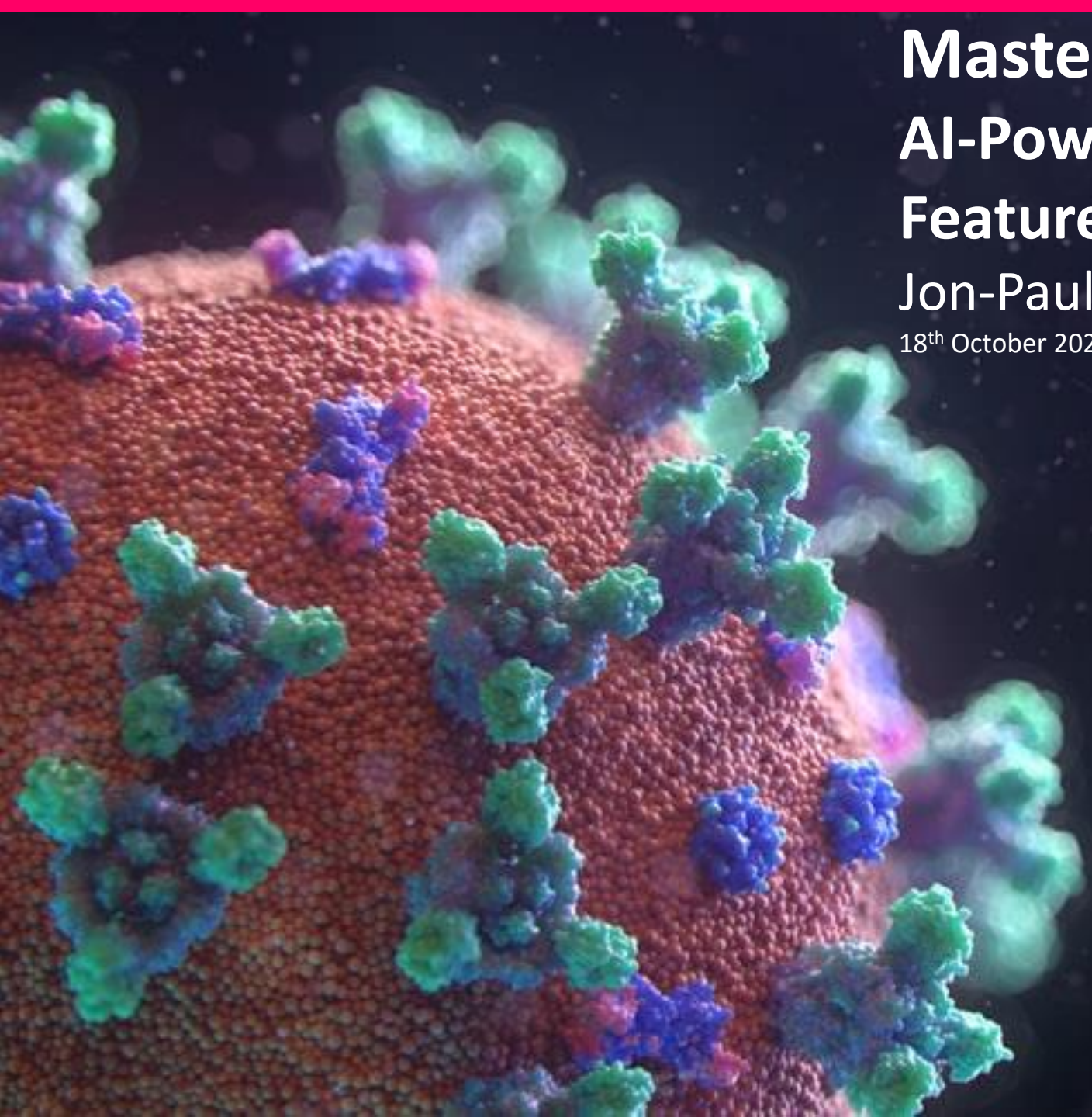


# Masters Thesis

## AI-Powered COVID-19 Research Cockpit

## Feature & Concept Prototyping

Jon-Paul Boyd

18th October 2020

DE MONTFORT UNIVERSITY LEICESTER

# Overview Feature & Concept Prototyping

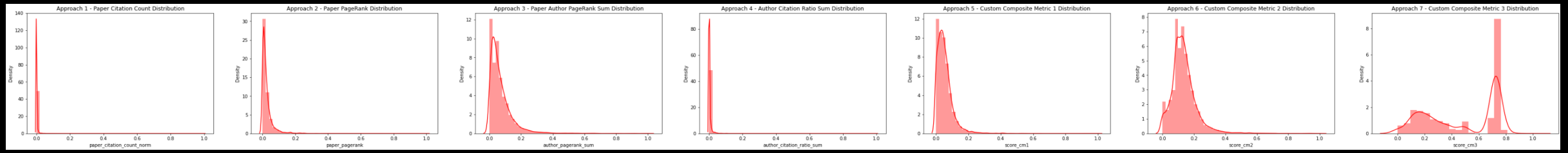
**Assessing viability of a cockpit empowering COVID-19 medical research using pillars of A.I.** Essential features including information extraction, literature clustering, search & summarization are prototyped in Jupyter notebooks, with results provided.

