 clearly separates the two groups.

On PC2, lower projection values are associated with CMF-related terms, such as “hospit” and “case”, while higher projection values are often associated with laboratory-based study.

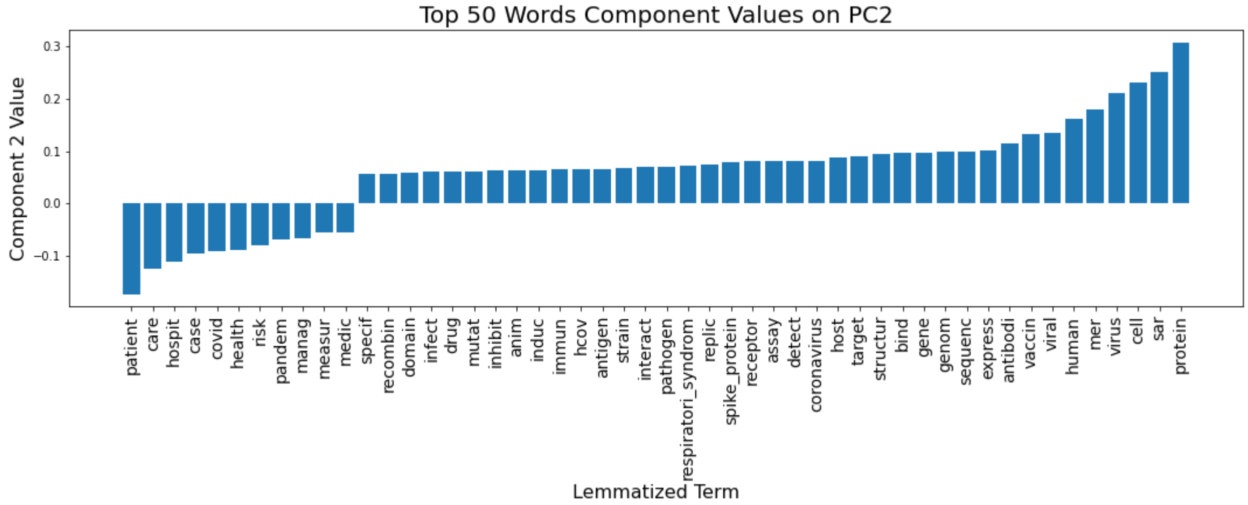


Figure SI.10.2. This bar chart displays the values of key lemmatized words on the second PC. We selected the 50 words associated with the largest component value magnitude for this plot to interpret the PC. The component values are represented by the y-axis and each individual word is plotted along the x-axis.

When compared with non-COVID-19 abstracts mentioning other coronaviruses, we observe that COVID-19 abstracts are much more likely to mention CMF-related terms and much less likely to mention terms related to laboratory-based study.

