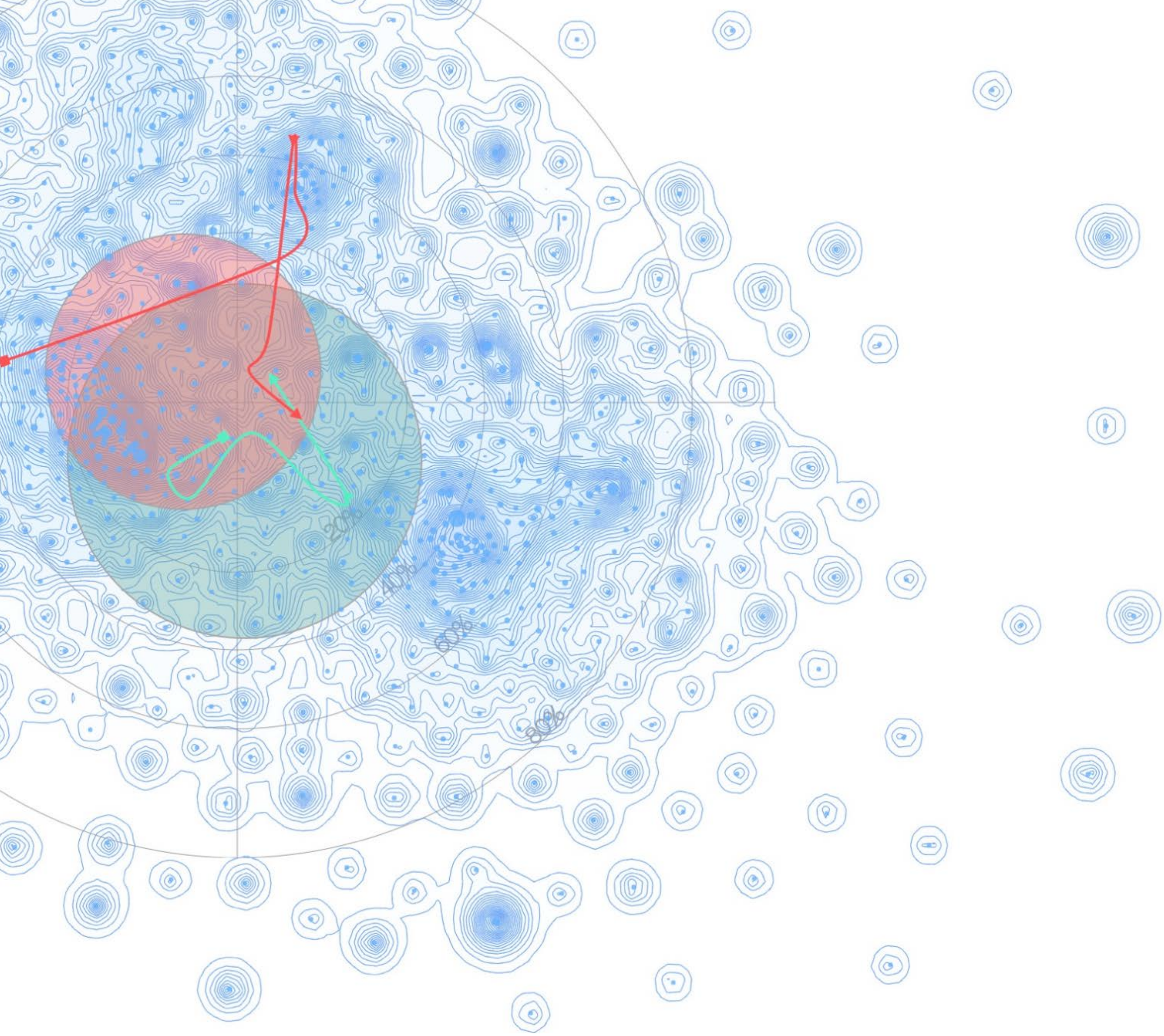

Insight on COVID-19 Research Directions

An Approach through Big Data Analytics

> The COVID-19 epidemic (also known as SARS-CoV-2) has impacted nearly every part of the world. Driven to help researchers further understand COVID-19, we aim to provide insights by exploring COVID-19 related public research documents using unique big data visualization methodologies.

The Research Goal

By aggregating the existing body of research on coronavirus, VALUENEX aims to make it easier for researchers to locate and understand new medical approaches for the rapidly changing world.

The Methodology

VALUENEX created data visualizations from large bodies of public research documents offering insight on the COVID-19 pandemic. The most popular areas of research were investigated, as well as growing areas and topics with relatively little prior attention.

The Results

Our data-driven methodology successfully captured the past and current trends of research in COVID-19 and points towards new directions in finding potential cures. Gauging data in previously under-exploited topics and drawing connections between highly relevant elements to COVID-19 allows researchers to gather information on existing literature quickly, while also guiding new possibilities for research.

