SOFTWARE Open Access

A pipeline for the retrieval and extraction of domain-specific information with application to COVID-19 immune signatures

Adam J. H. Newton[1,2,3,4], David Chartash[2,3,5], Steven H. Kleinstein[4,6,7] and Robert A. McDougal[2,3,7*]

*Correspondence: robert.mcdougal@yale.edu

[1] Department of Physiology and Pharmacology, SUNY Downstate Health Sciences University, Brooklyn, NY 11203, USA
[2] Yale Center for Medical Informatics, Yale School of Medicine, Yale University, New Haven, CT 06511, USA
[3] Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT 06511, USA
[4] Department of Pathology, Yale School of Medicine, Yale University, New Haven, CT 06511, USA
[5] School of Medicine, University College Dublin - National University of Ireland, Dublin, Co. Dublin, Republic of Ireland
[6] Department of Immunobiology, Yale School of Medicine, Yale University, New Haven, CT 06511, USA
[7] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

Abstract

Background: The accelerating pace of biomedical publication has made it impractical to manually, systematically identify papers containing specific information and extract this information. This is especially challenging when the information itself resides beyond titles or abstracts. For emerging science, with a limited set of known papers of interest and an incomplete information model, this is of pressing concern. A timely example in retrospect is the identification of immune signatures (coherent sets of biomarkers) driving differential SARS-CoV-2 infection outcomes.

Implementation: We built a classifier to identify papers containing domain-specific information from the document embeddings of the title and abstract. To train this classifier with limited data, we developed an iterative process leveraging pre-trained SPECTER document embeddings, SVM classifiers and web-enabled expert review to iteratively augment the training set. This training set was then used to create a classifier to identify papers containing domain-specific information. Finally, information was extracted from these papers through a semi-automated system that directly solicited the paper authors to respond via a web-based form.

Results: We demonstrate a classifier that retrieves papers with human COVID-19 immune signatures with a positive predictive value of 86%. The type of immune signature (e.g., gene expression vs. other types of profiling) was also identified with a positive predictive value of 74%. Semi-automated queries to the corresponding authors of these publications requesting signature information achieved a 31% response rate.

Conclusions: Our results demonstrate the efficacy of using a SVM classifier with document embeddings of the title and abstract, to retrieve papers with domain-specific information, even when that information is rarely present in the abstract. Targeted author engagement based on classifier predictions offers a promising pathway to build a semi-structured representation of such information. Through this approach, partially automated literature mining can help rapidly create semi-structured knowledge repositories for automatic analysis of emerging health threats.

Keywords: COVID-19, Biomarkers, Data mining, Immunity, Knowledge bases

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

The rapid growth in scientific publications [1] presents a challenge for researchers to seeking a comprehensive understanding of the literature. This challenge is of particular importance in emerging disciplines and domains without existing comprehensive reviews or widely accepted frameworks for representing the field. The COVID-19 pandemic is one such example of an emerging publication phenomenon. While machine learning has provided many solutions for search problems related to information retrieval (IR) [2], application of IR to specific scientific domains remains an active area of research [3, 4]. Researchers have leveraged search engines to retrieve relevant literature, with keywords searches [5] or alerts [6], but these approaches usually require substantial further refinement.

Once relevant sources have been retrieved, information has to be obtained from the text. For some domains, machine consumable structures make specific data types trivial to extract, e.g. genes [7] and proteins [8], however integrating this information with a more comprehensive data model remains challenging. There are many methods to obtain salient information from identified sources, including; manual curation e.g. HIPC [5], rule-based semi-automated extraction of metadata from an abstract, e.g. the metadata suggestions for ModelDB [9], and PICO (population, intervention, control, and outcomes) extraction [10], which tags words related to the PICO elements in randomized control trials. Given the novelty of the scientific domain of COVID-19 research, it is difficult to known what information characterizes this subfield and how it will be presented in the paper. Thus, a semi-automated human in the loop approach facilitates a solution.

COVID-19 may affect the human immune system in different ways. These effects—which could be at the level of changes of gene expression, of proteins, of metabolites, of antibodies, etc.—may vary by population (e.g. young vs old), disease severity (e.g. mild vs severe), etc, with each pattern of effects constituting an *immune signature* for the disease. For some diseases (e.g. cervical cancer [11]), immune signatures have shown potential as predictors of survival or other clinical outcomes. Unfortunately, identifying papers containing human immune signatures and locating those immune signatures within publications is non-trivial. Immune signatures can appear in the text, figures or tables, with dozens of distinct signatures in a single publication, and may not be presented as the principle finding.

We developed a semi-automated pipeline (Fig. 1), which utilized human-in-the-loop learning. As part of this pipeline we have created and validated a literature classifier that uses the abstracts and titles to retrieve papers likely to contain human COVID-19 immune signatures from a corpus of scientific literature. The pipeline then uses author solicitation: authors were asked to fill out a structured form describing the immune signature(s) in their papers. Author-supplied signatures from over thirty such papers are available on our website at covid-signatures.org [12].

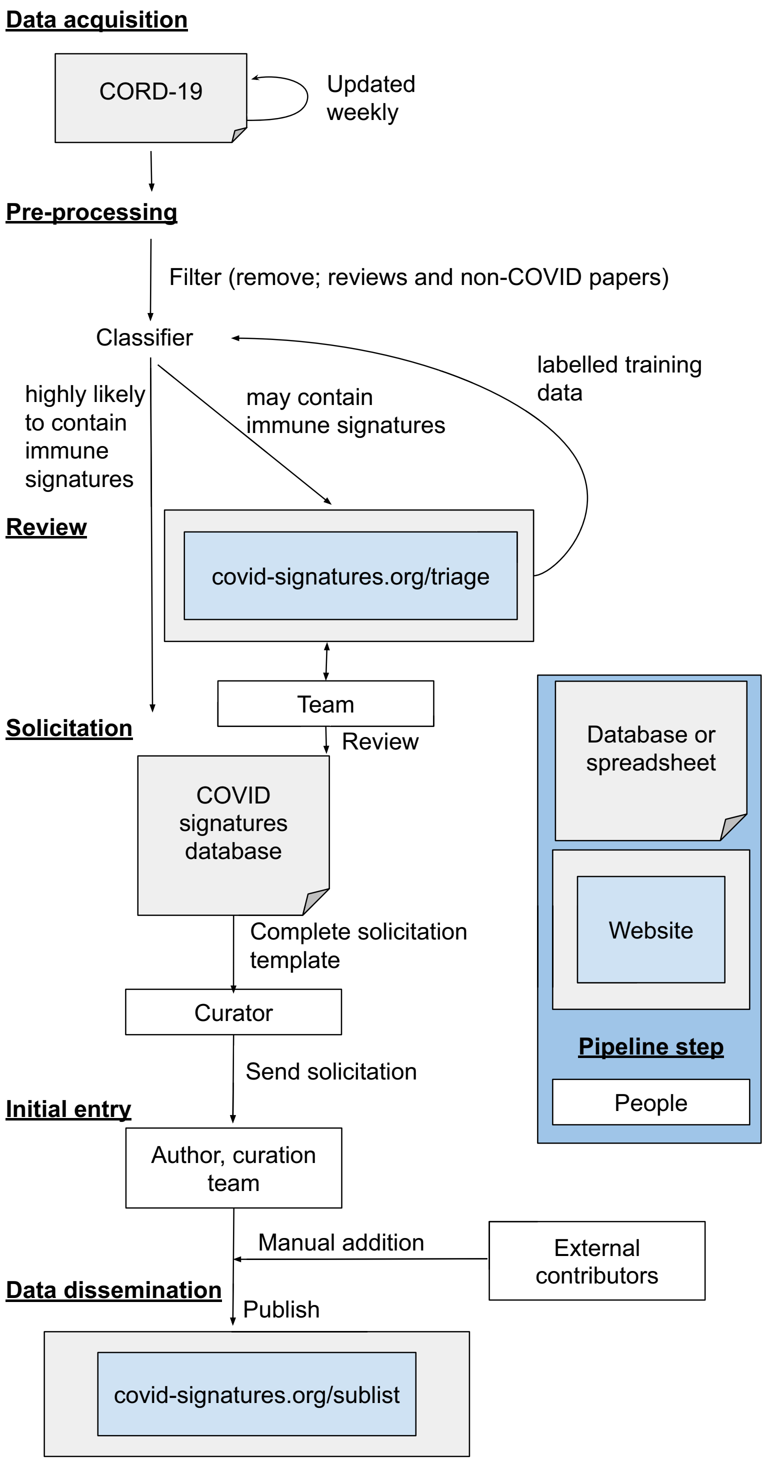


Fig. 1 Pipeline for semi-automated curation of COVID-19 immune signatures. Data acquisition utilized CORD-19 which required prepossessing; including substantially filtering of the dataset and removing duplicate entries. Triage divides the articles into one of 3 relevant ("Type A", "Type B" and "Type C") or 2 not-relevant ("review article" and "no signature") classes, which are then used to build a classifier to provide further articles for triage as well as directly for solicitation. Solicitation is done in a semi-automated fashion using an email template and an online form for the author to complete. After solicitation the information is made available via covid-signatures.org.