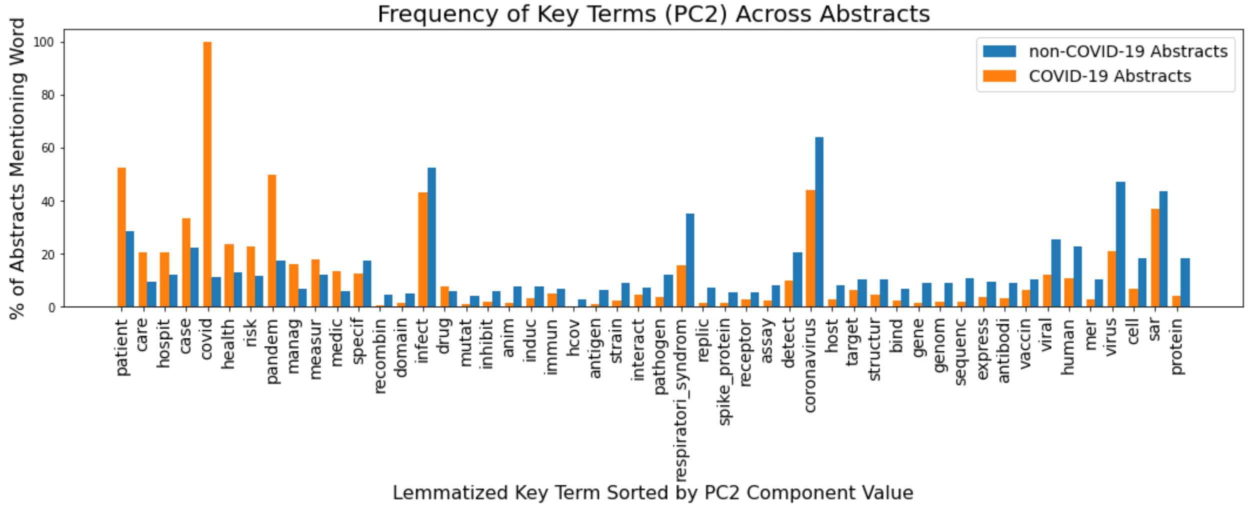


Figure SI.10.3. The top 50 key terms, selected by the magnitude of their component values on PC2, are unevenly distributed among the COVID -19 and non-COVID-19 abstracts. The proportion of abstracts in each group (orange for COVID-19 abstracts and blue for non-COVID-19 abstracts) mentioning each term is represented by the y-axis and each individual word is plotted along the x-axis.

LDA Analysis

SARS-CoV-2 abstracts focused on topics different from non-SARS-CoV-2 abstracts.

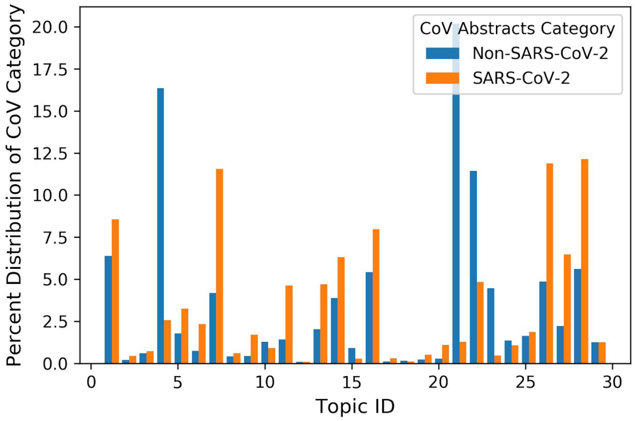


Figure SI.10.4. COVID-19 literature is distributed unevenly across the 30 topics identified via LDA. Topics identifiers (IDs) were assigned randomly by LDA. The percentage of abstracts in each of the two groups (orange for COVID-19 abstracts and blue for non-COVID-19 coronavirus abstracts) that are in each topic is represented by the y-axis, while each topic ID number is plotted against the x-axis.

In particular, we detected five major topic families, including (1) clinical issues: testing and diagnostics, (2) societies and outbreaks: responses to mitigate them and diseases' impact on society, (3) basic microbiological study, (4) general outbreak reporting, and (5) modeling of disease transmission. Basic microbiological research on SARS-CoV-2 continues to lag relative to CMF-related study.

