Patterns

Article

Machine Learning Maps Research Needs in COVID-19 Literature

Highlights

- AI/machine learning techniques can analyze coronavirus research at massive scale
- COVID-19 research has so far focused on non-lab-based (e.g., observational) research
- COVID-19 lab-based/basic microbiological research is less prevalent than expected

Authors

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

Correspondence

adoanvo@gmail.com (A.D.),
maimuna.majumder@childrens.harvard.edu (M.M.)

In Brief

An artificial intelligence/machine learning-based approach can be used to rapidly analyze COVID-19 literature and evaluate whether the research being produced at present addresses the existing knowledge gaps. We observe that COVID-19 research has been primarily clinical, modeling, or field based, and we observe significantly less laboratory-based research than expected when compared with other coronavirus (non-COVID-19) diseases. Our approach can be used to identify knowledge gaps and inform resource allocation decisions for research during future crises.

Doanvo et al., 2020, Patterns *1*, 100123
December 11, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.patter.2020.100123

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Machine Learning Maps Research Needs in COVID-19 Literature

Anhvinh Doanvo,[1,8,*] Xiaolu Qian,[2] Divya Ramjee,[3] Helen Piontkivska,[4] Angel Desai,[5] and Maimuna Majumder[6,7,*]
[1]COVID-19 Dispersed Volunteer Research Network, Washington, DC, USA
[2]University of Washington, Seattle, WA, USA
[3]Department of Justice, Law & Criminology, American University, Washington, DC, USA
[4]Department of Biological Sciences, Kent State University, Kent, OH, USA
[5]University of California, Davis, Sacramento, CA, USA
[6]Harvard Medical School, Boston, MA, USA
[7]Children's Hospital Computational Health Informatics Program (CHIP), Boston, MA, USA
[8]Lead Contact
*Correspondence: adoanvo@gmail.com (A.D.), maimuna.majumder@childrens.harvard.edu (M.M.)
https://doi.org/10.1016/j.patter.2020.100123

THE BIGGER PICTURE The impact of the COVID-19 pandemic has led scientists to produce a vast quantity of research aimed at understanding, monitoring, and containing the disease; however, it remains unclear whether the research that has been produced to date sufficiently addresses existing knowledge gaps. We use artificial intelligence (AI)/machine learning techniques to analyze this massive amount of information at scale. We find key discrepancies between literature about COVID-19 and what we would expect based on research on other coronaviruses. These discrepancies—namely, the lack of basic microbiological research, which is often expensive and time-consuming—may negatively impact efforts to mitigate the pandemic and raise questions regarding the research community's ability to quickly respond to future crises. Continually measuring what is being produced, both now and in the future, is key to making better resource allocation and goal prioritization decisions as a society moving forward.

1 2 **3** 4 5 **Production:** Data science output is validated, understood, and regularly used for multiple domains/platforms

SUMMARY

As of August 2020, thousands of COVID-19 (coronavirus disease 2019) publications have been produced. Manual assessment of their scope is an overwhelming task, and shortcuts through metadata analysis (e.g., keywords) assume that studies are properly tagged. However, machine learning approaches can rapidly survey the actual text of publication abstracts to identify research overlap between COVID-19 and other coronaviruses, research hotspots, and areas warranting exploration. We propose a fast, scalable, and reusable framework to parse novel disease literature. When applied to the COVID-19 Open Research Dataset, dimensionality reduction suggests that COVID-19 studies to date are primarily clinical, modeling, or field based, in contrast to the vast quantity of laboratory-driven research for other (non-COVID-19) coronavirus diseases. Furthermore, topic modeling indicates that COVID-19 publications have focused on public health, outbreak reporting, clinical care, and testing for coronaviruses, as opposed to the more limited number focused on basic microbiology, including pathogenesis and transmission.

INTRODUCTION

Since the beginning of 2020, investigators have published tens of thousands of studies on the coronavirus disease 2019 (COVID-19) pandemic, expanding the growing body of literature on the disease and its causative agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Research institutions have invested vast resources into closing key knowledge gaps regarding the pandemic, but the scope of current research remains unclear. In this paper, we aim to identify which topics COVID-19 research has focused on and which areas are likely to require additional attention.

This requires us to evaluate the breadth of COVID-19 research relative to past study. Thus far, attempts to do so have primarily considered citations, keyword co-occurrences, and other bibliometrics to identify influential literature.[1–3] These studies focus on (1) usage metrics to identify examples of literature that dominate the field and (2) manually generated metadata to explore narrow correlations between small groups of keywords that allude to large topics. Outside of traditional bibliometric studies, there are also large-scale applications of data science to coronavirus research that can indirectly help investigators analyze the research coverage of COVID-19. For example, ''LitCOVID,'' a literature hub created by the National Center for Biotechnology Information,[4] uses machine learning (ML) techniques to supplement manual review when curating and categorizing studies into discrete, predefined categories.

Despite previous bibliometric efforts, there is room for improvement as we attempt to address our key questions. First, in contrast with LitCOVID, our goals require techniques that can surface major topics without any *a priori* knowledge of what they might be. Predefined topics may bias analyses, which could detrimentally highlight insignificant topics and leave other important topics undetected. Second, we would ideally rely on natural language written by the publication authors themselves, rather than manually tagged keywords,[3] since such metadata may not be reliable or fully reflect latent issues discussed by the investigators who conducted the research. Third, defining primary topics in COVID-19 research solely by a select group of influential studies or on narrow correlations between a few metadata keywords at a time is insufficient because (1) topics may be broader than one or several highly influential studies and (2) topics may be comprised of complex correlations mapped between hundreds of different keywords. While a manual review might be desirable to capture this nuance,[5] this does not effectively scale over the tens of thousands of articles available.

Our methods address these issues by combining three techniques commonly used in natural language processing (NLP): document-term matrices (DTMs), dimensionality reduction, and topic modeling. Although these techniques are not methodologically novel, our specific application of them is: namely, we use them to analyze where there appears to be less COVID-19 research in comparison with existing research on other coronaviruses. Our DTMs allow us to draw on the *full text* of publication abstracts (as opposed to relying solely on keyword metadata). Our subsequent use of two ML techniques—dimensionality reduction and topic modeling—allows us to analyze complex information at scale *without any* a priori *knowledge of topics*, leveraging semantic trends between the tens of thousands of articles available to identify latent concepts and topics. This allows us to explore how the focus of COVID-19 studies differs from research on other coronaviruses by comparing the characteristics of COVID-19 articles identified through ML, with those pertaining to non-SARS-CoV-2 coronaviruses. These differences can then lend insight into possible gaps in research efforts for COVID-19.

We perform ML-aided analysis of research abstracts in the COVID-19 Open Research Dataset (CORD-19)[6] to automatically categorize ongoing research endeavors into *dynamically generated* categories, enabling us to identify topics that have received limited attention to date. By understanding the knowledge overlap between recently released abstracts on COVID-19 and abstracts related to other coronaviruses, we are able to gain insight into potential areas of SARS-CoV-2 research warranting further exploration. In addition, we propose a reusable framework for parsing an existing knowledge base about other emerging pathogens, such as the highly pathogenic avian influenza H5N1,[7,8] before they escalate to the level of a major epidemic or pandemic threat. In the future, such a framework will allow analysts to rapidly infer where research gaps might exist by comparing the cross-topic distribution of literature on an emerging disease with the distribution of research on related but previously explored pathogens.

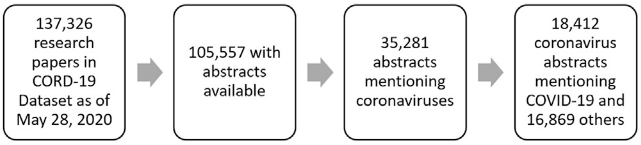

137,326 research papers in CORD-19 Dataset as of May 28, 2020
105,557 with abstracts available
35,281 abstracts mentioning coronaviruses
18,412 coronavirus abstracts mentioning COVID-19 and 16,869 others

Figure 1. The Data Filtering Process

This figure highlights the number of abstracts represented at each stage of the subsetting process.