Implementation

Generic online platform

We developed a general purpose online literature review, author solicitation, and information sharing. The platform is powered by the Django web framework [13] for templating and user management, MongoDB database backend [14], Bootstrap [15] for layout, and jQuery [16] for streamlined scripting (Fig. 1). The pipeline keeps track of timestamps, change history, and associated user IDs for auditing and error recovery. Visitor privacy is respected by not sending cookies or other tracking mechanisms to non-logged in users, while authentication status is preserve across page loads using cookies. This generic platform is freely available at [17].

Software architecture

The pipeline is powered by Django, a web framework that handles templating and user management, and MongoDB, a database backend for efficient data storage. Front-end components, such as Bootstrap and jQuery, enhance the user interface and facilitate streamlined scripting. The architecture includes modules for triage and solicitation, which are crucial components of the system. The modular architecture of the pipeline project enables flexibility, maintainability, and scalability.

The triage module focuses on effectively managing the document triaging process. It incorporates features such as document queues, review stages, and status tracking. By leveraging MongoDB, the module ensures access to document information and supports efficient retrieval and updating of metadata. The solicitation module enables communication and feedback collection from authors. It provides a form with a unique URL for each entry in the database, allowing authors to submit details of immune signatures in their papers. The solicitation module ensures secure and accountable data entry, with the entered data stored in the database and logged for auditing purposes.

COVID-19 immune signature pipeline

To adapt the generic platform for COVID-19 immune signatures, a JSON-encoded configuration file was used to specify database details, paper categories, explanatory text, data solicitation forms, email templates, etc (see Additional File 1). This semi-automated pipeline consists of several stages (Fig. 1). (1) Data acquisition. (2) Pre-processing to remove duplicates and identify papers for review or solicitation. (3) Review of papers by experts to identify papers for solicitation and build a labeled dataset. (4) Solicitation, contacting authors for immune signatures. (5) Initial entry, authors provide immune signatures via an online form. (6) Data dissemination, immune signatures are published online.

Data acquisition

For interoperability with other COVID-19 literature analysis efforts through the use of shared identifiers, we leveraged the Allen Institute's COVID-19 Open Research

Dataset (CORD-19) [18]. CORD-19 provides a corpus with clear information retrieval benchmarks (see TREC-COVID challenge [4, 19, 20]), standardized machine readable data and SPECTER document embedding of each title and abstract [21].

CORD-19 is regularly updated (often weekly); we use these updates to add new papers to our pipeline. The results presented here are base on the 8th November 2021 release of CORD-19.

Pre-processing

To ensure the quality and relevance of the dataset for COVID-19 immune signatures, we performed extensive pre-processing by applying filters and removing duplicates. To focus on primary sources that could contain COVID-19 immune signatures, we filtered this dataset to exclude:

- PubMed papers with "Comment", "Review", "Editorial", or "News" article type.
- Papers from journals whose journal title includes "rev" as a whole word or as the start of a word to avoid review journals.
- Papers published before December 1, 2019.
- Papers that do not explicitly mention "COVID" or a related term (e.g. "2019-nCoV") in the paper title or abstract.

Duplicate and near duplicate papers frequently end up in the corpus due to the presence of papers released on preprint services before their formal publication in a journal. The CORD-19 Unique Identifier is linked to a conceptual document, which may include multiple versions of the manuscript. Using the SPECTER embedding, we identified near-duplicate papers by grouping documents within a certain proximity. By experimenting with a 30-unit threshold in the SPECTER embedding space, we effectively excluded most duplicates while retaining distinct articles. For each group of papers, only the most recently released paper was used for further analysis. It is worth noting that some near duplicates may have missing metadata for one duplicate, resulting in discrepancies in processing. For this reason, we also remove entries with near duplicates in our excluded set.

We developed a two-stage Support Vector Machine (SVM) based classifier to determine the presence and then the type of immune signature for the filtered CORD-19 literature. For the first stage, the SVM model (polynomial kernel of degree 4) simply seeks to determine if a paper contains an immune signature or not; this SVM was trained by grouping the three immune signature classes into one super class (signature present) and the "review article" and "no signature" classes into another super class (signature not present). A second SVM model (polynomial kernel degree 5), trained on only the papers confirmed by our expert reviewers to contain a COVID-19 immune signature, was used to predict the type of immune signature that would be present for those papers predicted by the first classifier to contain an immune signature. Probabilities were obtained from the SVM classifiers using Platt scaling [22]. Feature vectors are used by SVMs to construct a decision boundary; we used the SPECTER embeddings [21] (768-dimensional feature vectors). SPECTER utilizes SciBERT [23], a transformer pre-trained on scientific literature, to create document level embeddings from the titles and abstracts of papers, not the full text. Both SVM models were trained and applied on the SPECTER embeddings of the title and abstract, not directly on the text. Source code for pre-processing, classification, and querying the database is included (Additional File 2). The SVMs used |sklearn| version 0.24.2 [24] with Python 3.6.10.

Review

Expert review was performed using the aforementioned platform (Fig. 1). A limited set of rules based on whole-word matching (e.g. the word "patients" implies that the paper studied humans) were used to tag the abstracts so that reviewers could examine by tag if desired (see Additional file 3: Table S1). Three expert reviewers each with at least five years graduate computational immunology training, examined the papers in the queue to determine whether or not they contained immune signatures and, if so, what type. The reviewers were presented with a title that links to the paper full text, abstract, selected metadata, and buttons to indicate their conclusions. To support reviewer corrections to automatic database population, an edit button allowed changes to the title and URL which were then pushed to the server via an AJAX call. For papers with a COVID-19 immune signature, reviewers were asked to choose from three broad classes of immune signatures. We included two additional review queues: "let's discuss" for papers where the category was not obvious, and "review article" for work that may have a human immune signature but not be the primary source. An additional, auto-saved notes field allowed reviewers to make notes for themselves and for any future discussion. After 288 papers were reviewed, we tasked one of the expert reviewers with re-reviewing the papers to identify key words from the abstract that, in their judgment, made it more (e.g."IL-8") or less (e.g. "influenza") likely that a paper would contain a COVID-19 immune signature. These identified terms were highlighted in the abstracts during the review phase (see Additional file 3: Table S2).