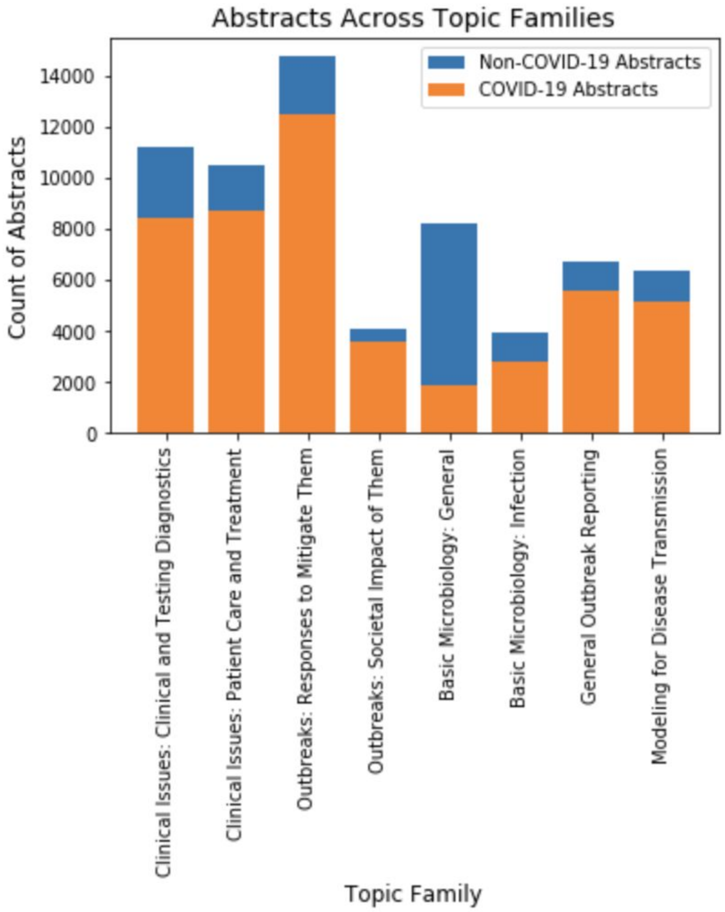
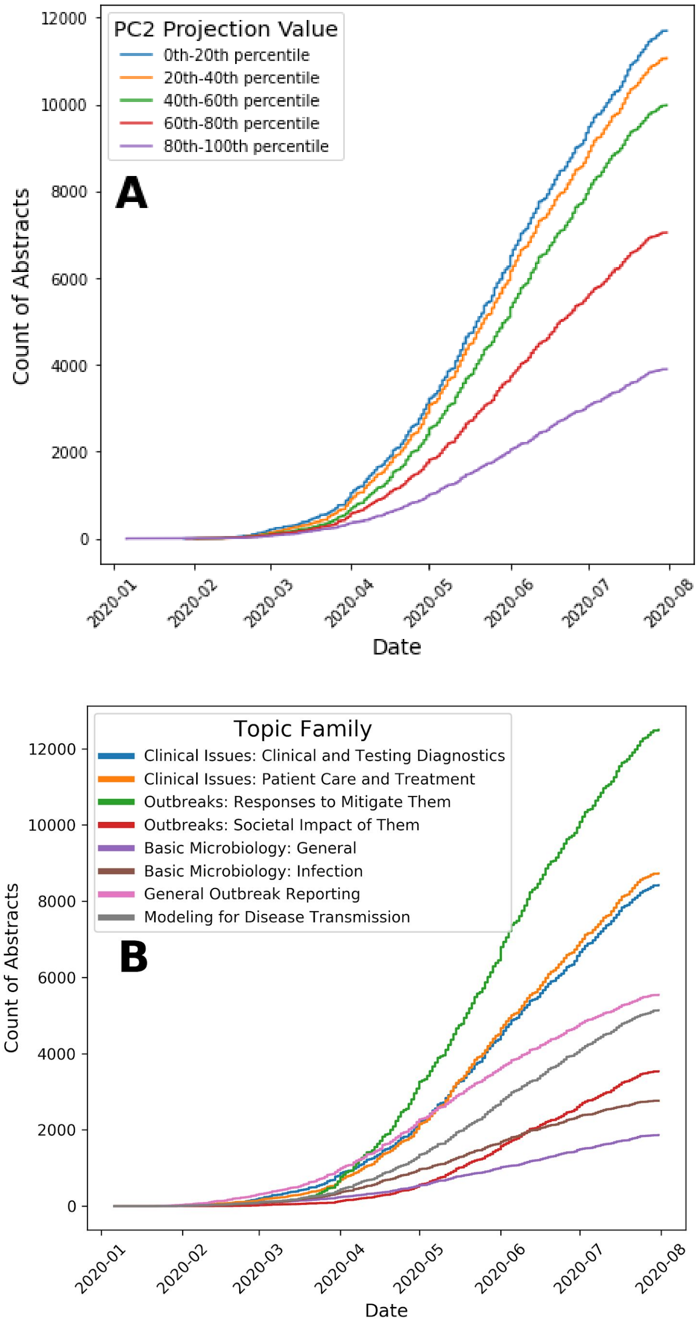


Figure SI.10.5. The distribution (y-axis) of abstracts of each category (blue for non-SARS-CoV-2 coronavirus abstracts and orange for SARS-CoV-2 abstracts) are plotted against each topic family (x-axis). Abstracts are unevenly distributed and SARS-CoV-2 abstracts tend to focus more on clinical, modeling, or field- (CMF) based study than basic microbiological research, even when compared with research on other coronaviruses.

Time Analysis

In both our PCA and LDA analysis, COVID-19 research has tended to focus more on CMF-based than laboratory-based study. However, some trends in the LDA analysis are particularly noteworthy: general outbreak reporting has slowed relative to research on clinical issues. Research on the societal impact of outbreaks and the modeling of disease transmission has also rapidly accelerated, relative to basic microbiological research. Study of public health responses to mitigate the pandemic continues to dominate the field.