# Paper Ranking Metrics With Author Influence & Paper Citations

**Several custom metrics enable paper importance weighting,** formulated using paper author & citation counts/ratios, the PageRank algorithm applied to assembled network graphs of authors & papers. Such metrics can weight paper scores in search.

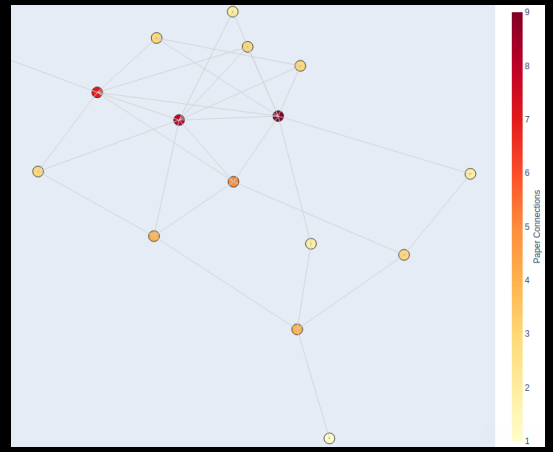
| | hash_title | paper_id | title | journal | authors | publish_year | publish_time | author_count | referenced_papers_count | paper_citation_count | paper_citation_count_norm | paper_pagerank | author_pagerank_sum | author_citation_ratio_sum | score_cm1 | score_cm2 | score_cm3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **69155** | 994a7643aa15553758a0b294580378ccb66d2f97 | PMC7122603 | Cardiovascular Activity | Drug Discovery and Evaluation | Vogel, Hans Gerhard | 2008 | 2008 | 1 | 3418.0 | 0.0 | 0.000000 | 1.000000 | 0.000524 | 0.000000 | 1.000000 | 0.871312 | 0.591382 |
| **70139** | 73aadaed0025eb62d9485ae9009e7b304f15a034 | PMC7339753 | Non-neoplastic diseases of the testis | Urologic Surgical Pathology | Nistal, Manuel; Paniagua, Ricardo | 2020 | 2020-06-22 | 2 | 1621.0 | 0.0 | 0.000000 | 0.521681 | 0.001890 | 0.000000 | 0.521681 | 0.535554 | 1.000000 |
| **27876** | ff35497603df24d07b6794f79b937b8e04055417 | PMC7168572 | Analysis of carbohydrates and glycoconjugates ... | Mass Spectrom Rev | Harvey, David J. | 2014 | 2014-05-26 | 1 | 1549.0 | 0.0 | 0.000000 | 0.497535 | 0.000330 | 0.000723 | 0.497535 | 0.450013 | 0.435206 |
| **2030** | 36a587842805886ddf7b6ca528a9a6838111a9a2 | PMC7427299 | Geographical and temporal distribution of SARS... | Euro Surveill | Alm, Erik; Broberg, Eeva K; Connor, Thomas; Ho... | 2020 | 2020-08-13 | 586 | 12.0 | 1.0 | 0.000287 | 0.001556 | 1.000000 | 0.133386 | 0.456102 | 1.000000 | 0.871627 |
| **26962** | 4200b0950b4371fa219305c8162613bc7d042f15 | PMC7173518 | Microorganisms Responsible for Neonatal Diarrhea | Infectious Diseases of the Fetus and Newborn I... | O'Ryan, Miguel L.; Nataro, James P.; Cleary, T... | 2010 | 2010-12-27 | 3 | 1291.0 | 0.0 | 0.000000 | 0.430953 | 0.003646 | 0.000047 | 0.432610 | 0.388158 | 0.335000 |
| **26963** | 4200b0950b4371fa219305c8162613bc7d042f15 | PMC7173613 | Microorganisms Responsible for Neonatal Diarrhea | Infectious Diseases of the Fetus and Newborn | O'Ryan, Miguel L.; Nataro, James P.; Cleary, T... | 2009 | 2009-05-19 | 3 | 1259.0 | 0.0 | 0.000000 | 0.430953 | 0.003646 | 0.000047 | 0.432610 | 0.386821 | 0.323587 |
| **26964** | 4200b0950b4371fa219305c8162613bc7d042f15 | PMC7173613 | Microorganisms Responsible for Neonatal Diarrhea | Infectious Diseases of the Fetus and Newborn I... | O'Ryan, Miguel L.; Nataro, James P.; Cleary, T... | 2009 | 2009-05-19 | 3 | 1259.0 | 0.0 | 0.000000 | 0.430953 | 0.003646 | 0.000047 | 0.432610 | 0.386821 | 0.323587 |
| **26961** | 4200b0950b4371fa219305c8162613bc7d042f15 | PMC7173518 | Microorganisms Responsible for Neonatal Diarrhea | Infectious Diseases of the Fetus and Newborn | O'Ryan, Miguel L.; Nataro, James P.; Cleary, T... | 2010 | 2010-12-27 | 3 | 1291.0 | 0.0 | 0.000000 | 0.430953 | 0.003646 | 0.000047 | 0.432610 | 0.388158 | 0.335000 |
| **42264** | 12675ba5050c3d59787def31dffdeef30fbd3ff4 | PMC7158344 | Disorders of the Gastrointestinal System | Equine Internal Medicine | Sanchez, L. Chris | 2017 | 2017-11-17 | 1 | 1401.0 | 0.0 | 0.000000 | 0.404194 | 0.000524 | 0.000187 | 0.404194 | 0.382515 | 0.498006 |
| **14967** | 9304e520b92a70f6ba4e57124589ceece9369c33 | PMC7223859 | 2019 HRS/EHRA/APHRS/LAHRS expert consensus sta... | J Interv Card Electrophysiol | Cronin, Edmond M.; Bogun, Frank M.; Maury, Phi... | 2020 | 2020-01-27 | 38 | 1145.0 | 0.0 | 0.000000 | 0.368848 | 0.058191 | 0.000914 | 0.395298 | 0.457257 | 0.931831 |