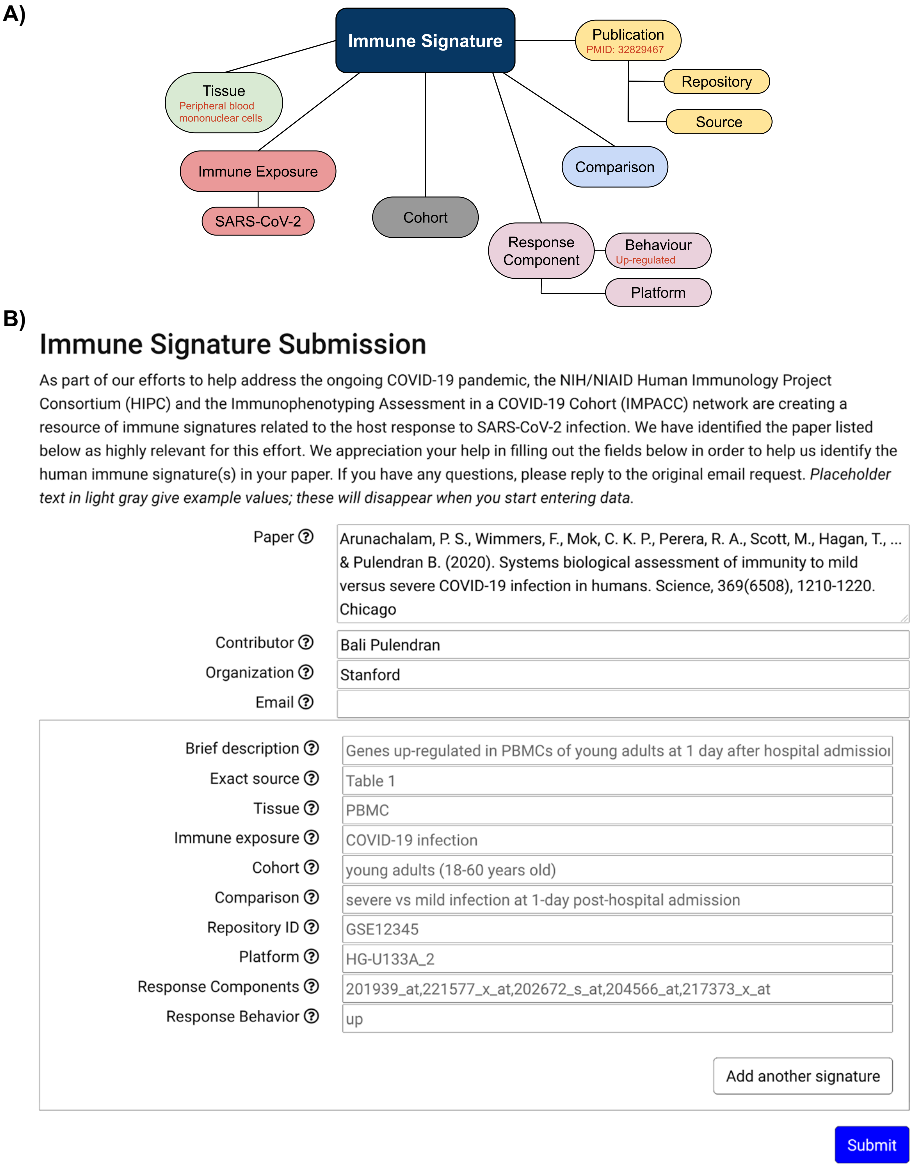
Solicitation

Once papers containing COVID-19 immune signatures have been identified, either manually or automatically, we contacted the corresponding author(s) to request details of the immune signatures in the paper, corresponding to our data model (Fig. 2).

Our platform provides a form with a unique URL for each entry in the database. This form identifies two classes of data—one that is global and applies to the entire form—and one that pertains to a specific fact about the paper (in our case, a specific immune signature), of which there can be many. The field names are configurable via the JSON configuration file, but for this project, the form asks for global data identifying the paper and contributor (reference, contributor, organization, and email address), and specific instance data about each immune signature (description, location in the paper, tissue, immune exposure, cohort, comparison, any repository ID, analysis platform, response components and response direction).

Initial entry

The process of providing details of an immune signature from a paper involves generating a customized email with field data, enabling recipients to access a paper-specific page for data entry and storage, ensuring security and accountability. The email



Immune Signature Submission

As part of our efforts to help address the ongoing COVID-19 pandemic, the NIH/NIAID Human Immunology Project Consortium (HIPC) and the Immunophenotyping Assessment in a COVID-19 Cohort (IMPACC) network are creating a resource of immune signatures related to the host response to SARS-CoV-2 infection. We have identified the paper listed below as highly relevant for this effort. We appreciation your help in filling out the fields below in order to help us identify the human immune signature(s) in your paper. If you have any questions, please reply to the original email request. *Placeholder text in light gray give example values; these will disappear when you start entering data.*

Paper ?	Arunachalam, P. S., Wimmers, F., Mok, C. K. P., Perera, R. A., Scott, M., Hagan, T., ... & Pulendran B. (2020). Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science, 369(6508), 1210-1220. Chicago
Contributor ?	Bali Pulendran
Organization ?	Stanford
Email ?	
Brief description ?	Genes up-regulated in PBMCs of young adults at 1 day after hospital admissio
Exact source ?	Table 1
Tissue ?	PBMC
Immune exposure ?	COVID-19 infection
Cohort ?	young adults (18-60 years old)
Comparison ?	severe vs mild infection at 1-day post-hospital admission
Repository ID ?	GSE12345
Platform ?	HG-U133A_2
Response Components ?	201939_at,221577_x_at,202672_s_at,204566_at,217373_x_at
Response Behavior ?	up

Fig. 2 Immune signatures model and solicitation. **A** Our data model for COVID-19 immune signatures is based on [5]. **B** The immune signature template provided to authors, with help text shown in gray

recipients receive a link to a page for just their specific paper, which does not require a login. Each field on the form has associated help text and examples. All fields are editable by the email recipient except for the paper reference. An arbitrary number of immune signatures may be entered for each paper. When the contributors press the "submit" button, the entered data is stored in the database and logged in a separate file, allowing administrators to revert to a previous version in the case of accidental or malicious changes after initial data entry. A typical data entry form is shown in Fig. 2.

To allow third-party manual solicitations, a "submit your immune signatures" button on the [12] homepage opens an entry form that is the same as one seen by solicited contributors except without the pre-filled global fields and with an editable paper reference field. These entries are assigned an automatically generated internal identifier which the website administrators can later map to a CORD-19 identifier.

Data dissemination

The contributor-entered details of the immune signatures are stored in a MongoDB database and are made available via [12] in both HTML and JSON. Internal users can access the full database entries, whereas the public version does not include edit history, contributor details, or internal identifiers (but does include the CORD-19 identifiers).

Data analysis

Python Data Analysis Library |pandas| 0.24.2 [25] was used to manage the data from CORD-19 and the pipeline for analysis. We used |sklearn| 0.24.2 [24] to perform k-means clustering, for SVM and logistic regression classifiers with tolerance set to 10^{-7}. Uniform manifold approximation and projection (|UMAP|) 0.4.6 [26] was used to visualize the clusters. |SciPy| 1.5.4[27] was used for statistical tests. To evaluate the classifier, we used Natural Language Toolkit (|NLTK|) 3.5 [28] for tokenization, excluding English stop words and words with less than three characters. WordNet [29] was used to lemmatize the words. TD-IDF computation was facilitated by the |sklearn| package, including 1- and 2-grams. The logistic regression classification used inverse of regularization strength 50. When comparing word frequencies, we excluded words that occurred fewer than five times in the titles and abstracts of the selected papers. We used |Gensim| 3.8.3 [30] to perform Latent Dirichlet Allocation (LDA) [31] with 1000 iterations and 100 passes, on the filtered CORD-19 abstracts and titles, excluding words that occurred in more than 80% of abstracts or fewer than 5%.

Results

Our overall workflow involved the development of a training set of papers containing COVID-19 immune signatures, SPECTER and SVM-powered identification of papers likely to have these immune signatures, expert review of a subset of papers, data solicitation from the authors, and then data dissemination on the covid-signatures.org site (summarized in Fig. 1). As we envision this workflow will generalize to other semi-structured data acquisition efforts, we developed a generic online platform to streamline its application.