Figures SI.10.6a and SI.10.6b. Panel A shows the distribution over time of COVID-19 abstracts with different projection values on the second PC (i.e, those likely reflecting CMF research versus laboratory research) and the different timelines for publication between these groups. Panel B shows COVID-19 research is predominantly focused on outbreak reporting, public health issues, and clinical issues. For both graphics, the y- and x-axes represent the count of abstracts and the date of each count respectively. Each line in both graphics is colored by the different groups of abstracts; for panel A, each group is comprised of abstracts within a certain range of projection values on PC2, and for panel B, each group is comprised of abstracts within a certain topic family.

Supplemental Information 5. Topic Families Across COVID-19 and non-COVID-19 Abstracts

ID	Topic Title	Topic Family	COVID-19 Count (percent)	Total (percent)
1	Treatment and patient care for COVID-19	Vaccine needs, patient care, and treatments	250 (1.4%)	338 (1%)
2	Biomolecular study of coronaviruses	Microbiology (general)	287 (1.6%)	3165 (9%)
3	Infection by coronavirus	Microbiology (transmission)	149 (0.8%)	311 (0.9%)
4	Porcine coronavirus microbiology and infection	Microbiology (general)	37 (0.2%)	199 (0.6%)
5	Infection by coronavirus	Microbiology (transmission)	226 (1.2%)	1695 (4.8%)
6	Public health issues of SARS transmission	Outbreaks (public health)	305 (1.7%)	1274 (3.6%)
7	Outbreaks in different countries	Outbreaks (general coverage)	231 (1.3%)	336 (1%)
8	Death and mortality due to COVID-19	Outbreaks (general coverage)	226 (1.2%)	284 (0.8%)
9	Human infection by coronaviruses	Microbiology (transmission)	430 (2.3%)	911 (2.6%)
10	Testing and infection of COVID-19	Testing (mixed with transmission)	210 (1.1%)	321 (0.9%)
11	Impact of COVID-19 on community services	Outbreaks (general coverage)	125 (0.7%)	208 (0.6%)
12	Infection by MERS-CoV	Microbiology (transmission)	36 (0.2%)	660 (1.9%)
13	General COVID-19 pandemic coverage	Outbreaks (general coverage)	432 (2.3%)	498 (1.4%)
14	COVID-19's impact on healthcare services	Outbreaks (public health)	2666 (14.5%)	3050 (8.6%)

15	Clinical testing and COVID-19 symptoms	Testing	1750 (9.5%)	2201 (6.2%)
16	Transmission and infection among coronaviruses	Microbiology (transmission)	980 (5.3%)	1903 (5.4%)
17	Modeling, statistics, and investigation of epidemic	Outbreaks (general coverage)	2006 (10.9%)	2580 (7.3%)
18	Drug studies and need for vaccines	Vaccine needs, patient care, and treatments	837 (4.5%)	1214 (3.4%)
19	Study of coronavirus genomes	Microbiology (general)	201 (1.1%)	1054 (3%)
20	Biomolecular study of coronaviruses	Microbiology (general)	123 (0.7%)	2931 (8.3%)
21	Treatment and patient care for coronaviruses	Vaccine needs, patient care, and treatments	2234 (12.1%)	2958 (8.4%)
22	Children infected by COVID-19	Outbreaks (general coverage)	125 (0.7%)	216 (0.6%)
23	Clinical testing for COVID-19	Testing	475 (2.6%)	767 (2.2%)
24	COVID-19 cases and deaths reported	Outbreaks (general coverage)	1190 (6.5%)	1408 (4%)
25	Lessons learned for epidemic preparedness	Outbreaks (public health)	2038 (11.1%)	2801 (7.9%)
26	Outbreak and public health response to COVID-19	Outbreaks (public health)	138 (0.7%)	221 (0.6%)
27	Biomolecular study of coronaviruses	Microbiology (general)	56 (0.3%)	89 (0.3%)
28	Testing and transmission of COVID-19	Testing (mixed with transmission)	121 (0.7%)	296 (0.8%)
29	Biomolecular study of coronavirus strains	Microbiology (general)	51 (0.3%)	191 (0.5%)
30	Outbreaks and public health responses	Outbreaks (public health)	477 (2.6%)	1200 (3.4%)

Supplemental Information 6. The percentage of all COVID-19 abstracts in each of the broad research topic.