RESULTS

We first filtered the CORD-19 dataset for publications that had abstracts available (Figure 1). Of these, a relatively small subset mentioned coronaviruses in their abstracts. Those that did not mention coronavirus search terms in their abstracts contained coronavirus-related terms somewhere else in the text, such as in its citations.

In August 2020, we updated our data analysis using the July 31, 2020, version of CORD-19. This dataset contained 65,929 abstracts mentioning coronaviruses, 48,670 of which mentioned COVID-19 or SARS-CoV-2 specifically.

Principal-Component Analysis Indicates a Limited Number of Laboratory Studies on Viral Mechanisms of SARS-CoV-2

While principal-component analysis (PCA) highlighted the abstracts' most prominent patterns in the first principal component (PC), which captured 0.25% of the data's variance, these patterns were not effective at distinguishing between COVID-19 and non-COVID-19 literature. Figure 2A demonstrates no meaningful difference between the two distributions of projection values from COVID-19 and non-COVID-19 abstracts onto the first PC, indicating a shared pattern of variance, i.e., both groups appear to discuss similar questions, approaches, and techniques using similar vocabulary within this pattern.

The patterns that successfully differentiated between the two groups were beneath the first PC and within the second PC, which captured 0.82% of the data's variance (see Supplemental Information 1 for why PC2 captures more variance than PC1 in this case), where the projection value distributions presented distinguishing patterns (Figure 2B). While there was considerable overlap between the two groups along this PC, the centers of the two distributions differed substantially, indicating that non-COVID-19 literature tended to cover different issues than those

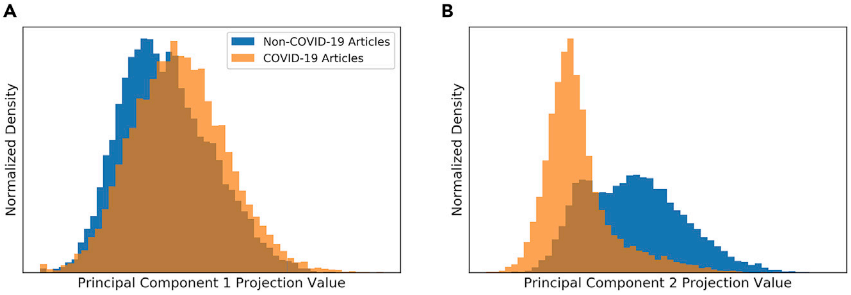


Figure 2. Distribution of COVID-19 (Orange) and Non-COVID-19 (Blue) Abstracts' Projection Values along the First Two PCs

The projection values are represented by the x axis, while the densities are represented by the y axis. (A) Shows PC1 and (B) PC2. PC1 does not effectively distinguish between COVID-19 and non-COVID-19 abstracts. PC2 shows distinct distributions between COVID-19 and non-COVID-19 abstracts, indicating distinct vocabularies used in these abstracts. Plots were generated via kernel density smoothing, across a linear scale and without dropping any outliers. Distributions were normalized by density, not raw counts.

covered by COVID-19 literature. Our interpretation of this PC relied on identifying terms that had values with the greatest magnitude (Supplemental Information 2 and Supplemental Information 3). Ultimately, Figure 2 indicates that while variance among non-COVID-19 abstracts (blue) stretched over much of the second PC, projection values of COVID-19 abstracts (orange) were concentrated in a smaller area, reflecting the narrower scope of COVID-19 abstracts considering that the virus and associated disease have only been studied since December 2019.

When we split the studies into subsets for the three human coronaviruses that have potential for severe infection, we found that the distributions of SARS-CoV and MERS-CoV abstracts in the PC projection space were unique to each virus (Figure 3). SARS-CoV-2 abstracts appeared to share a space in common with both MERS-CoV and SARS-CoV, likely reflecting some shared terminology and possible ongoing attempts to leverage existing knowledge of the other two viruses to learn about SARS-CoV-2. However, SARS-CoV-2 abstracts are much more concentrated among lower projection values. Notably, MERS-CoV and SARS-CoV abstracts were spread more evenly along the second PC, reflecting greater breadth and variation along these PCs that can be attributed to a broader range of studies focused on these pathogens as compared with SARS-CoV-2. This may be in part due to the much longer time that has been spent studying these viruses to date.

To identify terms associated with differences between COVID-19 and non-COVID-19 abstracts on PC2, we examined patterns of lemmatized terms from the respective abstracts (Figure 3). The projection values of COVID-19 abstracts on PC2 were lower and associated with emergent COVID-19 clinical-, modeling-, or field-based (CMF) research—such as observational, clinical, and epidemiological studies—exemplified by stem terms "patient," "pandem," "estim," and "case." Words in the opposite direction on PC2—such as "protein," "cell," "bind," and "express"—can be associated with viral biology and basic disease processes studied in biomolecular laboratories. COVID-19 abstracts were thus mostly associated with research conducted outside of laboratories, e.g., in hospitals, likely reflecting the pandemic reality of data collection alongside (and often secondary to) clinical care.

The high-level abstraction reflected by PC2 informed our designation of the extent that COVID-19 research included studies with any CMF design—ranging from epidemiological studies to retrospective reviews of clinical outcomes, case studies, and randomized clinical trials—or laboratory-driven research—including observational microscopy, experimentation with antiviral compounds, derivation of protein structures, and studies of animal or cell culture models. Overall, COVID-19 abstracts appeared more likely to have terms associated with CMF research rather than laboratory studies based on comparisons of distributions for key terms in the COVID-19 and non-COVID-19 abstracts (Figure 4; examples in Supplemental Information 3). This partition along research design for non-COVID-19 and COVID-19 abstracts was also evident in the abstract texts: 90% of the abstracts in the bottom 1% of projection values along the second PC were related to COVID-19; conversely, only 1% of the abstracts in the top 1% were related to COVID-19. In the future, we can implement PCA again to observe time-varying trends in CMF-based and laboratory-driven research. If COVID-19 research continues to focus significantly more on CMF-based study than laboratory-driven research, we would expect this to be reflected in the separation between COVID-19 and non-COVID-19 research along a new PC that separates these two categories of research.

When we reran dimensionality reduction and topic modeling on new data through July 31, 2020, we found that the body of CMF research has continued to grow far more quickly than laboratory-based research (figures available in Supplemental Information 4). Our PCA analysis found that PC2 strongly

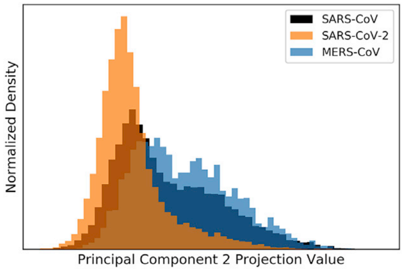


Figure 3. Distribution of SARS-CoV-2 (Orange), SARS-CoV (Black), and MERS-CoV (Blue) Abstracts' Projection Values along the Second PC

The projection values are represented by the x axis, while the densities are represented by the y axis. The second PC provides distinct separation of SARS-CoV-2, as well as mild separation between abstracts mentioning the two other human CoVs capable of causing severe illness (SARS-CoV and MERS-CoV). These distributions were separated by whether the studies mentioned one of the three viruses.

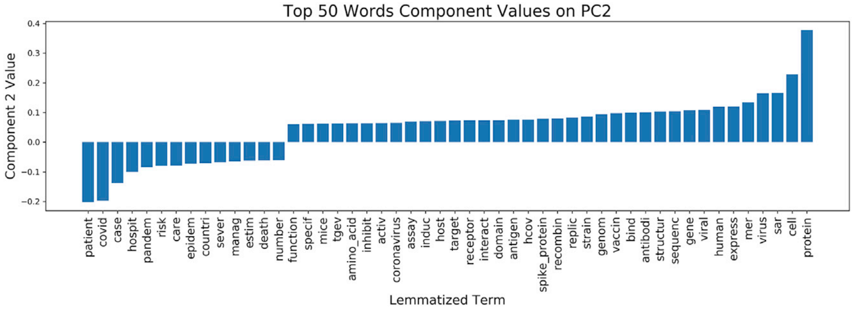


Figure 4. Component Values of Terms across PC2

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.

differentiated between abstracts related to SARS-CoV-2 and abstracts that mentioned other coronaviruses. SARS-CoV-2 abstracts tended to have lower PC2 projection values, which were associated with CMF-related terms, such as ''hospit,'' ''case,'' and ''risk.'' Conversely, non-SARS-CoV-2 studies tended to have higher projection values, which were associated with laboratory-based research, including ''antibodi,'' ''cell,'' and ''protein.''