Papers with immune signatures are clustered in SPECTER abstract embedding

We sought to determine whether SPECTER embedding preserved sufficient information to identify papers containing COVID-19 immune signatures. SPECTER provides document-level embeddings using a pre-trained language model. The CORD-19 dataset provides a SPECTER embedding, a 768 dimensional vector representation of the title and abstract, for each of the papers. We manually identified 5 papers from CORD-19 with COVID-19 immune signatures. We also considered an additional

69 papers with non-COVID immune signatures that were identified as part of the Human Immunology Project Consortium (HIPC) [5], together with a matched control group of papers for the HIPC immune signatures taken from the same volume of the journals. We used the pre-trained SPECTER model [21] to obtain an embedding for each of the additional papers from their title and abstract. *K*-means clustering applied to the SPECTER embeddings identified $k = 6$ clusters based on Akaike Information Criterion (AIC) [32]. Almost all of the papers with immune signatures were grouped in a single cluster, cluster 3. Four of the 5 papers with COVID-19 signatures and 68 of 69 papers with vaccination signatures were in cluster 3. The remaining papers with immune signatures were part of a second cluster, cluster 6 (1 of 5 papers with COVID-19 signatures, and 1 of 69 papers with vaccination signatures). In contrast, a control group of papers were more widely dispersed in the embedding space with a significantly different distribution compared to the papers with vaccination signatures ($\chi^2 = 14.42$, $p = 0.006$) (Fig. 3).

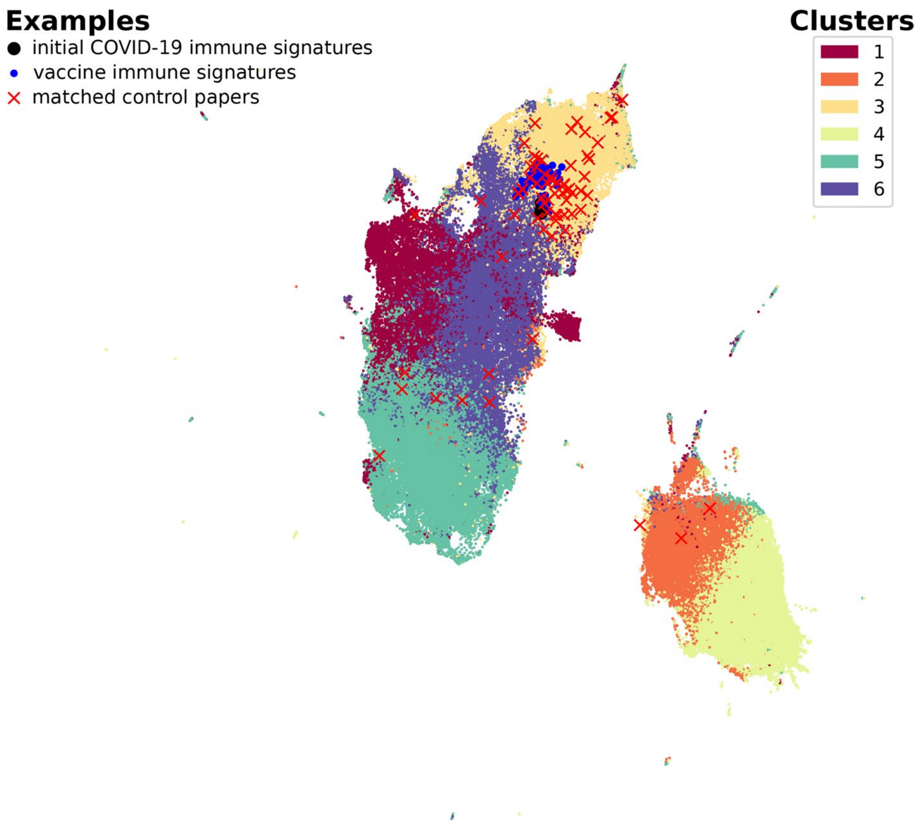


Fig. 3 Uniform manifold approximation and projection (UMAP) of $\sim 170,000$ papers from the CORD-19 dataset (small points) based on their SPECTER embedding. Manually selected papers with COVID-19 immune signatures ($n = 5$; large points), vaccine response immune signatures ($n = 69$; * markers), and matched control papers ($n = 68$, x markers) are marked. Each paper was assigned to a cluster via *k*-means clustering, with $k = 6$ based on the AIC

This significant sub-clustering suggests that the SPECTER embedding effectively preserves information on whether or not a paper contains an immune signature. These results suggest that we can use the SPECTER embedding as the basis to construct a classifier to identify papers with COVID-19 immune signatures.

Papers with COVID-19 signatures can be predicted with high accuracy

We built a classifier to determine if the SPECTER embedding could be used to predict which papers contained COVID-19 immune signatures. To create the classifier we iteratively built a training set, consisting of papers with a label indicating whether they contained immune signatures. We took advantage of the close grouping of immune signature papers in the SPECTER embedding by starting with five initial papers with COVID-19 signatures (shown in Fig. 3), and identifying the nearest 100 papers to each of these points in the embedding space. We also included the 100 papers from the filtered CORD-19 dataset that were nearest to the center of the vaccine immune signatures.

These papers were labeled by expert reviewers, using the pipeline interface. To allow simultaneous article review by multiple parties, only a random, small portion of the dataset is shown on the pipeline review platform at a time by default, minimizing the risk of duplicate review of the same paper. Overall, this identification and review process resulted in 271 papers, of which 140 contained immune signatures. An additional 52 papers from CORD-19 identified by rule-based filters were subjected to expert review, identifying another 11 relevant papers.

Inter-rater reliability of signature presence

To test the reliability of our expert reviews, we selected 100 articles at random from the set of papers to be reviewed independently by two reviewers. The two reviewers had 92% agreement on determining whether a paper contained a COVID-19 immune signature. Cohen's kappa coefficient [33], a robust measure of inter-rater reliability that accounts for chance agreements, shows very good agreement (0.84, 0.73–0.95 95% confidence interval). Papers where reviewers did not agree on the presence of an immune signature were not included in the training set for the classifier.

Classifier development and performance

A SVM classifier [24] was fit to the 316 papers unambiguously identified by reviewers as either containing or not containing immune signatures. To achieve reliable classifier predictions, we iteratively selected additional papers for review based on the prediction of the classifier. We added papers with estimated probability of containing human COVID-19 immune signatures between 0.80 and 1.0 to increase the number of positive examples (Fig. 4).

This iteration was deemed sufficient as the marginal improvements to the classifier's performance were small. Specifically, the average reduction in positive predictive value (PPV) based on Leave-One-Out Cross Validations (LOOCV) with one fewer paper was 0.10 ±0.28 (from 81.3 to 81.2%). Overall, this yielded a set of 216 papers containing immune signatures and 185 papers without immune signatures.