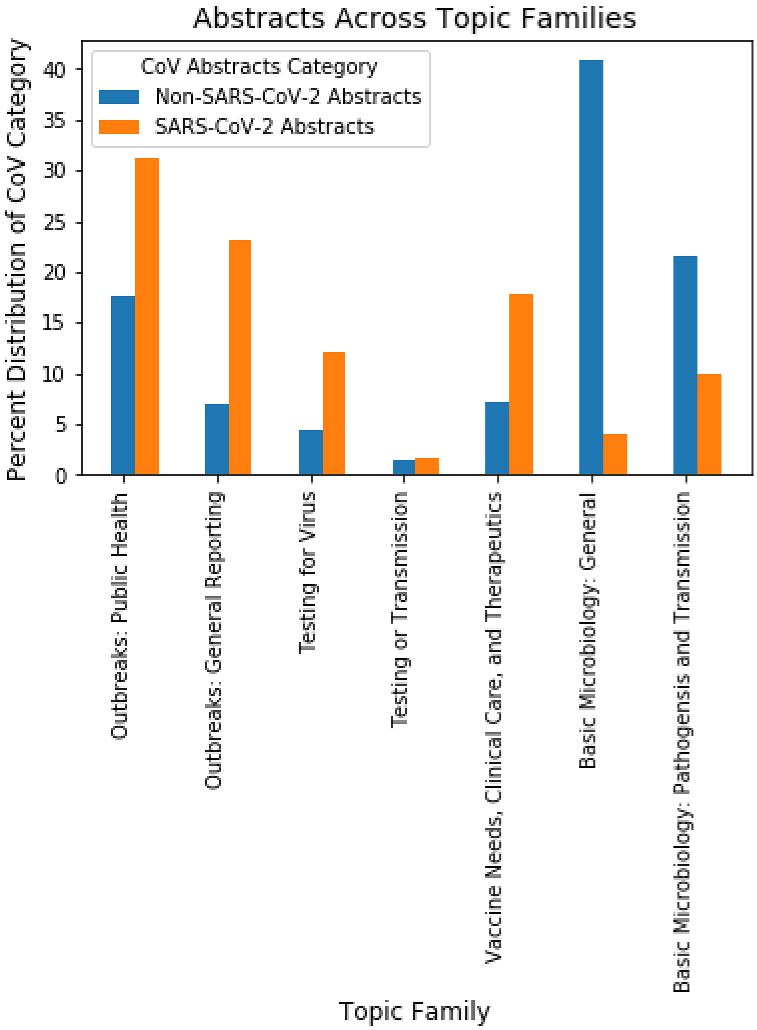


Figure SI.8.1. The distribution (y-axis) of abstracts of each category (blue for non-SARS-CoV-2 coronavirus abstracts and orange for SARS-CoV-2 abstracts) are plotted against each topic family (x-axis). Abstracts are unevenly distributed and SARS-CoV-2 abstracts tend to focus more on clinical, modeling, or field- (CMF) based study than basic microbiological research, even when compared with research on other coronaviruses.

Supplemental Information 7. Handling Incomplete Time Data

Of the 18,412 abstracts mentioning COVID-19 and its related terms, just 10 had no dates of any kind available. 5,837 were associated with the year 2020 and no further information. We assumed that these publications were evenly distributed throughout the year of 2020.

Supplemental Information 8. Specifications for Machine Learning Pipeline

We wrote a package that simplifies the preprocessing of the data, which is available on the Github repository[1]. It uses the nltk version 3.4.5 package's SnowBall stemmer to lemmatize words and the gensim package version 3.8.0 to preprocess text including the removal of punctuation, identification of bigrams, and creation of term frequency-inverse document frequency matrices (Supplemental Information 9). All cited software packages in this supplement are written in Python.

To implement dimensionality reduction in our pipeline, we used the package "scikit-learn" version 0.23.1. We specifically used the "TruncatedSVD" functionality, which enables the use of dimensionality reduction on sparse matrices like term-frequency-inverse-document-frequencies (Supplemental Information 9) and is analogous to principal components analysis.

We used the "gensim" package version 3.8.0 to conduct topic modeling with its LdaMulticore functionality.

All plots were created using matplotlib version 3.1.3.

[1] https://github.com/COVID19-DVRN/8-AI-Mapping-of-Relevant-Coronavirus-Literature/

Supplemental Information 9. Word Counts in Document-Term Matrices Serve as the Text's Computable Features

The smallest meaningful unit of semantic information in human language is a word. Therefore, we can infer that documents with very different words—perhaps shown through different frequencies of specific terms—discuss different topics; documents with similar frequencies for the same terms are likely focused on similar topics. This quantitative information feeds easily into classical machine learning algorithms, and so we captured it through document-term matrices (DTMs). Each cell in a DTM is filled by a metric for the frequency of a term (a column) in a specific document (a row). The DTM that we fed into PCA was a term frequency-inverse document frequency matrix, which down-weights terms that are more common across all products and thus aren't likely to be good differentiators between abstracts. LDA, on the other hand, expects simple term counts.