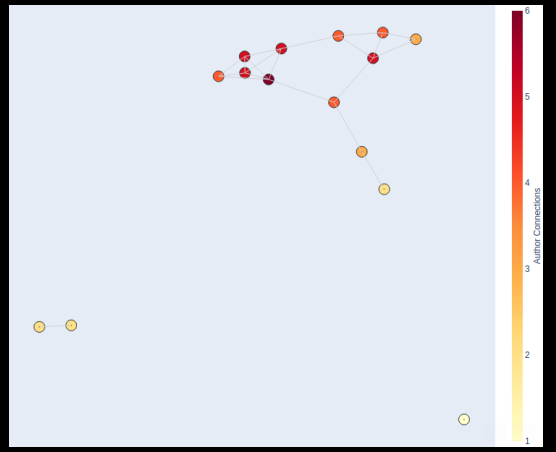
**Distribution of several ranking approaches across corpus,** can be used to fine-tune metrics




**Subset of paper & author network,** deeper colours indicate greater influence

Papers

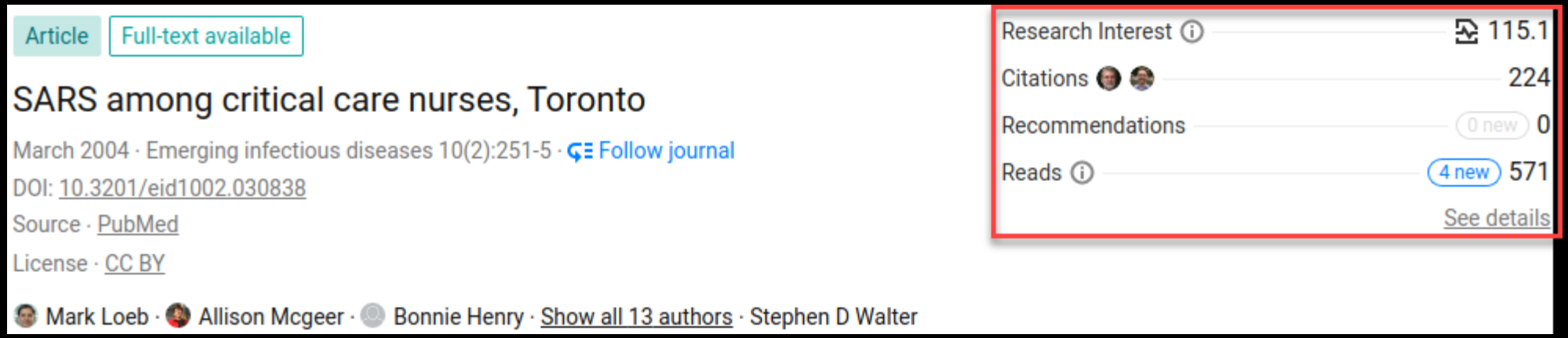



Authors

# Enriching CORD-19 Dataset ResearchGate Web Scraping

**Sourcing useful additional article data** from 3rd-party sources such as ResearchGate, which indicates each paper's research interest, citation count, recommendations & number of reads. Can be integrated into scoring metrics and filtering.