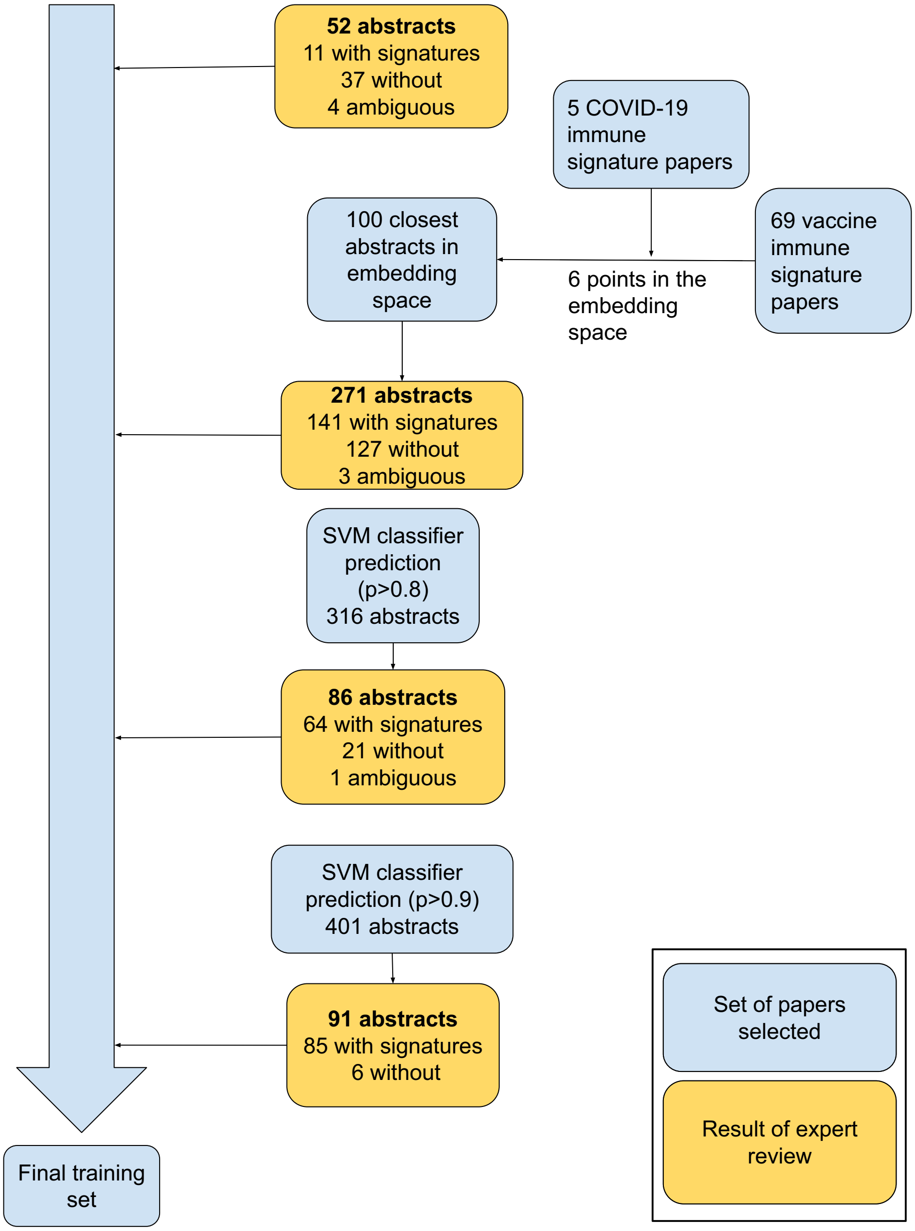


Fig. 4 Iterative identification of papers for developing the classifier. The iterative process was initiated with papers selected with rule-based filters for expert review. We then leveraged the SPECTER embedding to select papers for review based on their distances in the embedding space. Once an initial training set was labeled, we fit an SVM classifier with the 316 abstracts and used it to select 86 additional papers. An SVM classifier fitted to the 401 abstract, predicted 91 highly likely papers to validate the classifier. This gave a final training set of 492 papers, 301 with immune signatures, 8 of which were ambiguous (where reviewers did not agree on the presence or absence of immune signatures)

Using a SVM classifier trained on these 401 papers, we selected an additional 91 papers based on predicted probability ($>= 0.8$) to be highly likely to contain a COVID-19 immune signature. An expert reviewer (SK) then classified these articles, and determined that the majority (92%) contained immune signatures. Given the size of the

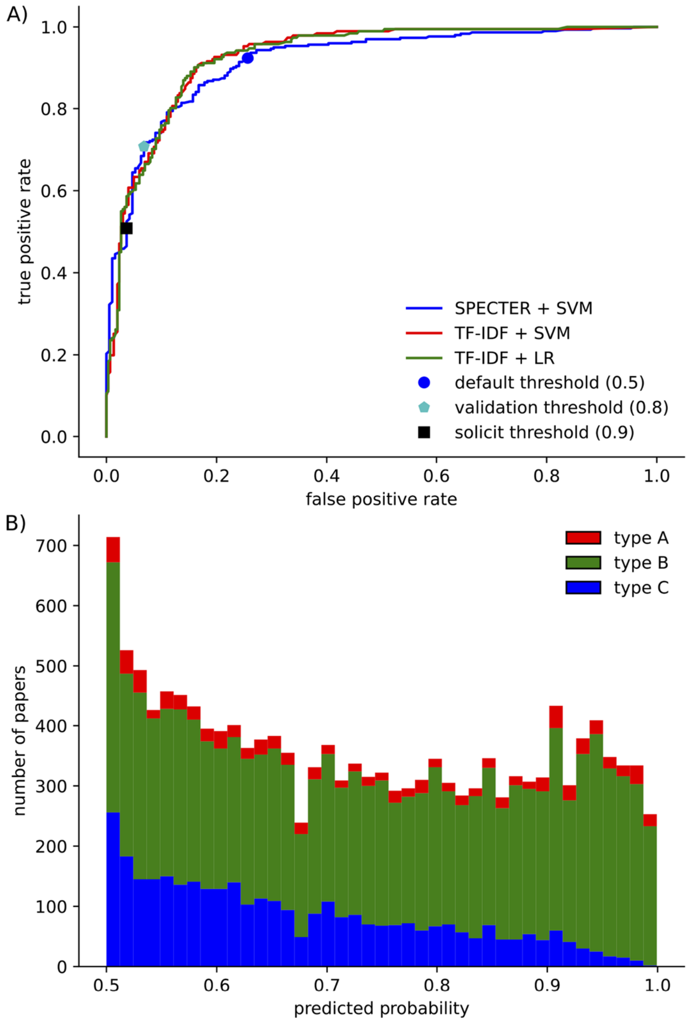


Fig. 5 The classifier can identify papers containing immune signatures from the SPECTER embedding **A** The receiver operating characteristic (ROC) curve based on LOOCV for the classifier with the default threshold (0.5), validation threshold (0.8) and direct solicitation threshold (0.9) highlighted. Different lines are shown for the SVM classifier with SPECTER, SVM classifier with TF–IDF and LR with TF–IDF. **B** Number of papers in the CORD-19 dataset predicted to have immune signatures of different types at different thresholds of the classifier. The first SVM classifier predicted the probability of the paper containing an immune signature at the threshold indicated. The second SVM classifier then predicted the type of immune signature

Table 1 Classifier confusion matrix based on LOOCV

	Present	Not present
Predicted present	**278**	49
Predicted not present	23	**142**

Here the "present" class represents the presence of any type of human COVID-19 immune signature in the paper

Correct predictions are highlighted in bold

corpus, and the fact that emailing corresponding authors included a manual step (i.e., a human-in-the-loop, see methods), we decided to use a greater specificity threshold to select papers for direct solicitation in order to increase efficiency. Considering a higher threshold (predicted probability $>=$ 0.9), of the 61 papers that met this threshold, 60 of them contained COVID-19 immune signatures (94%). These additional papers gave a final training set of 492 papers, with 301 containing COVID-19 immune signatures.

We evaluated the performance of the SVM classifier on our training set using LOOCV. The receiver operating characteristic (ROC) area under the curve (AUC) was 0.916 (Fig. 5A).With a probability threshold of 0.5, the SVM classifier had a PPV of 86%, with accuracy 85%, sensitivity 92%, selectivity 74% and F1-Score 89% (Table 1).Applying this SVM classifier to the filtered CORD-19 collection identified (at the probability threshold of 0.5) around 15,500 papers ($\sim$ 2% of the CORD-19 corpus) are likely to contain COVID-19 immune signatures.

After the iterative construction of the training set was complete, we evaluated whether the SPECTER embedding was essential for achieving ongoing high performance of the classifier. We compared the SVM classifier using the SPECTER embeddings to the widely used approach of Term Frequency–Inverse Document Frequency (TF–IDF) [34] of the titles and abstracts, with both a SVM classifier and with logistic regression. TF–IDF with SVM classifier achieved an AUC of 0.928 using LOOCV, which was very similar to the SPECTER approach. Likewise TF–IDF with logistic regression had similar performance with an AUC 0.928 (Fig. 5A). As these alternative approaches performed similarly, we chose to continue to use the SPECTER embedding for the rest of this study as it was supplied with the CORD-19 dataset and did not require any additional text processing.

The SPECTER embedding captures information about the type of COVID-19 signature

We next sought to determine if the SPECTER embedding of titles and abstracts contained sufficient information to distinguish between papers describing different types of COVID-19 immune signatures. We divided the COVID-19 immune signatures into three types: (A) Type A papers contained gene expression signatures, (B) Type B papers included signatures involving proteins, metabolites and/or cell types, and (C) Type C papers included all other COVID-19 immune signatures.

Inter-rater reliability of signature type

To verify that the different signature types were meaningful and distinct, two experts reviewed a set of 100 papers. Papers where reviewers agreed on the type of immune signature (38 of 46) were added to the training set for the second stage classifier. The classes of immune signatures were well recognized by our reviewers, with substantial (82%) agreement when including whether a paper contained an immune signature or not and the type of immune signature (κ 0.72, 0.61–0.83 95% confidence interval) (Table 2).