The Dataset

There are several datasets tied to the following analyses, which center on drugs and biological mechanism related to COVID-19:

1. Dataset from semanticscholar.org [1] provides publicly available datasets with publications related to COVID-19. This dataset ("metadata") consists of publications from four primary sources: CZI, PMC, BioRxiv, and MedRxiv. After conducting some initial processing, including removing entries with empty abstracts and adding name normalization, there were 37,575 documents available for analysis.

2. 9,011 preprints (not peer-reviewed) from BioRxiv with some standard preprocessing.

The Analyses

With VALUENEX proprietary algorithms that incorporate unsupervised machine learning, high-dimensional visualizations and precision clustering, the VALUENEX Radar output allowed for easy viewing of the 37,575 COVID-19-related research documents from the meta dataset. This was then grouped into 6,831 clusters based on full-text semantic similarities and was precisely plotted on a single contour map (see Figure 1). Distance between elements corresponds to the magnitude of difference between the documents, and the contour lines highlight levels of document grouping and topic density.

The majority of articles used were published between January 1st, 2010 and March 27th, 2020. These addressed epidemics including, but not limited to, severe acute respiratory syndrome (SARS), Middle East Respiratory Syndrome (MERS), and Ebola, all of which have similar patterns or symptoms to COVID-19.

Figure 1 depicts a high-level overview of the current state of research and provides a structure to uncover new insights from the literature. Major Areas capture related information on different types of diseases similar to COVID-19, and Growing Areas capture research that focuses on COVID-19 situations starting in Wuhan. In contrast, Sparse Areas contain research directions that haven't received as much attention but are surrounded by high-density areas. After some initial analyses on the areas, including investigating each of these areas and identifying the key trends in this radar, further investigation in the following direction seems valuable: Drugs related to COVID-19 and Pathology of COVID-19.

RADAR TERMINOLOGY

The labels next to the areas show keywords from the documents within:

- **Major Area**: Documents in these areas have many similar documents nearby. Major Areas have a high volume of documents in a relatively small section of the radar
- **Growing Area**: Indicates faster growth than other regions, indicating that research here has rapidly developed in the past few years
- **Sparse Area**: a low-density area surrounded by high document density areas, which is a potentially underdeveloped or newly-emerging field.

The Big Picture • COVID-19 Metadata Analysis

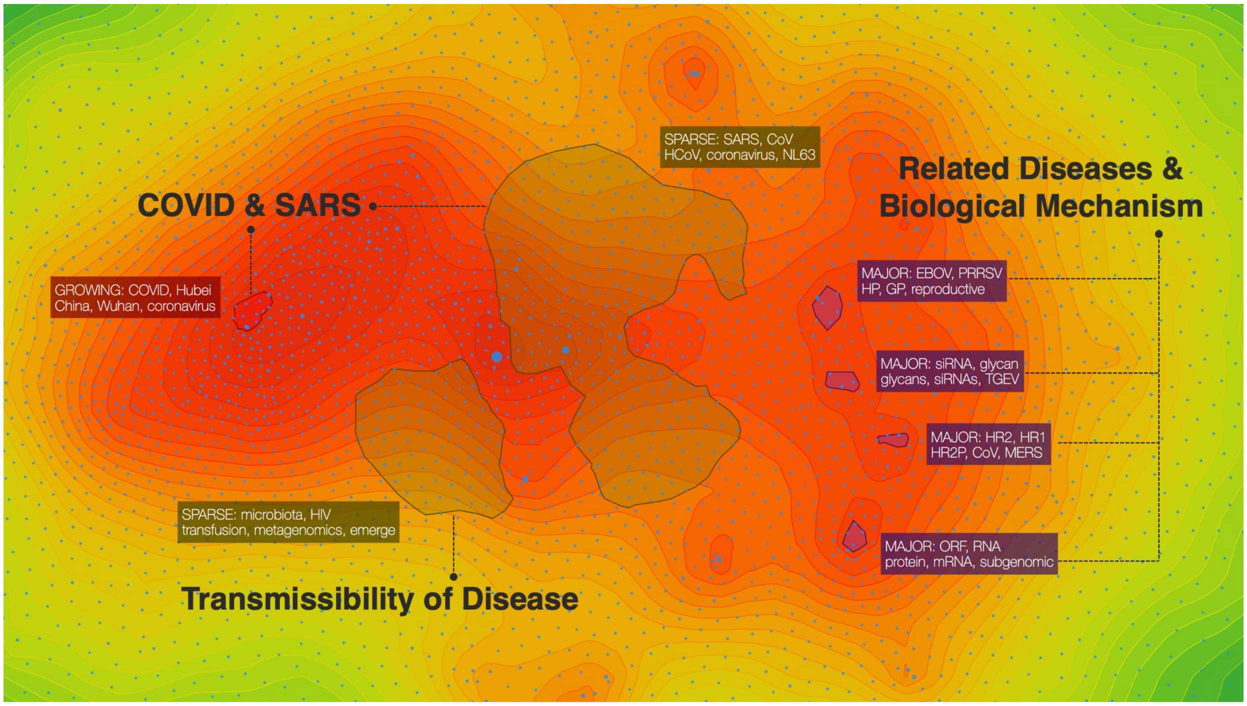

Figure 1: COVID-19 related research paper landscape (Metadata Radar)
[Link to Radar]

ROADMAP TO READING FIGURE 1

1. Begin by reading the Major Areas to understand where the higher density regions are and the topics most prominent in current research.
2. From here, identify the major research trends and direction. Most prominent are coronavirus and related respiratory diseases such as Middle East Respiratory Symptom (MERS) and Porcine Reproductive and Respiratory Syndrome Virus (PRRSV).
3. Then, find the fastest-growing region. Keywords include COVID and Wuhan and correspond to the growing academic shift towards the current COVID-19 epidemic.
4. Lastly, focus on sparse areas, the lower density areas surrounded by higher density regions. These areas typically contain information that bridges the gap among the major areas and suggests potentially important insights from topics with relatively little prior attention.

Figure 1 illustrates a sparse area directly related to SARS and Human Coronavirus NL 63 (HCoV), as well as one that focuses on anthroponosis and zoonosis-- how diseases are transmitted between humans and other non-human animals.

COVID-19 Sub-radar Analysis

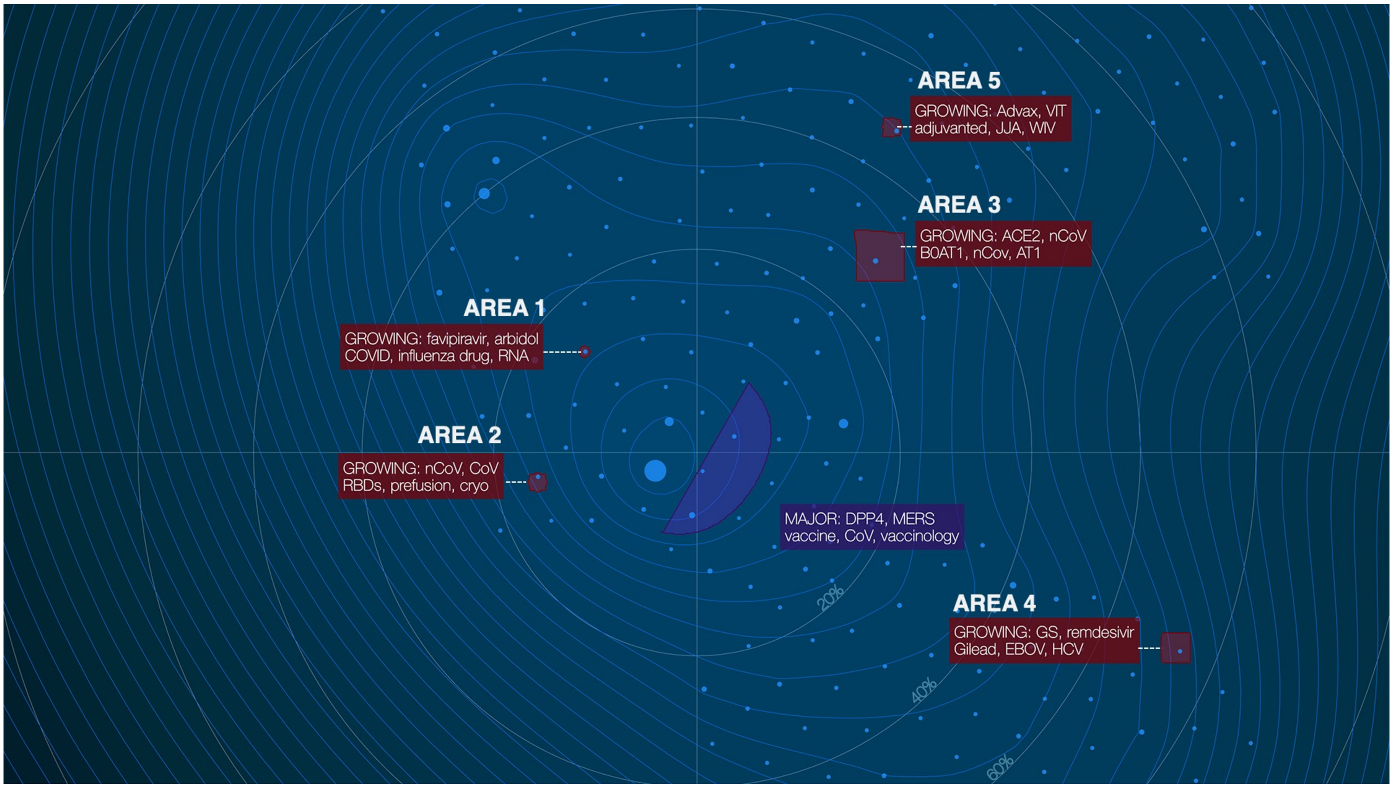

Figure 2: SARS-nCoV Sub-radar from Metadata
[Link to Radar]

ROADMAP TO READING FIGURE 2

A sub-radar is created from an area of interest in the original larger radar to generate more in-depth explorations of a specific region.