**Web page data scraped** and added to article metadata.

| title | doi | rg_stats_research_interest | rg_stats_citations | rg_stats_reads |
|---|---|---|---|---|
| Thiopurine analogues inhibit papain-like prote... | 10.1016/j.bcp.2008.01.005 | 24.9 | 47.0 | 153.0 |
| SARS among Critical Care Nurses, Toronto | 10.3201/eid1002.030838 | 115.1 | 224.0 | 571.0 |
| Short fractionation radiotherapy for early pro... | 10.1007/s11547-020-01216-9 | 2.2 | 2.0 | 35.0 |
| In vivo and in vitro Models of Demyelinating D... | 10.1159/000149316 | 6.5 | 13.0 | 5.0 |
| The role of high load herpes simplex virus in ... | 10.1186/s13054-020-2815-9 | 0.5 | 0.0 | 6.0 |

# Enriching CORD-19 Dataset Adding Clinical Trial Information

**Sourcing useful clinical trial data** including study type & design, trial phase, intervention, sponsor and outcome.

| id | url | study_type | study_design | intervention | sponsor | outcome | phase |
|---|---|---|---|---|---|---|---|
| NCT04400682 | https://ClinicalTrials.gov/show/NCT04400682 | Interventional | Allocation: Randomized\|Intervention Model: Cro... | Drug: FAVIRA 200 MG Film Tablet\|Drug: AVIGAN 2... | Novelfarma Ilaç San. ve Tic. Ltd. Sti.\|Novagen... | AUC0-tlast\|Favipiravir Cmax\|AUC0-inf of Favipi... | Phase 1 |
| NCT04502368 | https://ClinicalTrials.gov/show/NCT04502368 | Observational | Observational Model: Case-Only\|Time Perspectiv... | Procedure: Fiberoptic Bronchoscopy (FOB)\|Proce... | Erasme University Hospital | Regional Compliance Variation\|Regional Resista... | NaN |
| NCT04339842 | https://ClinicalTrials.gov/show/NCT04339842 | Observational | Observational Model: Case-Only\|Time Perspectiv... | Other: Assessment of Dietary Changes in Adults... | Eliz Arter\|Cyprus Science University\|Eastern M... | Changes in the Eating Habits of Adults during ... | NaN |
| NCT04476719 | https://ClinicalTrials.gov/show/NCT04476719 | Interventional | Allocation: Randomized\|Intervention Model: Cro... | Drug: ATAFENOVIR 200 MG KAPSUL\|Drug: ARBIDOL 1... | Atabay Kimya Sanayi Ticaret A.S.\|Novagenix Bio... | Primary PK Endpoint\|Secondary PK Endpoint | Phase 1 |
| NCT04407000 | https://ClinicalTrials.gov/show/NCT04407000 | Interventional | Allocation: Randomized\|Intervention Model: Cro... | Drug: Test: Favipiravir 200 mg (LOQULAR)\|Drug:... | World Medicine ILAC SAN. ve TIC. A.S.\|Novageni... | Primary PK End Points AUC0-tlast\|Primary PK En... | Phase 1 |
| NCT04417153 | https://ClinicalTrials.gov/show/NCT04417153 | Observational | Observational Model: Cohort\|Time Perspective: ... | Behavioral: Mindfulness Based Intervention | National University, Singapore\|Potential proje... | Change in Subjective measures of Sleep quality... | NaN |
| NCT03871491 | https://ClinicalTrials.gov/show/NCT03871491 | Interventional | Allocation: Randomized\|Intervention Model: Par... | Drug: Azithromycin\|Drug: Placebo | NICHD Global Network for Women's and Children'... | Maternal: Incidence of maternal death or sepsi... | Phase 3 |
| NCT04386876 | https://ClinicalTrials.gov/show/NCT04386876 | Interventional | Allocation: Randomized\|Intervention Model: Cro... | Drug: Lopinavir 200Mg/Ritonavir 50Mg FT Test\|D... | World Medicine ILAC SAN. ve TIC. A.S.\|Novageni... | Primary PK End Points | Phase 1 |
| NCT04276987 | https://ClinicalTrials.gov/show/NCT04276987 | Interventional | Allocation: N/A\|Intervention Model: Single Gro... | Biological: MSCs-derived exosomes | Ruijin Hospital\|Shanghai Public Health Clinica... | Adverse reaction (AE) and severe adverse react... | Phase 1 |
| NCT03474965 | https://ClinicalTrials.gov/show/NCT03474965 | Interventional | Allocation: N/A\|Intervention Model: Single Gro... | Drug: Crizanlizumab | Novartis Pharmaceuticals\|Novartis | PK (AUCd15) after 1st dose\|PD (AUCd15) after 1... | Phase 2 |