Table 2 Independent expert review of articles showed substantial agreement between the three immune signature classes

	Type A signature	Type B signature	Type C signature	No signature
Type A signature	**7**	2	0	0
Type B signature	2	**21**	2	1
Type C signature	1	3	**8**	1
no signature	1	0	5	**46**

Bold numbers indicate consensus between the expert reviewers

Table 3 Second stage classifier confusion matrix based on LOOCV

	Type A	Type B	Type C
Type A signature predicted	**42**	6	2
Type B signature predicted	15	**135**	45
Type C signature predicted	1	18	**27**

The correct predictions are highlighted in bold

Classifier performance

Using the papers that were initially identified as containing COVID-19 immune signatures, we fit a "second stage" SVM classifier to predict the type of signature. To evaluate performance, we tested the classifier using the 84 of the 91 papers used to validated the previous classifier. The 91 papers were predicted to contain immune signatures (probability threshold $>=0.8$) and the 84 used here are those that were validated by expert review as actually containing immune signatures. These papers included all three of the signature types: 10 type A, 58 type B, and 16 type C signatures. On this independent test set, the weighted averages (where each class's contribution is scaled by the fraction of papers in that class) were PPV 74%, with sensitivity of 58%, specificity 58% and F_1 score 61%. Incorporating these newly reviewed papers provided 291 annotated papers with immune signatures (out of the 506 total papers in our final training set). We then refit the classifier to the 291 papers with signature types. LOOCV of this "second stage" classifier (Table 3), gave weighted average PPV 69% with sensitivity 70%, selectivity 80%, and F_1-Score 69%.

Despite having the fewest examples in the training set, the classifier performed best on "Type A" signatures, the narrowest category. While the broadest category "Type C" had the greatest proportion of incorrect predictions. These performances strongly suggests that the embedding is able to capture the the differences between signature types.

Features driving the classification

To discover which features our classifiers were using to determine the presence of immune signature, we compared papers with high and low probabilities of containing COVID-19 immune signatures. For the high probability papers, we selected the 500 articles with the highest predicted probability to contain an immune signature based on the classifier. As a comparison group, we selected 500 papers with a low probability (~ 0.1) of containing an immune signature. This low probability was chosen for the comparison set to avoid selecting marginal papers from the corpus and those not written in English. A log-likelihood comparison of word frequencies [35] between these two groups identified 171 words with significantly different frequencies, of which 38% were more frequent in papers containing immune signatures (χ_1^2 with adjusted 5% threshold for 876 comparisons). The top 10 differences are shown (Additional file 3: Table S3) and are consistent with the focus of the classifier on human immunology, e.g. "patient", "health", "cell" and "severe".

To further interrogate the classifier, we applied LDA topic modeling to the CORD-19 corpus filtered to include only potential relevant entries as described in methods (Fig. 6).

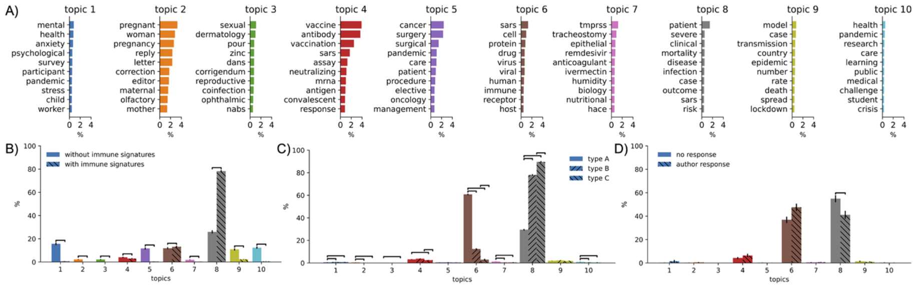


Fig. 6 LDA of the CORD-19 dataset with average scores given to a samples of papers based on SVM classifier predictions. **A** 10 topics identified from the abstracts and titles of the filtered CORD-19 corpus. **B** comparison of the average topic composition for a sample of 500 papers predicted to contain COVID-19 immune signatures and 500 papers predicted not to contain immune signatures. **C** comparison of average topic composition for three samples of 500 papers predicted to contain a specific type of COVID-19 immune signature. **D** comparison of average topic composition for the 31 solicited papers where authors provided immune signatures and the 69 where they did not. Error bars show the standard error of the mean. Statistically significant differences between topics is shown based on Mann–Whitney U tests at the 5% significance level using the Dunn–Šidák correction for multiple tests

When comparing the sets of papers described above, papers predicted to contain immune signatures were predominantly related to topics 6 and 8, which appear to describe immunological and clinical work. There were significant differences in the topic composition across all topics (Mann–Whitney U tests at the Dunn–Šidák corrected 5% significance level for 10 tests).

Next we consider which features may be driving the "second stage" classifier in assigning different signature types. We chose a random sample of 500 papers most likely to contain a particular immune signature type (prediction $p > 0.5$ of containing immune signatures and $p > 0.5$ for the type of immune signature). Comparing word frequencies between the titles and abstracts of the different signature types identified 413 statistically significantly different words for type A COVID-19 immune signatures (26% had higher relative frequency in type A sample). Similarly for the predicted type B titles and abstracts there where 343 statistically significantly different words (14% with greater relative frequency in the type B sample) and 326 for type C (with 15% with greater relative frequency in the type C sample). For each case, the ten words that differed the most are shown (see Additional file 3: Table S3), with bold typeface indicating they occurred with greater relative frequency in that sample. These word frequencies suggest one way that the classifier predicts the paper contains an immune signature may be the presence of certain words associated with immune signatures (e.g. "patient", "cell", etc). In contrast, the type of immune signature may be determined by a reduction in the frequency of confounding words, such as "patient", "clinical", etc for type A immune signatures. We performed LDA topic modeling for different types of immune signatures. While papers with all types of signatures were predominately associated with topics 6 and 8, there were significant differences between them, with type A being 61% topic 6, type B 12% and type *C* only 3%.

High author response rate was found for immune signature extraction

Previous work demonstrated many authors are willing to provide both data and metadata about their work [9]. To discover if this is also the case for COVID-19 immune signatures, we developed a semi-automated pipeline to contact the corresponding author(s) of papers and ask them to provide details of their published immune signatures via an online form (Fig. 2).

Articles were chosen via expert review or based on a predicted probability from the SVM classifier. Expert review is the gold standard for the presence of immune signatures, when using the classifier, to avoid an unnecessary burden corresponding with authors of potentially unrelated work, we required a higher specificity when retrieving papers for solicitation. We used probabilities > 0.9 for containing an immune signature (from the first SVM classifier) and combined probability > 0.8 of it being type A or B (from the second classifier). The high threshold for the predicted probability of having any immune signature changes the confusion matrix and subsequently increases the specificity to 96% at the cost of sensitivity (50%), with PPV 96%. Although the SVM classifiers were used to select articles for direct solicitation, a human was kept in-the-loop for quality control and validation. We selected a convenience sample of 100 papers to directly solicit information on the COVID-19 immune signatures (63 of these papers were selected by the classifier and 37 by expert review).

We sent solicitations to authors of the 100 selected papers; 31% of these authors contributed immune signature information. Notably, we sent only a single solicitation email to each author with no reminders. The majority of responses (20 of 31 papers) submitted a single immune signature, while 2 signatures were specified in 5 of the responses, 3 signatures were specified in 2 responses, and one contributor each specified 4, 12, 14, and 27 signatures for a single paper. While the submission form did not have any required fields, all of the responses included a free-text description of the signature, the comparison underlying the signature, and the source of the signature in the paper (e.g., the figure or table number). However, only 35% (11 or 31 responses) provided a list of the immune response components (e.g., gene or protein names) that comprise the signature. Thus, while the overall response rate to direct solicitation was reasonable, follow-up is needed to obtain signature details.