1. Start by reading the Major Area to understand the most prominent high-density regions.
2. Next, focus on the labeled areas in sequential order, taking note of the associated keywords.

THE FINDINGS

The sub-radar shown in Figure 2 (2,554 documents) was generated from the Sparse Area of the COVID-19 metadata radar (Figure 1). It emphasizes the keywords SARS, CoV, HCoV, coronavirus, and NL63 (a type of human coronavirus), where the Sparse Area suggested that there was previously little information.

The Major Area shows keywords that are directly relevant to COVID, such as MERS and vaccine; additional keyword Dipeptidyl Peptidase 4 Inhibitor (DPP4) was also notable and is described in more detail in the analysis of Figure 5. The most noteworthy regions of this sub-radar are the growing areas, which have seen the highest rate of growth over the past 5 years compared to other areas. In this radar, the growing areas hold research on the most recent approaches to treating coronavirus.

AREA 1 in the sub-radar above consists of ‘favipiravir & other influenza drugs’. Favipiravir is a Japanese anti-viral drug that has demonstrated potential in treating COVID-19, showing better therapeutic responses in terms of disease digression and viral clearance than some other treatments [2].

AREAS 2 & 3 include two important keywords: ‘angiotensin-converting enzyme 2’ (ACE2) and ‘receptor-binding domain’ (RBD). Coronavirus enters human cells through the RBD of a glycoprotein (GP) called SARS-CoV-2 spike. This GP directly binds to the human cell membrane protein ACE2 [3]. As these two areas are growing regions, researchers can use them to better understand the work being done to inhibit such binding for coronavirus treatments.

AREA 4 highlights ‘Remdesivir’, a drug developed by Gilead Sciences to treat Ebola Virus. In the initial period of COVID-19 epidemic, remdesivir showed some potential in becoming an effective treatment. However, Gilead Sciences recently decided to halt its COVID-trial in China, as the drug had not shown to be effective against coronavirus. [4]

AREA 5 pivots direction, where the main keyword ‘Advax adjuvant’ represents a novel human adjuvant that enhances both humoral and cellular immunity. Literature in this area focus on immunology, which could be a potential area for developing coronavirus vaccines.

THE IMPLICATIONS

This sub-radar has shown promising results in summarizing some of the most significant progress and directions that have been made to find cures and solutions for COVID-19. We believe this sub-radar can provide insights for researchers on existing literature, while also giving potential directions to research. The sub radar and meta-data radar open possibilities for research and efficiently allow researchers to find the most relevant work.

Interferons and Pro-inflammatory Cytokines

Further analysis of the sparse area within the filtered version of Metadata radar show two major areas – both related to the pathology of CoV (Figure 1). Articles that do not contain COVID-19 and related respiratory diseases were filtered out, resulting in a total of 516 documents with 122 clusters.