**Multiple trials per paper** is managed.

| | paper_id | trials | trial_url |
|---|---|---|---|
| 646 | PMC7252014 | [NCT04317092, NCT04320615, NCT04306705, NCT043... | [https://ClinicalTrials.gov/show/NCT04317092, ... |
| 647 | PMC7377794 | [NCT04328285, NCT04328441] | [https://ClinicalTrials.gov/show/NCT04328285, ... |
| 648 | PMC7295303 | [NCT04303507, NCT04308668] | [https://ClinicalTrials.gov/show/NCT04303507, ... |
| 649 | PMC7329292 | [ChiCTR2000029765, NCT04320615, ChiCTR20000297... | [http://www.chictr.org.cn/showproj.aspx?proj=4... |
| 650 | PMC7425928 | [NCT04347993, NCT04331808] | [https://ClinicalTrials.gov/show/NCT04347993, ... |

# Named Entity Recognition For Entity Classification

**Used to classify text entities into categories.** **U**seful in providing a summary overview of the article, clustering documents by entities, used as selectable facets in document filtering etc.

The examples below show the same article abstract processed through 2 different NER models.

PMC7189851 Candidate drugs against SARS-CoV-2 and COVID-19

Model 1

Abstract Outbreak and pandemic of coronavirus SARS-CoV-2 in 2019/2020 will challenge global health for the future. Because a vaccine against the virus will not be available in the near future, we herein try to offer a pharmacological strategy to combat the virus. There exists a number of candidate drugs that may inhibit infection with and replication of SARS-CoV-2. Such drugs comprise inhibitors of TMPRSS2 serine protease GENE_OR_GENE_PRODUCT and inhibitors of angiotensin-converting enzyme 2 GENE_OR_GENE_PRODUCT ( ACE2 GENE_OR_GENE_PRODUCT ). Blockade of ACE2 GENE_OR_GENE_PRODUCT , the host cell receptor for the S protein of SARS-CoV-2 and inhibition of TMPRSS2 GENE_OR_GENE_PRODUCT , which is required for S protein priming may prevent cell CELL entry of SARS-CoV-2. Further, chloroquine SIMPLE_CHEMICAL and hydroxychloroquine SIMPLE_CHEMICAL , and off-label antiviral drugs, such as the nucleotide analogue remdesivir GENE_OR_GENE_PRODUCT , HIV ORGANISM protease inhibitors lopinavir SIMPLE_CHEMICAL and ritonavir SIMPLE_CHEMICAL , broad-spectrum antiviral drugs arbidol and favipiravir SIMPLE_CHEMICAL as well as antiviral phytochemicals available to date may prevent spread of SARS-CoV-2 and morbidity and mortality of COVID-19 pandemic.

Model 2

Abstract Outbreak and pandemic of coronavirus SARS-CoV-2 PROTEIN in 2019/2020 will challenge global health for the future. Because a vaccine against the virus will not be available in the near future, we herein try to offer a pharmacological strategy to combat the virus. There exists a number of candidate drugs that may inhibit infection with and replication of SARS-CoV-2 PROTEIN . Such drugs comprise inhibitors of TMPRSS2 serine protease PROTEIN and inhibitors of angiotensin-converting enzyme 2 PROTEIN ( ACE2 PROTEIN ). Blockade of ACE2 PROTEIN , the host cell receptor PROTEIN for the S protein PROTEIN of SARS-CoV-2 PROTEIN and inhibition of TMPRSS2 PROTEIN , which is required for S protein PROTEIN priming may prevent cell entry of SARS-CoV-2 PROTEIN . Further, chloroquine and hydroxychloroquine, and off-label antiviral drugs, such as the nucleotide analogue remdesivir, HIV protease inhibitors lopinavir and ritonavir, broad-spectrum antiviral drugs arbidol and favipiravir as well as antiviral phytochemicals available to date may prevent spread of SARS-CoV-2 PROTEIN and morbidity and mortality of COVID-19 pandemic.

PMC7167713 Structure based computational assessment of channel properties of assembled ORF-8a from SARS-CoV

Model 1

ORF 8a is a short 39 amino acid bitopic membrane protein encoded by severe acute respiratory syndrome causing corona virus SIMPLE_CHEMICAL ( SARS-CoV ORGANISM ). It has been identified to increase permeability of the lipid membrane CELLULAR_COMPONENT for cations SIMPLE_CHEMICAL . Permeability is suggested to occur due to the assembly of helical bundles. Computational models of a pentameric assembly of 8a peptides SIMPLE_CHEMICAL are generated using the first 22 amino acids, which include the transmembrane CELLULAR_COMPONENT domain. Low energy structures reveal a hydrophilic pore mantled by residues Thr-8, and −18, Ser-11, Cys-13, and Arg-22 GENE_OR_GENE_PRODUCT . Potential of mean force (PMF) profiles for mono (Na(+), K(+), Cl(− GENE_OR_GENE_PRODUCT )) and divalent ( Ca(2+)) ions GENE_OR_GENE_PRODUCT along the pore are calculated. The data support experimental findings of a weak cation selectivity of the channel. Calculations on 8a are compared to data derived for a pentameric bundle consisting of the M2 helices CELLULAR_COMPONENT of the bacterial pentameric ligand gated ion channel GLIC (3EHZ). PMF curves of both, bundles 8a GENE_OR_GENE_PRODUCT and M2 GENE_OR_GENE_PRODUCT , show sigmoidal shaped profiles. In comparison to the data for the M2-GLIC model, data of the 8a bundle show lower amplitude of the PMF values between maximum and minimum and less discrimination amongst ions SIMPLE_CHEMICAL . Proteins 2015; 83:300–308. © 2014 Wiley Periodicals, Inc.