We next sought to determine if there were differences between the papers the led to successful solicitations compared with those that did not. The likelihood of obtaining an authors response for the 63 papers selected with our classifiers was not correlated with the predicted probability of the paper containing an immune signature (Point-biserial correlation coefficient coefficient $r = -8.09 \times 10^{-4}$, $t_{61} = 6.32 \times 10^{-3}$ $p = 0.995$). The lack of correlation may be because of the small probability range chosen for solicitation. We did not detect a difference in the predicted class ("Type A", "Type B" or "Type C") for papers with or without author responses ($\chi^2_2 = 0.47,\ p = 0.93$). Finally, although the response rate of papers chosen by the classifier was lower than those chosen by expert review (27% vs. 38%, respectively), this difference was not statistically significant ($\chi^2_1 = 0.89$, $p = 0.35$).

Topic modeling of the papers where we solicited further information showed they were predominately topics 6 and 8 (Fig. 6D). There were significant differences between the contribution of topic 8 where authors responded compared with papers where authors did not, 5% level with Dunn–Šidák correction for 10 comparisons (Topic 6; Mann–Whitney U $p = 0.015$, Topic 8; $p = 0.0045$).

Discussion

We systematically analyzed the CORD-19 dataset [18], a standardized collection of COVID-19 related literature, to identify papers that contained immune signatures and to classify the type of immune signatures. There were two key challenges to building such a classifier: (1) a limited number of examples of relevant papers, and (2) immune signatures themselves are generally not reported in the title or abstract, the text that was available to analyze. Using SPECTER embeddings [21] solved both of these challenges as this pre-trained, SciBERT-powered [23] model places similar papers (e.g. those that have COVID-19 immune signatures) near each other in the embedding space without requiring any custom training. Taking advantage of this property, we were able to iteratively identify a collection of COVID-19 immune signature papers from only 5 initially known examples, using a human-in-the-loop process. Once a large set of examples was gathered, SPECTER lost its advantage and conventional approaches like TF–IDF with logistic regression provided slightly better performance, at the cost of requiring an explicit training step. For papers predicted to contain immune signatures, we solicited additional details from the corresponding authors, and disseminated the information received on the covid-signatures.org website.

The platform we developed to do this iterative positive sampling, human-in-the-loop review, solicitation, and data sharing is available on GitHub at [17]. We designed this platform to be easily adaptable to other domains with all the immune signature-specific parts confined to either configuration files or data acquisition once an appropriate data model has been identified. In particular, the platform described here has also been applied to a computational neuroscience model repository, ModelDB [36]. As with COVID-19 immune signatures, the presence of results derived from a computational model can often be inferred from the abstract, but the model details are generally not themselves present in the abstract. With ModelDB, we found this platform to be an effective way to distribute the task of reviewing large numbers of papers across multiple reviewers who have varying availability to review. This process of review by domain experts, although time consuming, generates valuable data that can be helpful for training future classifiers. In early work, the emphasis was to identify relevant papers, so we sorted the candidate papers by a heuristic for probability of relevance based on the number of inclusion criteria met. We currently have an unmodified version of the platform on our development site http://modeldb.science with ModelDB providing its own interface for data sharing. In the process, we learned that it was important to distinguish between papers that are not relevant for inclusion in ModelDB because of text reasons (e.g. they do not involve computational neuroscience) from those that are not relevant for inclusion for non-text reasons (e.g. being a preprint instead of a journal article). This distinction is essential for enabling community reuse, which we envision to be focused on abstract concepts (in our case, if something involves computational neuroscience) and not on repository-specific choices.

In both of our use cases, engaging the authors is essential. In ModelDB's case, we have the advantage of reaching out on behalf of a community tool that has existed for over 25 years; we believe familiarity helps increase the willingness of the community to engage. Nonetheless, even for this new COVID-signatures project, with no publications or examples of how others have shared their data, 31% of solicited authors contributed immune signatures in response to a single solicitation. The willingness of authors to share information, likely varies by field and over time as the culture of the field shifts (see e.g. the discussion in [37]). Nonetheless, improving survey responses is a common challenge in social sciences research. To increase the number of survey responses, certain design decisions have been found to be effective [38]. These include using multiple contacts and mixed modes of invitation, ensuring that invitation text is complete and persuasive but not overly verbose, placing the survey URL at the end of the invitation, and providing accurate time/effort estimates and using an authoritative subject line. As the curation project grows, it may incentivize authors to provide more detailed findings to increase the visibility and impact of their work.

A high response rate is only useful if the responses are of sufficient quality to be reusable. A key challenge is balancing the need to minimize the burden on the contributor with the desire for structured information. In our case, we provided structured fields and examples, and requested authors report information as it appeared in the paper rather than attempting to translate terms to a standardized form (e.g., leveraging the Cell Ontology for cell types that are reported as response components) [39]. Thus, signatures included references to "NK cell", "CD3 T cells" and "gammadelta T cells," which could then be mapped to terms from the Cell Ontology: "natural killer cell" (CL:0000623), "T cell" (CL:0000084), and "gamma-delta T cell" (CL:0000798), respectively in post processing. While such data needs to be further processed to be standardized, it also gives us a data set that represents immune signatures as they are likely to appear in publications, which can assist in future development of methods to detect immune signatures and their component parts.

We focused the machine learning and natural language processing on the paper titles and abstracts, but there is more information available in the full text and, in particular, the COVID-19 immune signatures are themselves expressed in the body of the paper (figure captions, results, etc). However, the structured full-text is only available for about a quarter of the CORD-19 dataset and there is lower signal-to-noise ratio in the full-texts than in abstracts [9, 40]. Nonetheless, some types of data are relatively unambiguously identifiable from the full text. In particular, journals often require certain types of data to be archived in a repository, such as GEO, ArrayExpress and FlowRepository. These repositories use standardized formats for their accession IDs that can be identified using regular expressions. However if the paper contains multiple signatures in multiple repositories, manual curation is still necessary to correctly divide them between immune signatures.

More sophisticated language models facilitate higher performance on information extraction tasks, e.g. SciBERT (which underlies SPECTER) has shown success in retrieving Population Intervention Comparator Outcome (PICO) elements in the randomized controlled trial literature [10]. The PICO elements have been tested to both identify text spans and extract structured information for automated biomedical evidence extraction [41]. While PICO elements partially overlap the immune signature data model (Fig. 2), there are several challenges to leverage PICO extraction to further automate COVID-19 signature extraction. For example, information about immune signatures is often not present in the text, but rather in figures or tables [5]. Current PICO extraction methods also group the intervention and comparison categories, which makes it more difficult to apply them to our data model. Papers containing multiple signatures present the additional challenge of grouping the PICO elements for a particular immune signature. PICO predictions could be used to assist manual entry, such as by providing paper-specific autocomplete suggestions or placeholder text to reduce the burden of data entry. This approach risks, however, biasing the author responses. The author-supplied information may be used to improve machine learning for identifying PICO information by providing data for model training and testing, although this may be complicated by the variability with how author contributors specify their COVID-19 immune signatures.

Crowd-sourcing further curation offers a potential mechanism for reducing variability and providing consistently annotated human COVID-19 immune signatures similar to those made available through the HIPC Dashboard [5]. A portal is under development that facilities community annotation of COVID-19 immune signatures community.covid-signatures.org, utilizing the signatures retrieved by this pipeline.

Identifying papers containing COVID-19 immune signatures and collecting data and contextual information (metadata) regarding the signatures can help speed scientific progress in this area. As has been seen with vaccination [42] and inflammation [43], providing access to such a resource supports secondary and comparative analyses resulting in a broader understanding of immune system response. We have demonstrated this pipeline approach (Fig. 1) is able to identify and classify relevant papers from a large and varied corpus (Fig. 5), starting from only a few examples. Our method has also shown that authors are often willing to provide useful clinical and contextual information about their work.

Conclusion

This paper describes the development of a pipeline to retrieve papers containing COVID-19 immune signatures and for their semi-automated curation. Within this pipeline, we iteratively built a training set and incorporated machine learning to classify papers from an existing repository (CORD-19). We found that SPECTER embedding provides a good reduced representation of a paper and its relatedness to other papers that can be adopted for the purpose of identifying scientifically salient features of the paper (in this case immune signatures). However, SPECTER was not a necessary component, as TF–IDF with logistic regression has similar performance to the SPECTER approach. Thirty-one percent of authors of papers with immune signatures voluntarily provided semi-structured representations in response to a request from our team, regardless of the immune signature type.