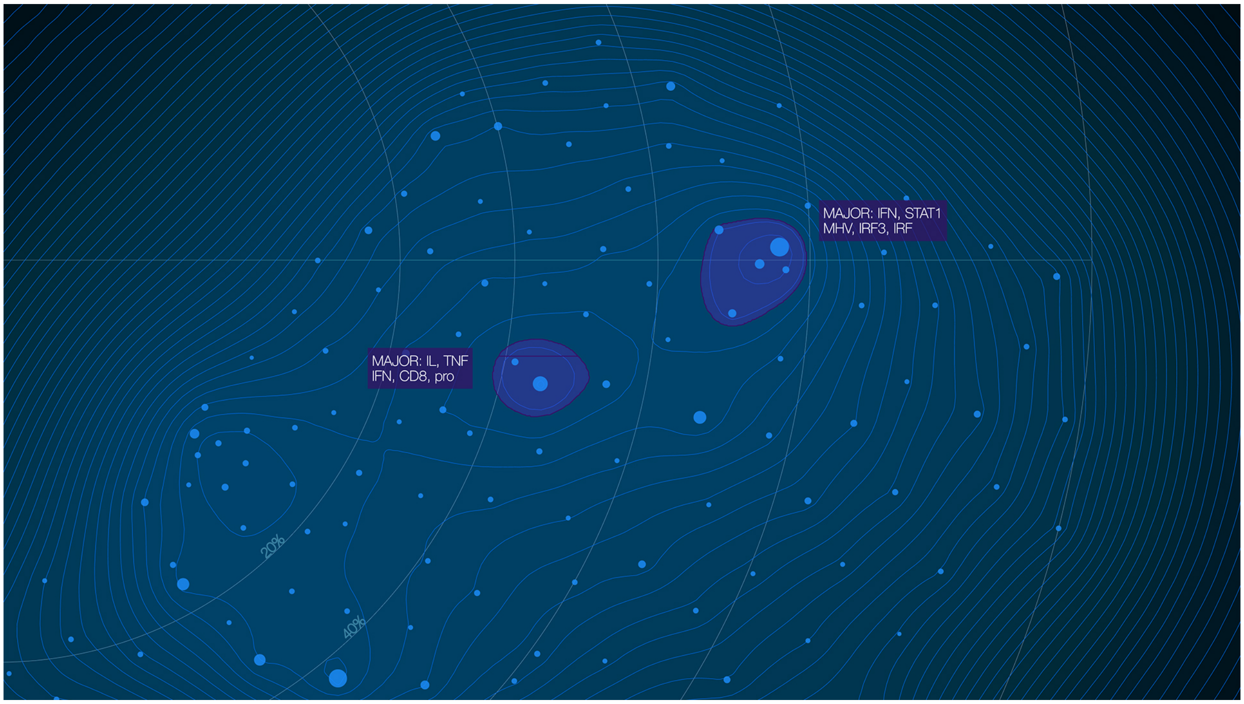

Figure 3: Filtered Metadata Sparse Area Sub-radar
[Link to Radar]

ROADMAP TO READING FIGURE 3

This sub-radar was created by isolating the sparse area in the filtered metadata set.

1. Start by reading the Major Areas to understand the topics of higher density regions and Sparse Areas. The keywords in these Major Areas include IFN, STAT1, IL, TNF, and CD8, all of which are biological components related to immunity.
2. In this sub-radar, there are two Major Areas: the first refers to the area on the right and the second refers to the area on the left. Less prominent keywords in the latter include CD4, CNS (Central Nervous System), and inflammation.

THE FINDINGS

The first Major Area on the right defines the relationship between Interferons (IFNs) and STAT1 (Signal Transducer and Activator of Transcription 1). Specifically, IFN-gamma mediates inflammation and immunity and activates STAT1, a cytoplasmic protein designed to defend the host. Together, IFNs and STAT1 are typically associated with antiviral defense and tumor suppression. However, the Murine Coronavirus Mouse Hepatitis Virus (MHV), another keyword in this Major Area, has a crucial role in circumventing this signaling pathway and thus in suppressing the immune response.

The second Major Area on the left is related to cytokines, signals that can either inhibit or encourage inflammatory responses. Interleukins (IL) and Tumor Necrosis Factors (TNF) are important cytokines in the signaling of components in the adaptive immune response, such as CD4/8 T-cells. A noteworthy paper found in this Major Area shows that IL-38, a critical therapeutic cytokine, may be able to inhibit inflammation in diseases like COVID-19 [6].

THE IMPLICATIONS

1. Given the relationship between IFNs and STAT1 in the innate immune response, this could be explored as a target for a treatment drug for CoV virus.
2. Understanding how the Central Nervous System (based on the second Major Area) might be impacted by COVID-19 could be an area for further study, given its appearance in the Sparse Area sub-radar (Figure 3).
3. Exploring IL and TNF signaling relationships can provide better insights into what factors can either inhibit or relieve symptoms of COVID-19. IL-38, an anti-inflammatory cytokine that can inhibit the inflammatory response to COVID-19, for example, was found in a research paper of the second Major Area [6], and further research on this topic can be novel and impactful.

COVID-19 Infection Mechanism

ACE2 was a keyword that appeared throughout the metadata radar, suggesting that it may play a role in COVID pathogenesis. Research has shown that ACE2 is the 'entry-point' for coronavirus [7]. A sub-radar was created from the filtered metadata radar to better understand ACE2 (Figure 4).

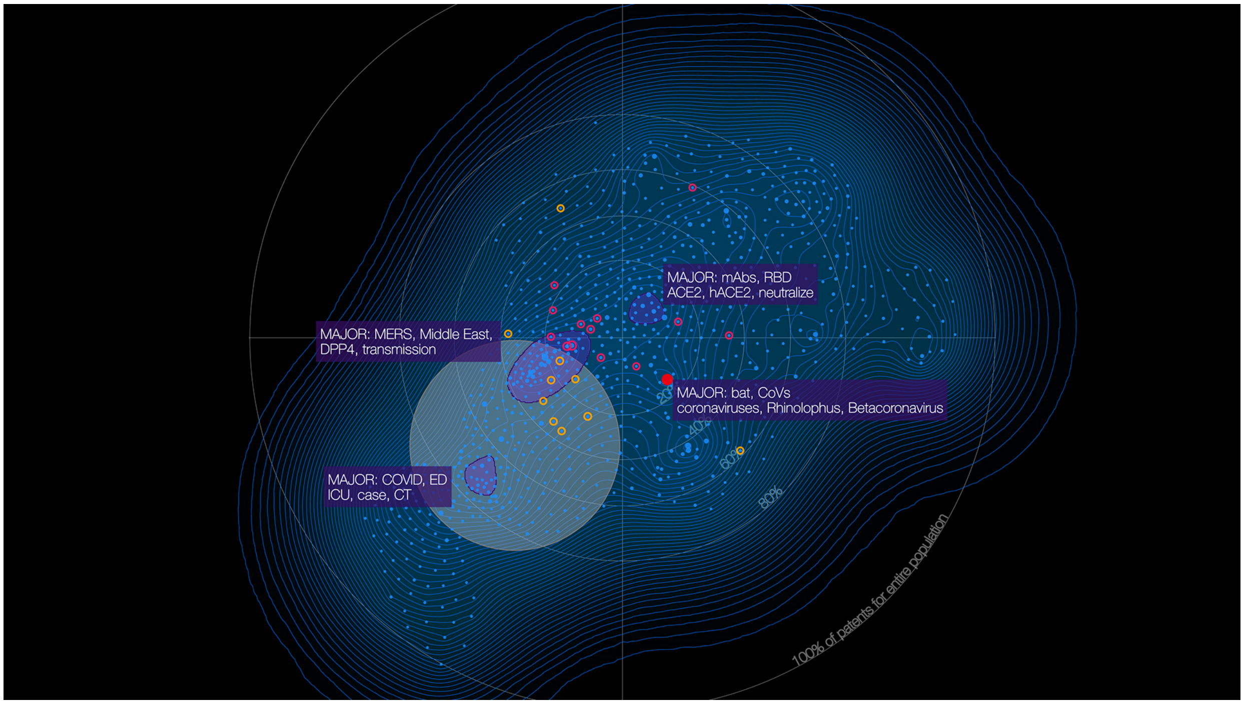

Figure 4: ACE-2 Sub-radar
[Link to Radar]

ROADMAP ON READING FIGURE 4

This sub-radar was created by filtering the metadata set for the 'ACE-2' keyword.

Read the Major Areas. Note that (from left to right), there is one Major Area related to hospital care (Keywords: ICU, CT), one related to MERS (Keywords: Middle East), one related to mAbs (a journal), and one related to bat transmission of coronavirus. The last area contains a very large, concentrated cluster, marked in red.

THE FINDINGS

A very large, concentrated cluster represents literature on the anthroponosis and zoonotic origin of coronavirus — particularly, the transmission between bats and humans (Figure 5, marked in red). The clusters on the left side are related to MERS and COVID cases, and how ACE2 is targeted in both.

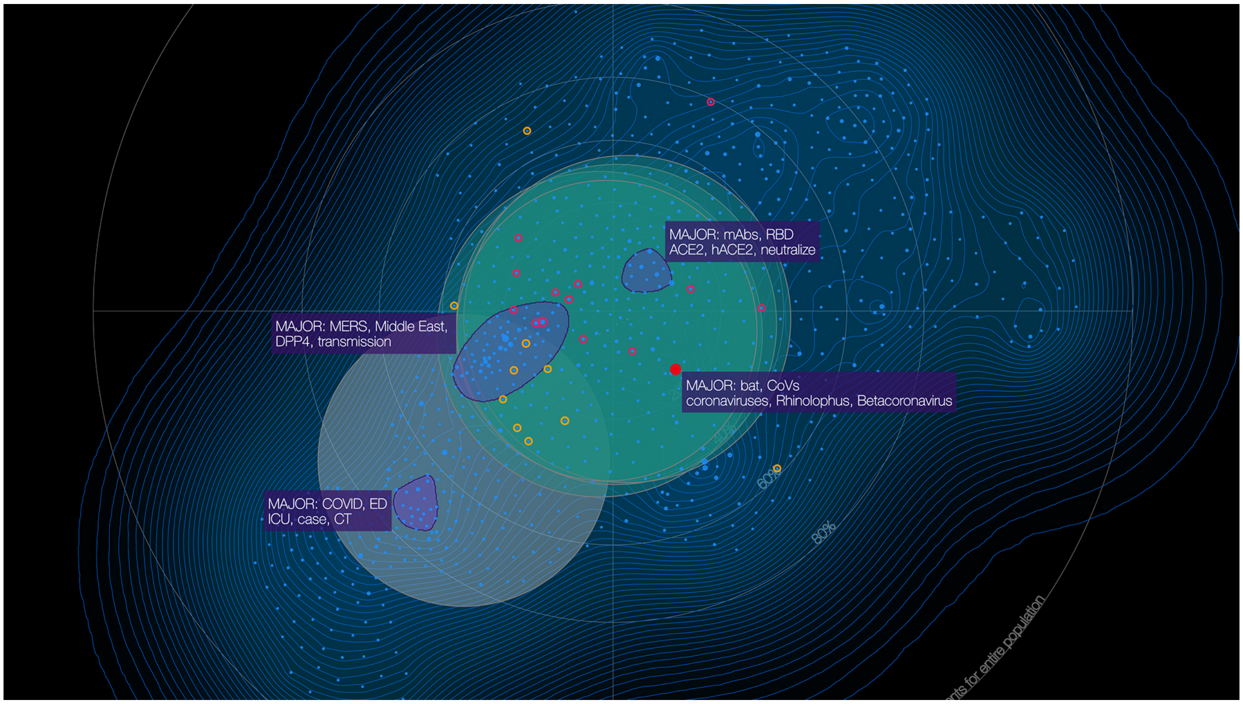

Figure 5: ACE-2 Sub-radar: 2020 Research Center of Gravity
[Link to Radar]

ROADMAP TO READING FIGURE 5

This sub-radar was created by isolating 'ACE-2' keyword and by creating a new sub-radar from the filtered metadata set.

1. Identify the Major Areas in this radar. The colored circles correspond to centers of gravity according to different time frames: the gray circle represents the center of gravity of research done in 2020, while the green circle represents the center of gravity of research done from 2016-2019. Given that the center of gravity of research done in 2020 is around the clusters on the left side of the radar, these were analyzed more deeply.

NOTE: Center of Gravity shows the breadth of a year's core coverage calculated by document ratio and relative distance. In this case, the center of gravity for 2020 shows where most of the research papers in 2020 are located and what topics they concentrated on.

2. Clusters containing keywords with DPP4 are marked in pink, while clusters with keywords including ribavirin are marked in orange. DPP4 is a tumor suppressor gene related to prostate cancer with the suppression of IL-1 and IL-6, and ribavirin is a hep C drug that treats severe respiratory/lung infections. We can see that these topics are close to the 2020 center of gravity. These topics were chosen by searching for keywords related to biological mechanisms.

THE FINDINGS

It is clear that articles on the diagnosis and hospital treatment of COVID-19 are markedly different than articles in the past few years: the keywords between the gray and green circles have little overlap, and the Centers of Gravity are distinct when comparing 2020 and 2016-2019. Moreover, articles related to mechanisms of biological pathogenesis, like CD4/8, IL, have a center of gravity and distribution on the upper right region of the radar. As COVID-19 emerged in 2020 there aren’t many pathology-related keywords in that area, keywords with 2020 Centers of Gravity adjacent to COVID keywords might be of interest in exploring pathology. Specifically, these keywords are DPP4 and ribavirin.

THE IMPLICATIONS

DPP4 is a tumor suppressor gene related to prostate cancer with the suppression of IL-1 and IL-6, and ribavirin is a hep C drug that treats severe respiratory/lung infections. ribavirin is often combined with IFN medicine. DPP4 and ribavirin may provide novel avenues for research, especially given the connection to IL/IFN pathology.

COVID Keywords Analysis

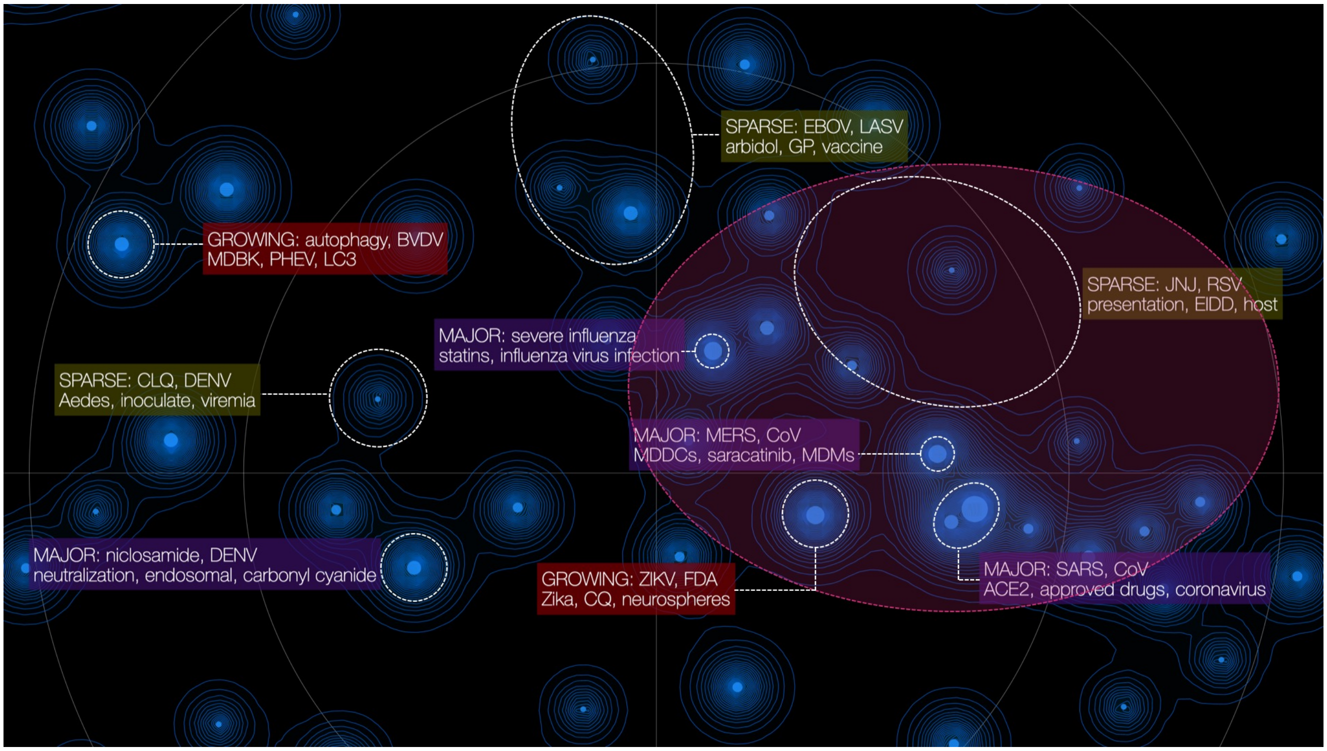

Figure 6: COVID Keywords Radar
[Link to Radar]

ROADMAP TO READING FIGURE 6

1. Start by noting the keywords for the Major Areas, the Growing Areas, and the Sparse Area sequentially.
2. Then, focus on the keywords and areas inside the pink enclosure, as most of the insights are concentrated there.

THE FINDINGS

This radar (194 documents) was created by filtering the metadata set (Figure 1) with the keyword list ("favipiravir", "chloroquine", "hydroxychloroquine", "azithromycin", "chlorpromazine", "niclosamide", "toremifine", "saracatanib", "gemcitabine").

These keywords are either drugs that have demonstrated some effectiveness against COVID-19, or drugs/substances that have been researched for their potential.

We find the area inside the pink enclosure to be of particular interest, since the distances between the Major, Growing, and Sparse Areas are relatively small. It is important to note that small distances between areas mean that these areas are more similar and potentially related. In this area, similar diseases are being situated together, with keywords capturing some of the most important information in COVID-19.

For example, one particularly insightful finding in the sparse area is EIDD (Emory Institute Drug Development), where EIDD-2801 was discovered. This is a drug developed before the COVID-19 pandemic in order to treat respiratory diseases. It has recently been shown (04/06/2020) that EIDD-2801 shows efficiency against COVID-19 in human/mice bodies. [5]

THE IMPLICATIONS

Keywords on the radar are a good starting point for researchers to orient their search in specific directions. Future approaches should focus on the keywords given in the radar in order to quickly synthesize information on the existing research directions on COVID-19.

> Our data-driven methodology allows researchers to gather information quickly, evaluate new insights, and find directions to investigate potential cures and vaccines. The radars captured current research trends on COVID-19 and revealed new directions for researchers to explore, such as the DPP4, IL-38, and EIDD-2801. Creating sub-radars can enable researchers to explore the literature in more detail, e.g., investigating Sparse Areas or filtering ACE-2 keyword to reveal insights on the biological mechanisms in COVID-19. We recommend further exploring radar areas that have received less prior-attention (e.g., filtering the dataset by a keyword of interest and creating a sub-radar to analyze) to find unexpected usage and connections for new ideas and pathways.

References

1. *CORD-19 | Semantic Scholar*. Semanticscholar.org. (2020). Retrieved from https://www.semanticscholar.org/cord19.

2. Cai, Q., Yang, M., Liu, D., Chen, J., Shu, D., & Xia, J. et al. (2020). Experimental Treatment with Favipiravir for COVID-19: An Open-Label Control Study. *Engineering*. https://doi.org/10.1016/j.eng.2020.03.007

3. Bergan, B. (2020). *For the First Time, Scientists Found How Coronavirus Infects Human Cells*. Interestingengineering.com. Retrieved from https://interestingengineering.com/for-the-first-time-scientists-found-how-coronavirus-infects-human-cells.

4. *Gilead shares slip as a 2nd remdesivir COVID-19 trial halted in China*. FierceBiotech. (2020). Retrieved from https://www.fiercebiotech.com/biotech/gilead-shares-slip-as-a-second-remdesivir-covid-19-trial-halted-china.

5. Sheahan, T., Sims, A., Zhou, S., Graham, R., Pruijssers, A., & Agostini, M. et al. (2020). An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Science Translational Medicine*, *12*(541), eabb5883. https://doi.org/10.1126/scitranslmed.abb5883

6. Conti P, Ronconi G, Caraffa A, et al. (2020). Induction of pro-inflammatory cytokines (IL-1 and IL-6) and lung inflammation by Coronavirus-19 (COVI-19 or SARS-CoV-2): anti-inflammatory strategies*. Journal of Biological Regulators and Homeostatic Agents, 34*(2). DOI: 10.23812/conti-e.

7. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C., Abiona, O., . . . Mclellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science, 367(6483), 1260-1263. doi:10.1126/science.abb2507

8. Khodarev, Nikolai N. , Bernard Roizman, and Ralph R. Weichselbaum. "Molecular Pathways: Interferon/Stat1 Pathway: Role in the Tumor Resistance to Genotoxic Stress and Aggressive Growth." Clinical Cancer Research (2012). doi: 10.1158/1078-0432

9. Hu, Xiaoyu, and Lionel B Ivashkiv. “Cross-regulation of signaling pathways by interferon-gamma: implications for immune responses and autoimmune diseases.” Immunity vol. 31,4 (2009): 539-50. doi:10.1016/j.immuni.2009.09.002