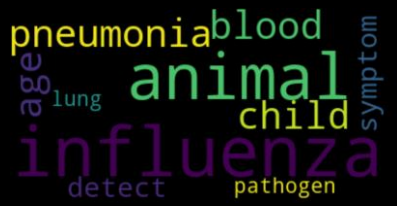
Model 2

ORF 8a PROTEIN is a short 39 amino acid bitopic membrane protein encoded by severe acute respiratory syndrome causing corona virus (SARS-CoV). It has been identified to increase permeability of the lipid membrane for cations. Permeability is suggested to occur due to the assembly of helical bundles. Computational models of a pentameric assembly of 8a peptides are generated using the first 22 amino acids, which include the transmembrane domain PROTEIN . Low energy structures reveal a hydrophilic pore mantled by residues Thr-8 PROTEIN , and −18, Ser-11, Cys-13 DNA , and Arg-22 PROTEIN . Potential of mean force (PMF) profiles for mono ( Na(+ PROTEIN ), K(+ PROTEIN ), Cl(− PROTEIN )) and divalent (Ca(2+)) ions along the pore are calculated. The data support experimental findings of a weak cation selectivity of the channel. Calculations on 8a are compared to data derived for a pentameric bundle consisting of the M2 helices PROTEIN of the bacterial pentameric ligand gated ion channel GLIC (3EHZ) PROTEIN . PMF curves of both, bundles 8a and M2 PROTEIN , show sigmoidal shaped profiles. In comparison to the data for the M2-GLIC model, data of the 8a bundle show lower amplitude of the PMF values between maximum and minimum and less discrimination amongst ions. Proteins 2015; 83:300–308. © 2014 Wiley Periodicals, Inc.

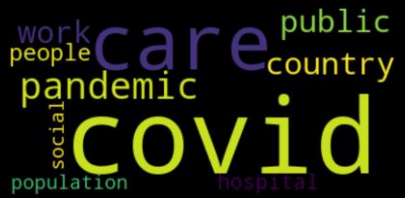
# Improved Topic Modelling For Document Categorisation & Clustering

**Refined over previous feature presentation.** For purposes of demo, example illustrates a 2-layer topic modelling design over 50K Cord-19 article abstracts. First, 3 main topic themes are generated, followed by 10 sub-topic themes. Each document is categorised with both a main & sub-theme topic. This provides a hierarchical topic model that can be drilled down. In reality, sub-themes could comprise hundreds of topics to facilitate exploration by topic filter & determining related articles during search result compilation. The inter-topic distance map, reducing topics to 2 dimensions using PCA, shows good cluster separation (except 6 & 10, ), one tool to help determine best number of topics in each layer.

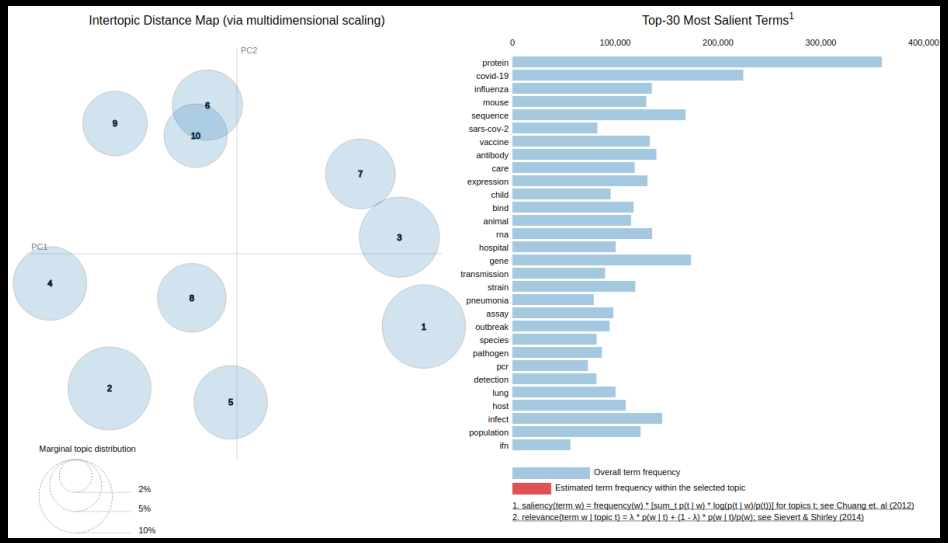
**Main Topic 1**




**Main Topic 2**




**Main Topic 3**


**Sub Topic 7**
7803 docs

care
covid-19
hospital
pandemic
medical
staff
contact
healthcare
participant
work

**Sub Topic 4**
7047 docs

covid-19
sars-cov-2
mortality
symptom
trial
sars
ace2
outcome
lung
therapy

**Sub Topic 2**
6762 docs

public
country
social
people
global
development
work
policy
government
community

**Sub Topic 9**
5156 docs

protein
sequence
rna
bind
gene
acid
genome
structure
site
domain

**Sub Topic 8**
4659 docs

expression
activity
protein
immune
mouse
gene
receptor
Ifn
pathway
induce

**Sub Topic 6**
3978 docs

solution
concentration
data
point
air
particle
structure
average
node
performance

**Sub Topic 1**
3794 docs

antibody
vaccine
mouse
animal
strain
serum
infect
antigen
culture
assay

**Sub Topic 3**
3781 docs

population
transmission
species
outbreak
infect
host
animal
epidemic
sequence
bat

**Sub Topic 5**
3559 docs

influenza child
detection
pcr
detect
pneumonia
pathogen
assay
rsv
bacterial

**Sub Topic 10**
3059 docs

blood
lung
therapy
tissue
lesion
common
pulmonary
diagnosis
normal
chronic

# Search State-Of-The-Art

**Implemented using deep bidirectional transformers for language understanding.** Results ranked by score, with the question directly answered with the most relevant excerpt from the paper (highlighted red).

Unfortunately model training limited to only 5K articles due to local lab system GPU VRAM constraints. Better results expected with larger training set.

Tested with 3 transformer models pre-trained on biomedical datasets.

**Question** Which have reported the highest levels of neutralizing Abs after a second, boost vaccination?

**Paper ID** PMC7115484
**Title** Nasal delivery of Protollin-adjuvanted H5N1 vaccine induces enhanced systemic as well as mucosal immunity in mice
**Rank** 1
**Score** 83.97
**Authors** Cao, Weiping; Kim, Jin Hyang; Reber, Adrian J.; Hoelscher, Mary; Belser, Jessica A.; Lu, Xiuhua; Katz, Jacqueline M.; Gangappa, Shivaprakash; Plante, Martin; Burt, David S.; Sambhara, Suryaprakash
**Answer Excerpt** However Protollinadjuvanted vaccine even at the lowest vaccine dose 03 µg significantly reduced weight loss and completely protected mice against the lethal challenge Fig1C and D. Therefore Protollin also enhanced crossclade protection with significant antigen dosesparing. Next we assessed whether the improved protection conferred by Protollinadjuvanted H5N1 vaccine was due to enhanced antibody titers. Mice were immunized as described in Fig 1 and sera were collected at 3 weeks following primary immunization week 3 and booster immunization week 7 to measured HI titers against a homologous AVN120304 virus as well as heterologous AIN0505 virus. Mice immunized with unadjuvanted H5N1 vaccine did not induce detectable antibody titers following primary immunization

**Paper ID** PMC7115571
**Title** SARS coronavirus spike polypeptide DNA vaccine priming with recombinant spike polypeptide from Escherichia coli as booster induces high titer of neutralizing antibody against SARS coronavirus
**Rank** 2
**Score** 82.86
**Authors** Woo, Patrick C.Y.; Lau, Susanna K.P.; Tsoi, Hoi-wah; Chen, Zhi-wei; Wong, Beatrice H.L.; Zhang, Linqi; Chan, Jim K.H.; Wong, Lei-po; He, Wei; Ma, Chi; Chan, Kwok-hung; Ho, David D.; Yuen, Kwok-yung
**Answer Excerpt** coli were able to generate high titer of neutralizing antibody against SARSCoV Table 2 Groups 4 and 6. This indicates that the type of vaccine used for priming is crucial in determining the type of immune response developed. Subsequent doses will booster the immune response generated by the first dose of vaccine. Of note is that the humoral immune response developed in mice primed with spike polypeptide DNA vaccine and boosted with Speptide from E. coli was not particularly of the Th1 type as compared to that developed in mice immunized with Speptide from E

**Paper ID** PMC7125630
**Title** Frontiers of transcutaneous vaccination systems: Novel technologies and devices for vaccine delivery
**Rank** 3
**Score** 82.73
**Authors** Matsuo, Kazuhiko; Hirobe, Sachiko; Okada, Naoki; Nakagawa, Shinsaku
**Answer Excerpt** 001 following the first vaccination using the TCI formulation indicating that a single application of our TCI formulation could induce an immune response in humans. We also administered a second vaccination to five subjects in whom neither antibody titer was significantly increased by the first vaccination. The IgG titers increased in a part of subjects following the second vaccination suggesting that an additional application increases the efficacy of the TCI formulation. Antibody titers on day 365 after application of the TCI formulation were maintained at a higher level than those on day 0 in all subjects examined although antibody titers tended to be lower on day 365 than on day 60 62. Conventional patchbased TCI systems require the pretreatment of disrupting or removing SC but our hydrogel patch achieved Ag penetration into skin without removing the SC and Agspecific antibodies were produced in some subjects by a single application in humans which represents a safety and efficacy advantage

**Question** What mutations have been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-2 virus?

**Paper ID** PMC2443636
**Title** Mutation in murine coronavirus replication protein nsp4 alters assembly of double membrane vesicles
**Rank** 1
**Score** 82.96
**Authors** Clementz, Mark A.; Kanjanahaluethai, Amornrat; O'Brien, Timothy E.; Baker, Susan C.
**Answer Excerpt** However the mechanism by which this substitution in nsp4 causes the defect in RNA synthesis in Alb ts6 is not known. A schematic diagram of nsp4 topology indicating the position of the two asparagine residues modified by Nlinked glycosylation and an asparagine to threonine change predicted to be responsible for the temperature sensitive phenotype are depicted in Fig 1D. To determine if nsp4N176 nsp4N237 or nsp4N258 is important for nsp4 function we generated virus encoding each specific substitution. Each substitution was introduced into the MHVA59 genome using a reverse genetics approach pioneered by Yount et al. 2002 as described in the Materials and methods

**Paper ID** PMC7168560
**Title** Emergence and adaptive evolution of Nipah virus
**Rank** 2
**Score** 82.81
**Authors** Li, Kemang; Yan, Shiyu; Wang, Ningning; He, Wanting; Guan, Haifei; He, Chengxi; Wang, Zhixue; Lu, Meng; He, Wei; Ye, Rui; Veit, Michael; Su, Shuo
**Answer Excerpt** It is located very close to the fusion peptide the distance to I122 the Cterminal residue of the fusion peptide is only 5 Å Figure 4b. Furthermore residue 42 is also part of the strap region composed of βsheets which is implicated in interactions with the G protein. To analyse where these amino acids are located in the postfusion structure of F we labelled the analog residues in the F protein of Newcastle Disease virus. Upon activation the F protein undergoes largescale refolding mostly in DIII. The heptad repeat region A HRA extends into a long αhelix which forms a sixhelix bundle with the HRB region of the stalk

**Paper ID** PMC4125544
**Title** Nidovirus papain-like proteases: multifunctional enzymes with protease, deubiquitinating and deISGylating activities
**Rank** 3
**Score** 82.08
**Authors** Mielech, Anna M.; Chen, Yafang; Mesecar, Andrew D.; Baker, Susan C.
**Answer Excerpt** The authors showed that PRRSV P2 can block Sendai virus induced IFNβ and inhibit NFκB by preventing IκBα degradation by its deubiquitination in cell culture. In addition they generated several mutant versions of P2 that had reduced ability to inhibit NFκBactivation. To test if P2 can block NFκB activation in the context of the virus the authors introduced these mutations into PRRSV. Although some of the mutations that altered P2 activity in vitro could not be recovered as viable virus the authors did report two single amino acid P2 mutant viruses that showed decreased ability to inhibit NFκBreporter activity and a decrease in the level of IκBα in infected cells Sun et al 2010. However the authors did not address whether protease or DUB activity was responsible for the effect they observed in infected cells

# Summary

**Main research cockpit concept is viable.** Main features realised as working prototypes. Value of NLP processing complex, biomedical specific problem domain, free text dataset is demonstrated.

**Clarity on NLP pipeline requirements** via prototyping at granular level with Jupyter notebooks, will help Azure industrialisation.

**Better understanding of compute resource requirements**. For example, through prototyping main features of each NLP pipeline stage, steps which are more CPU bound (e.g. data cleansing) or GPU bound (e.g. transformer model training) have been identified. This supports infrastructure design of the pipeline & more appropriate/cost effective allocation of compute resources.

**How to unambiguate authors?** How do you separate one author John Doe from another author John Doe, so that each can be uniquely identified and associated with the correct citations and co-author network?

**Main challenge has been demands on local lab compute resources.** Data cleansing steps time to complete limited by CPU clock speed and IPC (instructions per cycle). Dataset size to be encoded by transformer models for search limited by available GPU VRAM. Justifies migration to cloud (and possibly upgrade of lab resources).

**Next Steps** Confident that prototyping has proven the main required features viable, next step is to design the required Azure infrastructure and migrate prototyping for a cloud-hosted minimum viable product with user interface.