Since the words represent the feature space of DTMs, we took care to identify terms in a meaningful manner. Prior to computing the DTMs, we removed all punctuation and numbers from the text and lemmatized the remaining words so that words with the same stem are consolidated. This reduced the noise in the dataset and enhanced the consistency between machine-derived metrics and their semantic meaning. We also leveraged existing natural-language-processing packages (Gensim) to identify potentially useful word pairs, or "bigrams", as terms to feed into the DTMs.

However, these DTMs are ultimately extremely large.With 35,281 coronavirus abstracts and 69,667 unique words or bigrams identified, there are billions of elements in our matrices. Over 99.9% of these elements are simply 0 (i.e., instances where a word does not appear in one document, though it appears in others), and so we stored these DTMs in a "sparse" format. This means that we only store the coordinates and values of nonzero elements, and assume that all other elements are zero. The result is massive memory savings and thus computational feasibility, but this has significant algorithmic implications in dimensionality reduction (Supplemental Information 1).

Supplemental Information 10. COVID-19-related Keywords for Filtering Subsetted Abstracts

Item	Value(s)	Description and rationale
General search terms	Case sensitive: MERS Not case sensitive: “covid-19”, “coronavirus”, “corona virus”, “2019-ncov”, “sars-cov”, “mers-cov”, “severe acute respiratory syndrome”, “middle east respiratory syndrome”	● The presence of these search terms in an abstract indicated that the abstract was relevant to the study ● Mentioning these terms in an abstract made it more likely that a coronavirus was central to the research
Search terms for COVID-19	“COVID-19”, “COVID”, “2019-nCoV”, “SARS-CoV-2” (case sensitive)	● The presence of these terms in an abstract indicated that it was relevant to COVID-19
Search terms for MERS	Case sensitive: MERS Not case sensitive: “middle east respiratory”	● The presence of these terms in an abstract indicated that it was relevant to MERS-CoV
Search terms for SARS	Case sensitive: SARS Not case sensitive: “severe acute respiratory syndrome”	● The presence of these terms in an abstract indicated that it was relevant to SARS-CoV

Supplemental Information References

1. Fogel, P., Hawkins, D.M., Beecher, C., Luta, G., and Young, S.S. (2013). A Tale of Two Matrix Factorizations. The American Statistician *67*, 207–218.

2. Halko, N., Martinsson, P.G., and Tropp, J.A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review *53*, 217–288.

3. Wang, D.-Y., Guo, J.-M., Yang, Z.-Z., You, Y., Chen, Z.-C., Chen, S.-M., Cheng, H., Zhang, Y.-S., Jiang, D.-Z., Zuo, X.-L., et al. The first report of the prevalence of COVID-19 in Chronic myelogenous leukemia patients in the core epidemic area of China:multicentre, cross-sectional survey.

4. Su, J.-W., Wu, W.-R., Lang, G.-J., Zhao, H., and Sheng, J.-F. (2020). Transmission risk of patients with COVID-19 meeting discharge criteria should be interpreted with caution. Journal of Zhejiang University-SCIENCE B *21*, 408–410.

5. Cai, Q., Huang, D., Ou, P., Yu, H., Zhu, Z., Xia, Z., Su, Y., Ma, Z., Zhang, Y., Li, Z., et al. (2020). COVID-19 in a designated infectious diseases hospital outside Hubei Province, China. Allergy.

6. Chagnon, F., Lamarre, A., Lachance, C., Krakowski, M., Owens, T., Laliberté, J.F., and Talbot, P.J. (1998). Characterization of the expression and immunogenicity of the ns4b protein of human coronavirus 229E. Can. J. Microbiol. *44*, 1012–1017.

7. Zakhartchouk, A.N., Viswanathan, S., Mahony, J.B., Gauldie, J., and Babiuk, L.A. (2005). Severe acute respiratory syndrome coronavirus nucleocapsid protein expressed by an adenovirus vector is phosphorylated and immunogenic in mice. Journal of General Virology *86*, 211–215.

8. Taguchi, F., Ikeda, T., and Shida, H. (1992). Molecular cloning and expression of a spike protein of neurovirulent murine coronavirus JHMV variant c1-2. Journal of General Virology *73*, 1065–1072.