Topic Modeling Suggests Additional Differences in Specific Research Subareas

Topic modeling helped characterize differences between research topics discussed in COVID-19 and non-COVID-19 abstracts. Results from the latent Dirichlet allocation (LDA) model suggested that, similar to the pattern observed in Figure 5, there was clear differentiation between COVID-19 and non-COVID-19 abstracts across 30 topics (Figure 6; Supplemental Information 5). There were five topics in particular—(1) Topic 14: outbreaks' impact on healthcare services, (2) Topic 15: testing for coronaviruses, (3) Topic 17: epidemic cases and modeling, (4) Topic 21: clinical care and therapeutics, and (5) Topic 25: lessons learned for epidemic preparedness—that accounted for 58% of all COVID-19 abstracts and just 17% of non-COVID-19 abstracts. COVID-19 abstracts were thus disproportionately concentrated in these five topics relative to non-COVID-19 abstracts.

Across the 30 topics, we grouped several topics into topic families based on internal commonalities (Supplemental Information 5), including (1) updates on the spread of and events related to coronavirus outbreaks (including two subfamilies: general updates versus public health responses); (2) testing for coronaviruses; (3) clinical care, therapeutics, and the need for vaccinations; and (4) basic microbiological research (which included two subfamilies: a general catch-all subfamily versus a subfamily specific to pathogenesis and transmission). When divided by topic family, the disparity between COVID-19 and non-COVID-19 research in the first and fourth topic families showed that COVID-19 abstracts appeared to be heavily concentrated on topics that typically included field-based data (the first topic family, on outbreak reporting) and excluded laboratory-based studies (the fourth topic family, on basic microbiology). However, one exception was that COVID-19 was overrepresented in studies on testing (especially diagnostics; the second topic family), which included both the laboratory development of the tests and their field application (Supplemental Information 6).

We observed similar results in our update with data through July 31, 2020, where we detected five topic families, including (1) clinical issues: testing and diagnostics, (2) societies and outbreaks: responses to mitigate them and their impact on society, (3) basic microbiological study, (4) general outbreak reporting, and (5) modeling disease transmission. Basic microbiological study in COVID-19 research has lagged behind research in the other topic families (Supplemental Information 4).

Documents Analyzed through ML Highlight Trends over Time

We also examined the rates of publication and preprint submission for COVID-19 abstracts along PC2 (Figure 7A) and the previously mentioned topic families (Figure 7B) (for details on how we used incomplete time data, see Supplemental Information 7). From the beginning of 2020, COVID-19 abstracts tended to have lower projection values for the second PC, reflecting the relatively higher number of CMF studies emerging during the early stages of the pandemic compared with laboratory-based studies.

Likewise, the growth of studies in different topic families for COVID-19 was unevenly distributed (Figure 7B). From January 2020 through the end of May 2020, publications related to COVID-19 were dominated by studies involving (1) outbreak and responses and (2) patients and healthcare services, similar to the observed faster pace of CMF research in the PCA results. Publications regarding viral mechanisms and biomolecular processes related to SARS-CoV-2 grew at a slower pace.

We observed similar trends when we reran PCA and LDA on data through July 31, 2020. Data up through that point in time continues to suggest that COVID-19 research has been dominated by CMF-based investigation and that laboratory-driven study has lagged.

DISCUSSION

Our findings demonstrate the utility of our novel NLP-driven approach for determining potential areas of underrepresentation in current research efforts for COVID-19. By applying unsupervised ML methods to CORD-19, we identified overarching key research topics in existing coronavirus- and COVID-19-specific abstracts, as well as the distribution of abstracts among topics and over time. Our results support a previous bibliometric study that also found more frequent appearances of epidemiological keywords in COVID-19 research compared with research on other coronaviruses.[3] However, our study presents the unique finding that laboratory-based COVID-19 studies, including those on genetic and biomolecular topics, are underrepresented

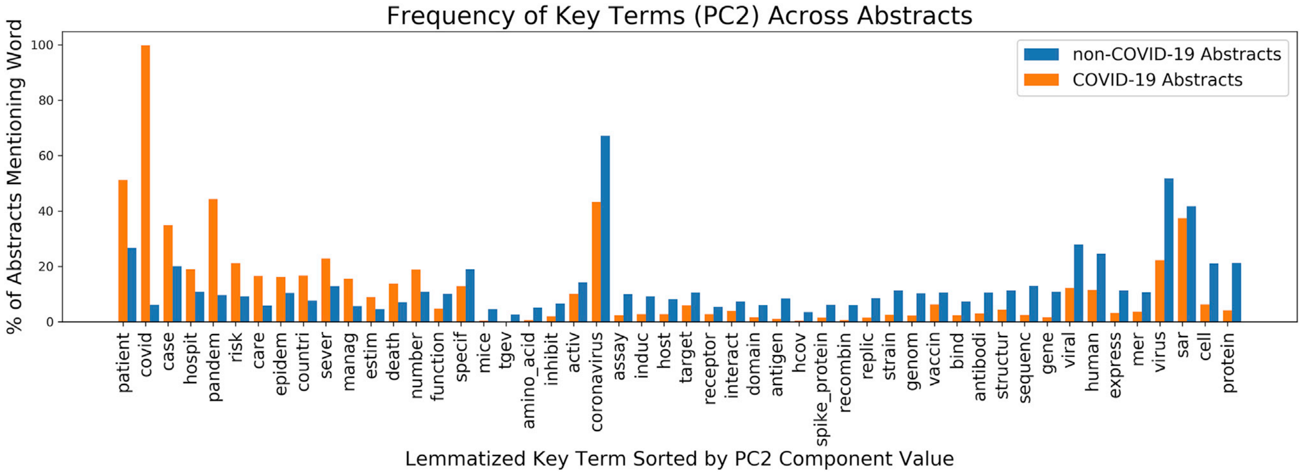


Figure 5. Top Terms across PC2 and Their Distribution in COVID-19 and Non-COVID-19 Abstracts

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.

relative to studies of epidemiological and clinical issues, particularly when compared with the distribution of previous research on other coronaviruses. We continued to observe this trend when we updated our May 30, 2020, analysis with data through July 31, 2020. In particular, the pace of basic microbiological study has lagged behind that of research in other areas (e.g., topic families derived from LDA, including clinical issues, societal impacts and policies, general reporting, and transmission modeling), all of which are CMF-based.

Furthermore, we developed a framework that improves upon existing bibliometric studies in three key ways; namely, our approach (1) maps connections between publications by relying directly on the abstracts instead of the narrow information gained from metadata as in other bibliometric analyses, including those from other fields[9,10]; (2) uses ML to explore latent semantic information of vast scale and complexity to identify hidden trends; and (3) does not rely on any *a priori* knowledge of what topics we expect coronavirus literature to cover but rather highlights them without any preconceived assumptions. We believe this methodology can be reused to rapidly explore possible research gaps during future epidemics and pandemics. More specifically, NLP and ML could serve as a way to identify the major concepts and topics covered by past research in comparison against present efforts; if certain topics identified in earlier research are not well represented in more recent studies on the emerging pathogen, they could be interpreted as potential research gaps.

The distribution of COVID-19 and non-COVID-19 abstracts from our PCA results suggests that, at the time of writing (CORD-19 dataset release on July 31, 2020), the breadth of published research for COVID-19 is relatively narrow compared with that of published non-COVID-19 studies (Figures 1 and 2). As shown in our results, keywords associated with biomolecular processes (e.g., viral structure, pathogenesis, and host cell interactions) appeared more frequently in non-COVID-19 abstracts than in COVID-19 abstracts. This finding reflects the emergent nature of SARS-CoV-2. Nonetheless, the availability of laboratory studies for other coronaviruses represents an opportunity for generating hypothesis-driven research questions grounded in empirical research.

It is worth discussing two possible explanations for the lack of SARS-CoV-2 laboratory study. First, researchers may be working under the assumption that biological processes of SARS-CoV-2, including life cycle and interactions with the human host, are comparable with those of SARS-CoV due to their genetic similarity and relatedness.[11–13] For example, several previous SARS-CoV studies on host cell entry helped identify the angiotensin converting enzyme 2 (ACE2) protein as a mediator for SARS-CoV-2 infection.[14] Likewise, CD147 and GRP78 proteins have been hypothesized to play a role in cell entry for SARS-CoV-2 based on earlier SARS-CoV and MERS-CoV findings, although additional studies are needed.[15–18] An alternative explanation, however, is that, by researching specific subjects on other coronaviruses, researchers might have learned which areas warrant less effort for new coronaviruses, such as SARS-CoV-2.

But while relying upon assumed similarities is an important first step, it becomes increasingly critical to identify features that are unique to each virus as work progresses. A full exploration of the characteristics associated with SARS-CoV-2—regardless of whether certain areas have been highlighted by other coronavirus research as promising or not—is essential to the development of vaccines, therapeutics, and tactics to mitigate transmission of this particular pathogen. Yet the scope of literature for biological processes unique to SARS-CoV-2 is currently quite limited as a whole, and perhaps even more limited than what our PCA results suggest if most SARS-CoV-2 literature relies heavily on other coronavirus research.

This underrepresentation of studies on biomolecular processes could also be attributed to the rapid worldwide spread of SARS-CoV-2 that occurred within mere months of its emergence, necessitating an unprecedented response from healthcare and public health infrastructures globally. Our PCA results reflect an overwhelming concern regarding the exponential

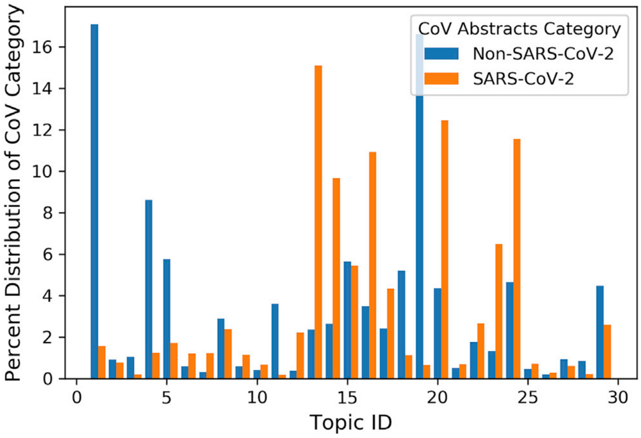


Figure 6. Distribution of Literature across Topics

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.

spread of the virus and risks for transmission involved with more frequent appearances of stem terms, such as “pandem,” “outbreak,” “estim,” “countri,” “number,” and “risk” in COVID-19 abstracts. This was also supported by our topic modeling results, which indicated that 58% of COVID-19 abstracts fell into just 5 of 30 topics, generally related to healthcare services, the pandemic’s public health issues, and testing for coronaviruses (Figures 7A and 7B). The more rapid growth of CMF research, relative to laboratory-driven research, mirrors the current response to the pandemic in the United States where the initial focus on pressing epidemiological and clinical concerns is now followed by interest in experimental investigations, including those of structural mechanisms for host cell entry and possible therapeutic targets.

Overall, our findings reflect a clear divide between COVID-19 and non-COVID-19 abstracts based upon research design; unlike CMF research, laboratory-driven SARS-CoV-2 research is either still underway or has only just been initiated. This can be attributed in part to the fact that laboratory research is often a labor-intensive process within a federally regulated infrastructure that depends on the availability of timely, project-based funding as well as longer-term funding. Our findings also suggest that the pace of research on SARS-CoV-2 biomolecular processes is potentially insufficient given the global threat posed by the virus (Figures 7A and 7B). This lag may adversely impact the development of antivirals and other therapeutic interventions, adding strain to already overwhelmed healthcare systems. Furthermore, these trends raise questions about the readiness of institutions supporting the research community in times of extraordinary stress. Previous experiences with global pandemics, such as H1N1, have resulted in various policy recommendations[19] to maintain and enhance readiness in laboratory-based research, and analysis on the effectiveness of recommendations arising from these experiences may be worthwhile.

While PCA identified a prominent pattern that differentiated between COVID-19 and non-COVID-19 literature, the topic families derived from LDA refined our understanding of knowledge gaps and research needs in COVID-19 literature by delineating specific research areas. This included an underrepresentation of studies on basic microbiological examination of SARS-CoV-2, including its pathogenesis and transmission. Research on these issues is published at a slower pace than CMF studies (e.g., those on clinical topics, outbreak response, and statistical reporting) and research on testing (Figure 7B). Even when compared with the distribution of non-COVID-19 research, COVID-19 research was more heavily focused on topics within the CMF realm (Supplemental Information 6).

We recognize that the number of abstracts in each of these topics does not necessarily represent scientific progress made in these areas, but they *do* reflect the pace of research and potential availability of public knowledge. This indicates either a mismatch between the level of effort in these issues and the urgency of work or time lags inherent to these fields that constrain the responsiveness of the scientific community. Increased and consistent funding of emerging pathogens research, including support of basic research even when there is no immediate threat of an outbreak, would allow us to maintain a proactive posture in accumulating available knowledge rather than overreliance on reactivity.

These conclusions must be caveated by several limitations that must be acknowledged. First, while CORD-19 includes a vast quantity of coronavirus-related publications, it potentially omits relevant literature from other databases, such as the Social Science Research Network (SSRN) or arXiv (a preprint server for studies in mathematics, computer science, and quantitative biology, among other topics). This may have constrained the representativeness of our analysis on COVID-19 literature, thus affecting the external validity of our findings. Second, analyzing abstracts inherently excludes ongoing research efforts because not all relevant studies are publicly available or have released preprints. Third, the number of publications does not directly represent progress in research areas. Fourth, the high-level trends we observed through our unsupervised ML approaches may not completely align with how researchers identify and process specific research topics. The counts of words in DTMs informing the ML algorithms may not directly capture the ideas researchers are trying to convey and may therefore gloss over nuances in the literature.

These four limitations are somewhat mitigated by both the nature of the data sources and the needs of the research community. For the first, the excluded sources (SSRN and arXiv) heavily focus on research within the CMF arena, indicating that our conclusions on the rapid pace of CMF COVID-19 research (versus lab-based research) are conservative. For the second, existing research pipelines have been accelerated in the pandemic, especially with the proliferation of preprint services. This reduces the lag between the discovery of knowledge and the availability of an abstract to ingest in our data pipeline. Third, the number of publications in each area may imply a relative difference in research productivity for different topics, and thus may still serve as a proxy for indicating such progress or the attention given to specific issues. And finally, our ML-based method offers the chance to quickly review large quantities of text at scale and highlight underlying trends. Both speed and scale are crucial to informing time-sensitive decisions on policy and priorities to facilitate the most impactful research.

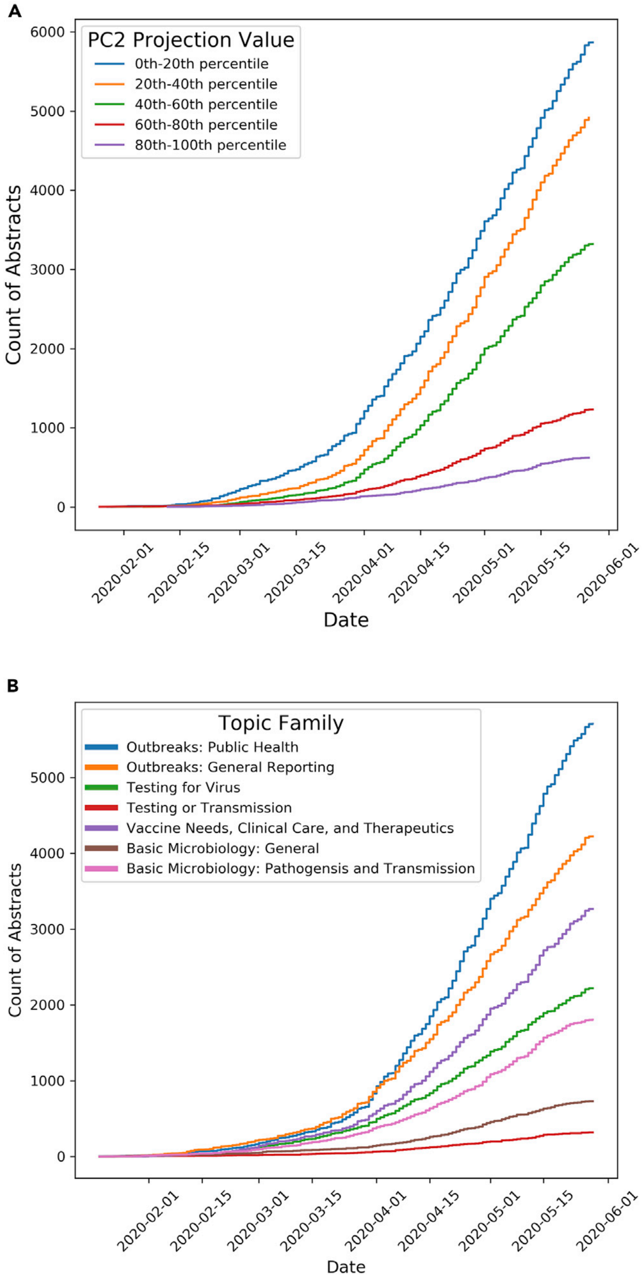


Figure 7. Trends over Time

(A) 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.
(B) COVID-19 research is predominantly focused on outbreak reporting and public health 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 (A), each group is comprised of abstracts within a certain range of projection values on PC2, and for (B), each group is comprised of abstracts within a certain topic family.

Our ML-based study offers insights into potential areas for research opportunities to tackle key gaps in our knowledge regarding SARS-CoV-2 and COVID-19. Our findings showcase the need for institutions to support laboratory-driven research on an ongoing basis—not only during a crisis—to enable a proactive preparedness posture. While we would prefer future pandemics to be prevented through comprehensive surveillance and mitigation of new pathogens, if a crisis emerges in the future, the urgency to understand knowledge gaps will remain. Our approach can be reused in such scenarios to rapidly explore potential research gaps and to inform future efforts for other emergent pathogens. By using previous research or studies focused on related pathogens as a baseline, the trends and gaps in knowledge regarding an emergent pathogen can be monitored to ensure that key areas in research are not under-resourced in the middle of a crisis.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Anhvinh Doanvo (adoanvo@gmail.com).

Materials Availability

This study did not generate any unique reagents.

Data and Code Availability

The data of CORD-19 are available to download here: https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html. We used the data released on May 30, 2020, for our initial analysis; for our update, we used the data released on July 31, 2020. All of our code is available for download from GitHub here: https://github.com/covid19-dvrn/8-ai-mapping-of-relevant-coronavirus-literature.

Overview

Without using any pre-existing knowledge about the abstracts' topics, we used unsupervised ML to determine differences between COVID-19 and non-COVID-19 abstracts in our corpus of documents. A dimensionality reduction approach was used to identify principal patterns of variation in the abstracts' text, followed by topic modeling to extract high-level topics discussed in the abstracts.[20] Our data pipeline is available on GitHub (https://github.com/COVID19-DVRN/8-AI-Mapping-of-Relevant-Coronavirus-Literature) and the specific software packages we used are described in Supplemental Information 8.

Dataset and Preprocessing

We obtained research abstracts from CORD-19 on May 28, 2020. Generated by the Allen Institute for artificial intelligence, and in partnership with other research groups, CORD-19 is updated daily with coronavirus-related literature. Peer-reviewed studies from PubMed/PubMed Central, as well as preprints from bioRxiv and medRxiv, are retrieved using specific coronavirus-related keywords ("COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"). At the time of writing, CORD-19 contained approximately 137,000 articles, including both full text and metadata for all coronavirus research articles, with ∼40% of the dataset classified as virology-related.[6] We focused our analysis on the abstracts of articles in CORD-19.

As some of the CORD-19 abstracts were neither relevant to SARS-CoV-2 nor other coronaviruses, we first filtered the CORD-19 data to isolate coronavirus-specific abstracts by searching for abstracts that mentioned relevant terms. These abstracts served as our "documents" associated with the sparse DTMs in our NLP pipeline (DTMs and sparse matrices are described in more detail in Supplemental Information 9). We also identified abstracts for only COVID-19-related studies by filtering for COVID-19-related keywords within this subset (Supplemental Information 10).

Methodology

We used two ML techniques to identify key trends in coronavirus literature: dimensionality reduction and topic modeling, discussed below. For software packages and additional details behind the data pipeline, see Supplemental Information 8.

Dimensionality Reduction

PCA is a dimensionality reduction algorithm that summarizes data by determining linear correlations between variables.[21] PCA identifies individual patterns of variance, or PCs, in DTMs that differentiate documents from one another, highlighting key trends in the data. For example, in a simple corpus with two mutually exclusive topics, such as ML and health infrastructure, the terms "machine" and "learning" would be correlated with one another. PCA would recognize these terms as an important source of variation, providing a way to differentiate documents about either topic ("machine learning" versus "health infrastructure") by the frequency of these terms.

When PCA is applied to DTMs, PCs represent patterns differentiating different documents, typically ordered by their prominence. This means that earlier PCs almost always capture more variance than later PCs. However, in some cases, PC1 may capture less variance than PC2 if certain precomputation processing is not conducted (see Supplemental Information 1 for more details). Each detected pattern reflects both the contextual links between words and their level of importance within the texts. Words with component values of the greatest magnitude on each PC most strongly drive the pattern that each individual PC recognizes. For example, if "machine" and "healthcare," respectively, have highly negative and highly positive values on a particular PC, then that PC detects the pattern that when "machine" appears in a text, "healthcare" appears less often. Another PC may detect a different pattern of variance, such as when some documents mention "deep learning" more often than others.

The projection values of the text corpus onto the PCs suggest what concept each document discusses and to what extent, relative to the average document within the corpus. Following the previous example, strongly negative projection values on the first PC, which would capture the data's most prominent patterns, indicate that the document mentions "machine" more often than the average and thus is more likely to focus on ML. In addition, projection values on the second PC could distinguish between ML documents by focus, or lack thereof, on deep learning or other techniques. This approach enables us to delineate between different groups of abstracts by visualizing differences in their projections on the top PCs. After applying PCA to the DTMs of our abstracts, we identified which PCs successfully separated COVID-19 and non-COVID-19 abstracts. We then used the component values with the largest magnitude on these PCs to interpret them.

Applying PCA to DTMs can be computationally expensive and sometimes infeasible because of their extremely high dimensionality (i.e., many different words are being counted). Furthermore, traditional implementations of PCA that rely on calculating covariance matrices[21] cannot be used on sparse matrices, and thus would not be applicable to our sparse DTMs where instances in which words do not appear in specific documents are implied (Supplemental Information 9) but not recorded. For information on the modified procedure we used to mitigate these limitations, see Supplemental Information 1.

Topic Modeling

After establishing high-level trends using PCA, we used LDA, a topic modeling method, to add nuance to observed differences between COVID-19 and non-COVID-19 literature and examine potential topics of interest. LDA is an unsupervised probabilistic algorithm that extracts hidden topics from large volumes of text.[22] Once trained to discover words that separate documents into a predetermined number of topics, LDA can estimate the "mixture" of topics associated with each document. These mixtures suggest the dominant topic for a document that is then used to assign a document to an overarching topic category. For example, LDA may separate documents into two topics, one on "machine learning" and another on "healthcare," and if a particular document's mixture is 60% "machine learning" and 40% "healthcare," it would assign that document to a "machine learning" topic category.

The predetermined number of topics is the most important hyperparameter in an LDA model, as models with sub-optimal number of topics fail to summarize data in an efficient manner.[22,23] The number of topics can be determined by (1) identifying a model that has a low perplexity score and high coherence value when applied to an unseen dataset or (2) conducting a principled, manual assessment of the topics that arise. Perplexity is a statistical measure of how imperfectly the topic model fits a dataset, and a low perplexity score is generally considered to provide better results.[23] Similarly, topic models with high coherence values are considered to offer meaningful, interpretable topics.[24,25] Thus, a model with a low perplexity score and a high coherence value is more desirable when choosing the optimal number of topics. Our initial implementation of LDA showed no optimal value for the number of topics, even as it approached ~100, potentially reflecting a relatively shallow yet broad pool of COVID-19 publications. We ultimately identified 30 topics via manual review of topics from topic models with different numbers of topics to identify which model satisfied two criteria: (1) topics that were relatively specific, focusing on a single subject matter and (2) topics that would typically be non-redundant with one another.

Reusability

Our framework can be reused to identify literature gaps for other fields, including emerging pathogens. This would require researchers to preprocess their data to create a document-term matrix for literature on the emergent pathogen and that of related but previously observed pathogens. Investigators can then conduct PCA and LDA to identify (1) a PC that separates abstracts in the two bodies of literature, (2) the terms that enable them to interpret the meaning of that PC, and (3) the distribution of literature in each of the two categories across several topics. PCA and LDA together can quickly identify concepts and topics that separate the two bodies of literature by tapping data from correlations between numerous different words at once across all the literature considered.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100123.

ACKNOWLEDGMENTS

This project is part of the COVID-19 Dispersed Volunteer Research Network (COVID-19-DVRN) led by M.M. and A. Desai. M.M. is supported by a grant from the NIH (T32HD040128) and is a faculty member at both Harvard Medical School and Boston Children's Hospital's Computational Health Informatics Program (CHIP). H.P. is also supported by a grant from the NIH (R21AG064479-01). The funders for this grant had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The authors thank Shagun Gupta for her thoughtful feedback on an earlier version of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, A. Doanvo, X.Q., D.R., H.P., A. Desai, and M.M.; Data Curation, Methodology, Software, and Visualization, A. Doanvo and X.Q.; Formal Analysis and Investigation, A. Doanvo, X.Q., D.R., and H.P.; Writing – Original Draft, A. Doanvo, D.R., X.Q., and H.P.; Writing – Review & Editing, A. Doanvo, D.R., H.P., X.Q., A. Desai, and M.M.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: June 17, 2020
Revised: August 13, 2020
Accepted: September 14, 2020
Published: September 16, 2020

REFERENCES

1. Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A.A., Salhab, H., Fares, M., and Khachfe, H.H. (2020). A bibliometric analysis of COVID-19 research activity: a call for increased output. Cureus *12*, e7357.

2. Golinelli, D., Nuzzolese, A.G., Boetto, E., Rallo, F., Greco, M., Toscano, F., and Fantini, M.P. (2020). The impact of early scientific literature in response to COVID-19: a scientometric perspective. medRxiv. https://doi.org/10.1101/2020.04.15.20066183.

3. Hossain, M.M. (2020). Current Status of Global Research on Novel Coronavirus Disease (COVID-19): A Bibliometric Analysis and Knowledge Mapping. F1000Research *9*, https://doi.org/10.12688/f1000research.23690.1.

4. Chen, Q., Allot, A., and Lu, Z. (2020). Keep up with the latest coronavirus research. Nature *579*, 193.

5. Liu, N., Chee, M.L., Niu, C., Pek, P.P., Siddiqui, F.J., Ansah, J.P., Matchar, D.B., Lam, S.S.W., Abdullah, H.R., Chan, A., et al. (2020). Coronavirus disease 2019 (COVID-19): an evidence map of medical literature. BMC Med. Res. Methodol. *20*, 177.

6. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., et al. (2020). CORD-19: the COVID-19 open research dataset. arXiv, [cs.DL].

7. Kilpatrick, A.M., Chmura, A.A., Gibbons, D.W., Fleischer, R.C., Marra, P.P., and Daszak, P. (2006). Predicting the global spread of H5N1 avian influenza. Proc. Natl. Acad. Sci. U S A *103*, 19368–19373.

8. Kissler, S.M., Gog, J.R., Viboud, C., Charu, V., Bjørnstad, O.N., Simonsen, L., and Grenfell, B.T. (2019). Geographic transmission hubs of the 2009 influenza pandemic in the United States. Epidemics *26*, 86–94.

9. de Oliveira, O.J., da Silva, F.F., Juliani, F., Barbosa, L.C.F.M., and Nunhes, T.V. (2019). Bibliometric method for mapping the state-of-the-art and identifying research gaps and trends in literature: an essential instrument to support the development of scientific projects. In Scientometrics Recent Advances (IntechOpen).

10. Campbell, D., Picard-Aitken, M., Côté, G., Caruso, J., Valentim, R., Edmonds, S., Williams, G.T., Macaluso, B., Robitaille, J.-P., Bastien, N., et al. (2010). Bibliometrics as a performance measurement tool for research evaluation: the case of research funded by the National Cancer Institute of Canada. Am. J. Eval. *31*, 66–83.

11. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

12. Petrosillo, N., Viceconte, G., Ergonul, O., Ippolito, G., and Petersen, E. (2020). COVID-19, SARS and MERS: are they closely related? Clin. Microbiol. Infect. *26*, 729–734.

13. Zhang, Y.-Z., and Holmes, E.C. (2020). A genomic perspective on the origin and emergence of SARS-CoV-2. Cell *181*, 223–227.

14. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell *181*, 271–280.e8.

15. Wang, K., Chen, W., Zhou, Y.-S., Lian, J.-Q., Zhang, Z., Du, P., Gong, L., Zhang, Y., Cui, H.-Y., Geng, J.-J., et al. (2020). SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. bioRxiv. https://doi.org/10.1101/2020.03.14.988345.

16. Chen, Z., Mi, L., Xu, J., Yu, J., Wang, X., Jiang, J., Xing, J., Shang, P., Qian, A., Li, Y., et al. (2005). Function of HAb18G/CD147 in invasion of host cells by severe acute respiratory syndrome coronavirus. J. Infect. Dis. *191*, 755–760.

17. Ibrahim, I.M., Abdelmalek, D.H., Elshahat, M.E., and Elfiky, A.A. (2020). COVID-19 spike-host cell receptor GRP78 binding site prediction. J. Infect. *80*, 554–562.

18. Chu, H., Chan, C.-M., Zhang, X., Wang, Y., Yuan, S., Zhou, J., Au-Yeung, R.K.-H., Sze, K.-H., Yang, D., Shuai, H., et al. (2018). Middle East respiratory syndrome coronavirus and bat coronavirus HKU9 both can utilize GRP78 for attachment onto host cells. J. Biol. Chem. *293*, 11709–11726.

19. French, M.B., Loeb, M.B., Richardson, C., and Singh, B. (2009). Research preparedness paves the way to respond to pandemic H1N1 2009 influenza virus. Can. J. Infect. Dis. Med. Microbiol. *20*, 63–e66.

20. Ziegler, A. (2016). An Introduction to Statistical Learning with Applications. R.G. James, D. Witten, T. Hastie, and R. Tibshirani (2013). Berlin: Springer. 440 pages, ISBN: 978-1-4614-7138-7. Biom. J. *58*, 715–716.

21. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. *24*, 417–441.

22. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

23. Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics *16*, S8.

24. Aletras, N., and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (Association for Computational Linguisitics), pp. 13–22.

25. Newman, D., Bonilla, E.V., and Buntine, W. (2011). Improving topic coherence with regularized topic models. In Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 496–504.

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.

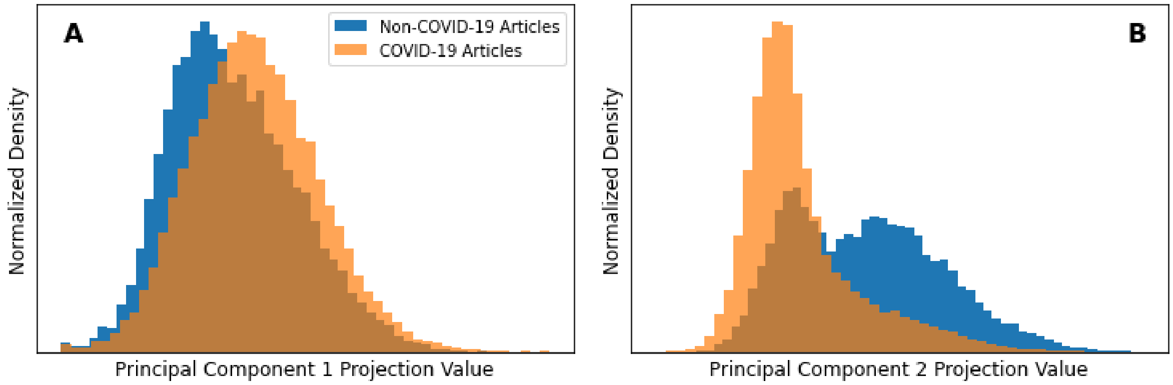


Figure SI.10.1. The distributions (y-axis) of non-COVID-19 abstracts (blue) and COVID-19 abstracts (orange) are plotted against the PC projection values (x-axis) in each panel. PC2 clearly separates the two groups.

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.

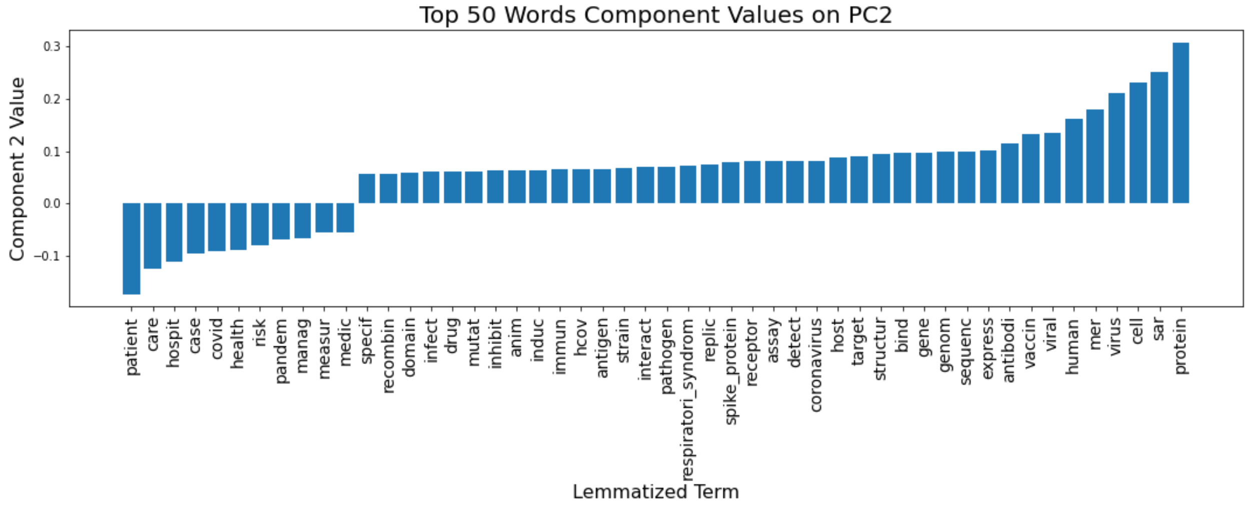


Figure SI.10.2. This bar chart displays the values of key lemmatized words on the second PC. We selected the 50 words associated with the largest component value magnitude for this plot to interpret the PC. The component values are represented by the y-axis and each individual word is plotted along the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

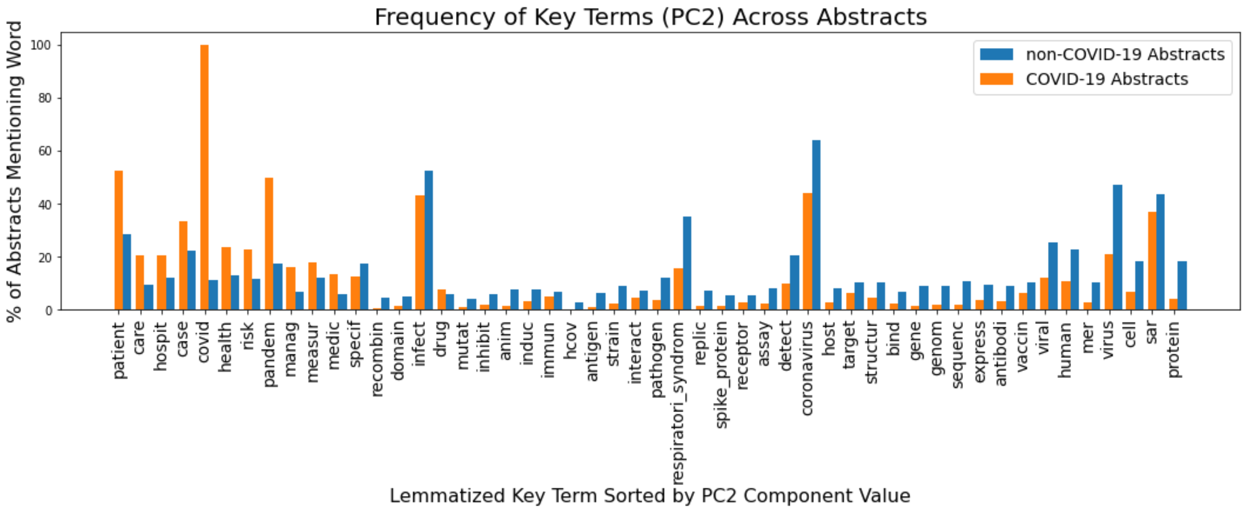


Figure SI.10.3. The top 50 key terms, selected by the magnitude of their component values on PC2, are unevenly distributed among the COVID -19 and non-COVID-19 abstracts. The proportion of abstracts in each group (orange for COVID-19 abstracts and blue for non-COVID-19 abstracts) mentioning each term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

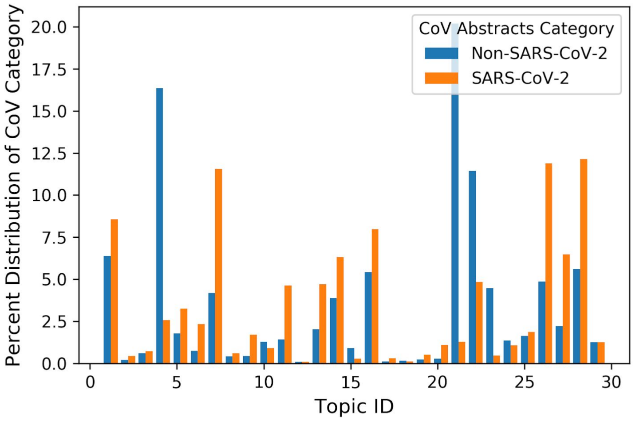


Figure SI.10.4. COVID-19 literature is distributed unevenly across the 30 topics identified via LDA. Topics identifiers (IDs) were assigned randomly by LDA. The percentage of abstracts in each of the two groups (orange for COVID-19 abstracts and blue for non-COVID-19 coronavirus abstracts) that are in 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.

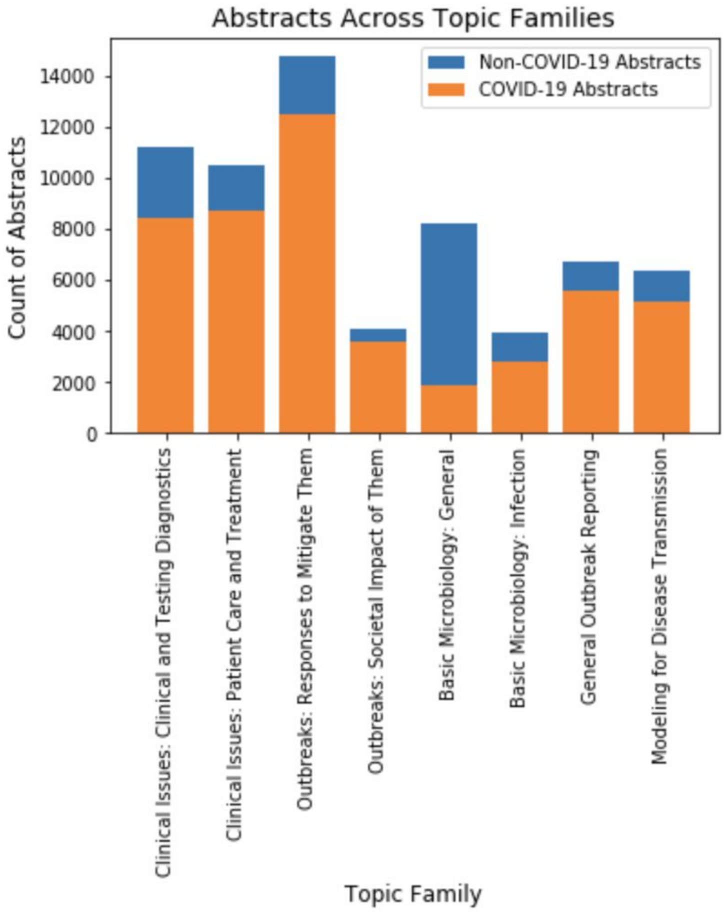
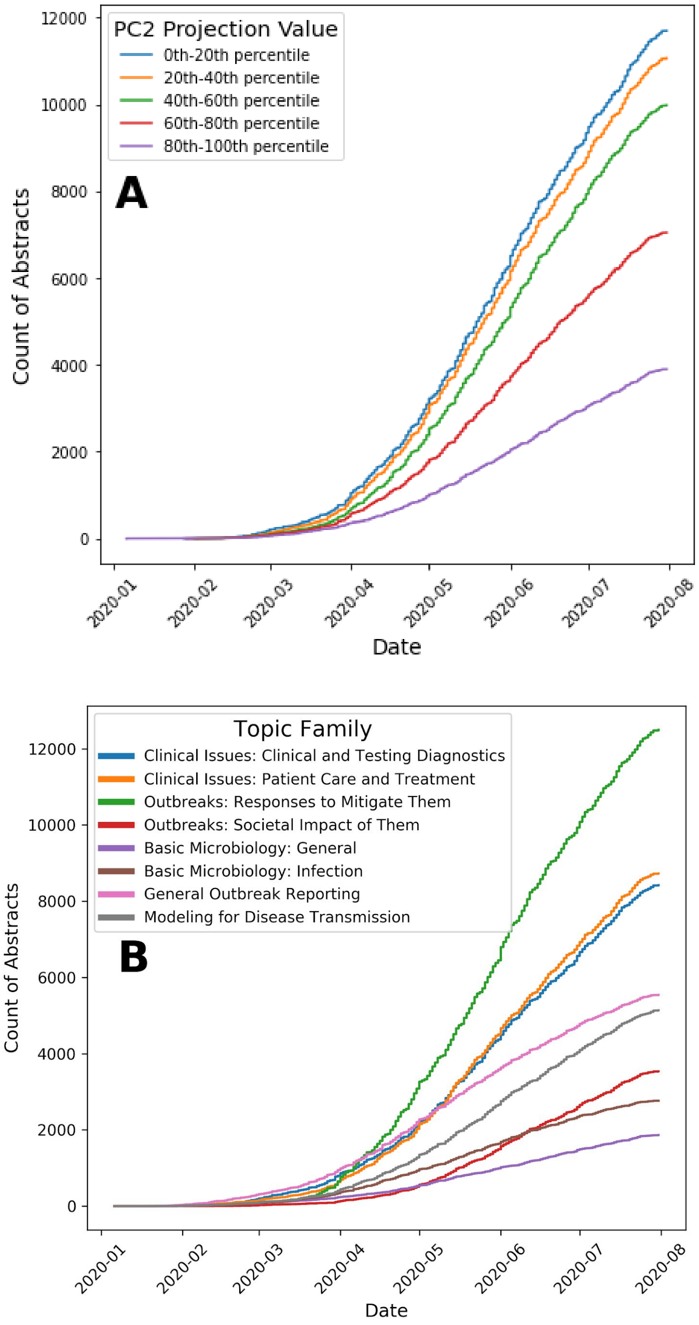


Figure SI.10.5. The distribution (y-axis) of abstracts of each category (blue for non-SARS-CoV-2 coronavirus abstracts and orange for SARS-CoV-2 abstracts) are plotted against each topic family (x-axis). Abstracts are unevenly distributed and SARS-CoV-2 abstracts tend to focus more on clinical, modeling, 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.

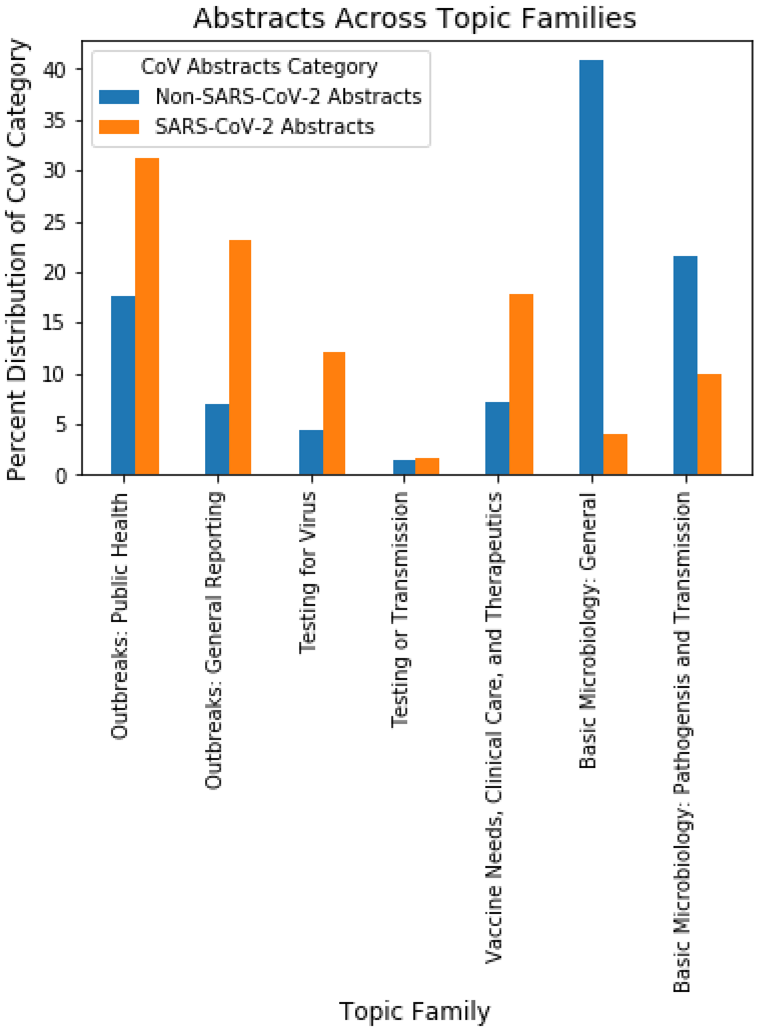


Figure SI.8.1. The distribution (y-axis) of abstracts of each category (blue for non-SARS-CoV-2 coronavirus abstracts and orange for SARS-CoV-2 abstracts) are plotted against each topic family (x-axis). Abstracts are unevenly distributed and SARS-CoV-2 abstracts tend to focus more on clinical, modeling, 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.