Given its start as a neuroinformatics tool, the successful application to COVID-19 demonstrates that this pipeline approach is readily adaptable for other fields to identify papers containing scientifically relevant features, which can be further processed—by data solicitation, manual curation, or automated means-to extract the relevant data for presentation in a unified knowledge base or dashboard.

Abbreviations

COVID-19	Coronavirus disease 2019
SVM	Support vector machine
IR	Information retrieval
HIPC	Human Immunology Project Consortium

PICO	Population, intervention, control, and outcomes
URL	Uniform resource locator
AJAX	Asynchronous JavaScript and XML
IL-8	Interleukin 8
CORD-19	COVID-19 Open Research Dataset
TREC-COVID	Text REtrieval Conference COVID
JSON	JavaScript Object Notation
HTML	HyperText Markup Language
UMAP	Uniform manifold approximation and projection
NLTK	Natural language toolkit
TF–IDF	Term frequency–inverse document frequency
LDA	Latent Dirichlet allocation
AIC	Akaike information criterion
PPV	Positive predictive value
LOOCV	Leave-one-out cross validations
ROC	Receiver operating characteristic
AUC	Area under the curve
NK	Natural killer
CD3	Cluster of differentiation 3

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05397-8.

Additional file 1. A copy of the generic pipeline (github.com/mcdougallab/pipeline) that has been adapted for COVID-19 immune signatures.

Additional file 2. Python code use to query the pipeline, build and compare the classifiers.

Additional file 3. Supplementary tables of the tags and key words used in the pipeline, and the words frequency that differ between classes.

Acknowledgements

We thank Daniel Chawla and Bram Gerritsen for reviewing papers and providing feedback about the pipeline interface.

Author contributions

All authors wrote the manuscript and contributed to the design and data analysis of the project. RAM envisioned and developed the pipeline. SHK devised the project and constructed the data model. DC guided the data analysis. AJHN performed the data analysis.

Funding

Research reported by this publication was supported by the National Institute of Allergy and Infectious Diseases (NIAID), the National Institute on Deafness and Other Communication Disorders, and the National Library of Medicine of the National Institutes of Health under Award Numbers U19AI089992, R01DC009977, and 5T15LM007056. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

The signatures gathered during the current study are available in the COVID Signatures website, covid-signatures.org/sublist. The dataset generated to build the classifiers in the study are available from the corresponding author on reasonable request. Project name: COVID-19 immune signature pipeline Project homepage: covid-signatures.org Operating systems: Platform independent Programming language: HTML, CSS, Python. Other requirements: Django web framework, MongoDB, Bootstrap and jQuery. License: BSD 2-Clause "Simplified" License Any restrictions to use by non-academics: None

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

S.H.K. receives consulting fees from Peraton. A.J.H.N, D.C., and R.A.M declare that they have no competing interests.

Received: 26 January 2023 Accepted: 23 June 2023
Published online: 20 July 2023

References

1. National Science Board NSF. Publication output: US trends and international comparisons. Science and Engineering Indicators 2019 (2020).
2. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval, vol. 463. New York: ACM Press; 1999.
3. Khalid S, Khusro S, Ullah I, Dawson-Amoah G. On the current state of scholarly retrieval systems. Eng Technol Appl Sci Res. 2019;9(1):3863–70.
4. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, Soboroff I, Voorhees E, Wang LL, Hersh WR. TREC-COVID: rationale and structure of an information retrieval shared task for covid-19. J Am Med Inform Assoc. 2020;27(9):1431–6.
5. Smith KC, Chawla DG, Dhillon BK, Ji Z, Vita R, van der Leest E, Weng J, Tang E, Abid A, Peters B, et al. A curated collection of human vaccination response signatures. bioRxiv (2021).
6. Morse TM, Wang R, Carnevale NT, Shepherd GM, McDougal RA. Pipeline to promote discovery and sharing of computational neuroscience research. In: Program number 814.07 neuroscience meeting planner. Society for Neuroscience (2017).
7. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):884–91.
8. UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):506–15.
9. McDougal RA, Dalal I, Morse TM, Shepherd GM. Automated metadata suggestion during repository submission. Neuroinformatics. 2019;17(3):361–71.
10. Kang T, Zou S, Weng C. Pretraining to recognize pico elements from randomized controlled trial literature. Stud Health Technol Inform. 2019;264:188.
11. Yang S, Wu Y, Deng Y, Zhou L, Yang P, Zheng Y, Zhang D, Zhai Z, Li N, Hao Q. Identification of a prognostic immune signature for cervical cancer to predict survival and response to immune checkpoint inhibitors. Oncoimmunology. 2019;8(12):1659094.
12. COVID-19 immune signature pipeline. http://covid-signatures.org/. Accessed 08 Nov 2022.
13. Django: the web framework for perfectionists with deadlines. https://djangoproject.com. Accessed 08 Nov 2022.
14. MongoDB: the developer data platform. https://mongodb.com. Accessed 08 Nov 2022.
15. Bootstrap. the most popular HTML, CSS, and JS library in the world. https://getbootstrap.com. Accessed 08 Nov 2022.
16. jQuery: write less, do more. https://jquery.com/. Accessed 08 Nov 2022.
17. Paper processing pipeline. https://github.com/mcdougallab/pipeline. Accessed 08 Nov 2022.
18. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W. et al.: Cord-19: the covid-19 open research dataset. ArXiv (2020).
19. Voorhees E, Alam T, Bedrick S, Demner-Fushman D, Hersh WR, Lo K, Roberts K, Soboroff I, Wang LL. TREC-COVID: constructing a pandemic information retrieval test collection. In: ACM SIGIR Forum, vol 54. New York: ACM; 2021. p. 1–12.
20. Chen JS, Hersh WR. A comparative analysis of system features used in the TREC-COVID information retrieval challenge. J Biomed Inform. 2021;117:103745.
21. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. Specter: document-level representation learning using citation-informed transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics; 2020. p. 2270–82.
22. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif. 1999;10(3):61–74.
23. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019).
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
25. Pandas Development Team T. Pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134.
26. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018).
27. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J. Scipy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17(3):261–72.
28. Bird S. NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 interactive presentation sessions; 2006. p. 69–72.
29. Fellbaum C, editor. WordNet: an electronic lexical database. Language, speech, and communication. Cambridge: MIT Press; 1998.
30. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. ELRA, Valletta, Malta; 2010. p. 45–50. http://is.muni.cz/publication/884893/en.
31. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
32. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–23.
33. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
34. Jones KS. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28:11–21.
35. Rayson P, Garside R. Comparing corpora using frequency profiling. In: The workshop on comparing corpora; 2000. p. 1–6.
36. McDougal RA, Morse TM, Carnevale T, Marenco L, Wang R, Migliore M, Miller PL, Shepherd GM, Hines ML. Twenty years of ModelDB and beyond: building essential modeling tools for the future of neuroscience. J Comput Neurosci. 2017;42(1):1–10.
37. Ascoli GA. Turning the tide of data sharing. Neuroinformatics. 2019;17(4):473–4. https://doi.org/10.1007/s12021-019-09437-8.

38. Kaplowitz MD, Lupi F, Couper MP, Thorp L. The effect of invitation design on web survey response rates. Soc Sci Comput Rev. 2012;30(3):339–49.
39. Overton JA, Vita R, Dunn P, Burel JG, Bukhari SAC, Cheung K-H, Kleinstein SH, Diehl AD, Peters B. Reporting and connecting cell type names and gating definitions through ontologies. BMC Bioinform. 2019;20(5):259–64.
40. Tirupattur N. Text miner for hypergraphs using output space sampling. Ph.D. thesis (2011).
41. Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, Wallace BC. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: Proceedings of the conference. association for computational linguistics. Meeting. NIH Public Access; 2018. p. 197.
42. Brusic V, Gottardo R, Kleinstein SH, Davis MM. Computational resources for high-dimensional immune analysis from the human immunology project consortium. Nat Biotechnol. 2014;32(2):146–8.
43. Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, Mesirov JP, Haining WN. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. Immunity. 2016;44(1):194–206.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC