bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Linguistic Analysis of the bioRxiv Preprint Landscape

This manuscript (permalink) was automatically generated from greenelab/annorxiver_manuscript@2034e45 on May 12, 2021.

Authors

- **David N. Nicholson**
  0000-0003-0002-5761 · danich1 · dnicholson329
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, Philadelphia PA, USA · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (T32 HG000046)

- **Vincent Rubinetti**
  · vincerubinetti · vincerubinetti
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, Philadelphia PA, USA; Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (R01 HG010067)

- **Dongbo Hu**
  · dongbohu
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, Philadelphia PA, USA · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (R01 HG010067)

- **Marvin Thielk**
  0000-0002-0751-3664 · MarvinT · TheNeuralCoder
  Elsevier, Philadelphia PA, USA

- **Lawrence E. Hunter**
  0000-0003-1455-3370 · LEHunter · ProfLHunter
  Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora CO, USA · Funded by The Gordon and Betty Moore Foundation (GBMF4552)

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, Philadelphia PA, USA; Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora CO, USA; Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (R01 HG010067)

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Preprints allow researchers to make their findings available to the scientific community before they have undergone peer review. Studies on preprints within bioRxiv have been largely focused on article metadata and how often these preprints are downloaded, cited, published, and discussed online. A missing element that has yet to be examined is the language contained within the bioRxiv preprint repository. We sought to compare and contrast linguistic features within bioRxiv preprints to published biomedical text as a whole as this is an excellent opportunity to examine how peer review changes these documents. The most prevalent features that changed appear to be associated with typesetting and mentions of supplementary sections or additional files. In addition to text comparison, we created document embeddings derived from a preprint-trained word2vec model. We found that these embeddings are able to parse out different scientific approaches and concepts, link unannotated preprint-peer reviewed article pairs, and identify journals that publish linguistically similar papers to a given preprint. We also used these embeddings to examine factors associated with the time elapsed between the posting of a first preprint and the appearance of a peer reviewed publication. We found that preprints with more versions posted and more textual changes took longer to publish. Lastly, we constructed a web application (https://greenelab.github.io/preprint-similarity-search/) that allows users to identify which journals and articles that are most linguistically similar to a bioRxiv or medRxiv preprint as well as observe where the preprint would be positioned within a published article landscape.

Introduction

The dissemination of research findings is key to science. Initially, much of this communication happened orally [1]. During the 17th century, the predominant form of communication shifted to personal letters that were shared from one scientist to another [1]. Scientific journals didn't become a predominant mode of communication until the 19th and 20th centuries, when the first journal was created [1,2,3]. Although scientific journals became the primary method of communication, they added high maintenance costs and long publication times to scientific discourse [2,3]. Some scientists' solutions to these issues has been to communicate through preprints, which are scholarly works that have yet to undergo peer review process [4,5].

Preprints are commonly hosted on online repositories, where users have open and easy access to these works. Notable repositories include arXiv [6], bioRxiv [7] and medRxiv [8]; however, there are over 60 different repositories available [9]. The burgeoning uptake of preprints in life sciences has been examined through research focused on metadata from the bioRxiv repository. For example, life science preprints are being posted at an increasing rate [10]. Furthermore, these preprints are being rapidly shared on social media, routinely downloaded, and cited [11]. Some preprint categories are shared on social media by both scientists and non-scientists [12]. About two-thirds to three-quarters of preprints are eventually published [13,14] and life science articles that have a corresponding preprint version are cited and discussed more often than articles without them [15,16,17]. Preprints take an average of 160 days to be published in the peer-reviewed literature [18], and those with multiple versions take longer to publish[18].

The rapid uptake of preprints in the life sciences also poses challenges. Preprint repositories receive a growing number of submissions [19]. Linking preprints with their published counterparts is vital to maintaining scholarly discourse consistency but is challenging to perform manually [16,20,21]. Errors and omissions in linkage result in missing links and consequently erroneous metadata. Furthermore, repositories based on standard publishing tools are not designed to show how textual content of preprints is altered due to the peer review process [19]. Certain scientists have expressed concern that competitors could scoop them by making results available before publication [19,22]. Preprint

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

repositories by definition do not perform in-depth peer review, which can result in posted preprints containing inconsistent results or conclusions [17,20,23,24]; however, an analysis of preprints posted at the beginning of 2020 revealed that most underwent minor changes as they were published [25]. Despite a growing emphasis on using preprints to examine the publishing process within life sciences, how these findings relate to the text of all documents in bioRxiv has yet to be examined.

Textual analysis uses linguistic, statistical, and machine learning techniques to analyze and extract information from text [26]. For instance, scientists analyzed linguistic similarities and differences of biomedical corpora [27,28]. Scientists have provided the community with a number of tools that aide future text mining systems [29,30,31] as well as advice on how to train and test future text processing systems [32,33,34]. Here, we use textual analysis to examine the bioRxiv repository, placing a particular emphasis on understanding the extent to which full-text research can address hypotheses derived from the study of metadata alone.

To understand how preprints relate to the traditional publishing ecosystem, we examine the linguistic similarities and differences between preprints and peer-reviewed text and observe how linguistic features change during the peer review and publishing process. We hypothesize that preprints and biomedical text are pretty similar, especially when controlling for the differential uptake of preprints across fields. Furthermore, we hypothesize that document embeddings [35,36] provide a versatile way to disentangle linguistic features along with serving as a suitable medium for improving preprint repository functionality. We test this hypothesis by producing a linguistic landscape of bioRxiv preprints, detecting preprints that change substantially during publication, and identify journals that publish manuscripts that are linguistically similar to a target preprint. We encapsulate our findings through a web app that projects a user-selected preprint onto this landscape and suggests journals and articles that are linguistically similar. Our work reveals how linguistically similar and dissimilar preprints are to peer-reviewed text, quantifies linguistic changes that occur during the peer review process, and highlights the feasibility of document embeddings with respect to preprint repository functionality and peer review's effect on publication time.

Materials and Methods

Corpora Examined

Text analytics is generally comparative in nature, so we selected three relevant text corpora for analysis: the BioRxiv corpus, which is the target of the investigation, the PubMedCentral Open Access corpus, which represents the peer-reviewed biomedical literature, the New York Times Annotated Corpus, which is used a representative of general English text.

BioRxiv Corpus

BioRxiv [7] is a repository for life sciences preprints. We downloaded an XML snapshot of this repository on February 3rd, 2020, from bioRxiv's Amazon S3 bucket [37]. This snapshot contained the full text and image content of 98,023 preprints. Preprints on bioRxiv are versioned, and in our snapshot, 26,905 out of 98,023 contained more than one version. When preprints had multiple versions, we used the latest one unless otherwise noted. Authors submitting preprints to bioRxiv can select one of twenty-nine different categories and tag the type of article: a new result, confirmatory finding, or contradictory finding. A few preprints in this snapshot were later withdrawn from bioRxiv; when withdrawn, their content is replaced with the reason for withdrawal. As there were very few withdrawn preprints, we did not treat these as a special case.

PubMed Central Open Access Corpus

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PubMed Central (PMC) is a digital archive for the United States National Institute of Health's Library of Medicine (NIH/NLM) that contains full text biomedical and life science articles [38]. Paper availability within PMC is mainly dependent on the journal's participation level [39]. PMC articles can be closed access ones from research funded by the NIH appearing after an embargo period or be published under Gold Open Access [40] publishing schemes. Individual journals have the option to fully participate in submitting articles to PMC, selectively participate sending only a few papers to PMC, only submit papers according to NIH's public access policy [41], or not participate at all. As of September 2019, PMC had 5,725,819 articles available [42]. Out of these 5 million articles, about 3 million were open access (PMCOA) and available for text processing systems [30,43]. PMC also contains a resource that holds author manuscripts that have already passed the peer review process [44]. Since these manuscripts have already been peer-reviewed, we excluded them from our analysis as the scope of our work is focused on examining the beginning and end of a preprint's life cycle. We downloaded a snapshot of the PMCOA corpus on January 31st, 2020. This snapshot contained many types of articles: literature reviews, book reviews, editorials, case reports, research articles, and more. We used only research articles, which aligns with the intended role of bioRxiv, and we refer to these articles as the PMCOA corpus.

The New York Times Annotated Corpus

The New York Times Annotated Corpus (NYTAC) is [45] is a collection of newspaper articles from the New York Times dating from January 1st, 1987, to June 19th, 2007. This collection contains over 1.8 million articles where 1.5 million of those articles have undergone manual entity tagged by library scientists [45]. We downloaded this collection on August 3rd, 2020, from the Linguistic Data Consortium (see Software and Data Availability section) and used the entire collection as a negative control for our corpora comparison analysis.

Mapping bioRxiv preprints to their published counterparts

We used CrossRef [46] to identify bioRxiv preprints linked to a corresponding published article. We accessed CrossRef on July 7th, 2020, and were able to link 23,271 preprints to their published counterparts successfully. Out of those 23,271 preprint-published pairs, only 17,952 pairs had a published version present within the PMCOA corpus. For our analyses that involved published links, we only focused on this subset of preprints-published pairs.

Comparing Corpora

We compared the bioRxiv, PMCOA, and NYTAC corpora to assess the similarities and differences between them. We used the NYTAC corpus as a negative control to assess the similarity between two life sciences repositories when compared with non-life sciences text. All corpora contain both words and non-word entities (e.g., numbers or symbols like $\pm$), which we refer to together as tokens to avoid confusion. We calculated the following characteristic metrics for each corpus: the number of documents, the number of sentences, the total number of tokens, the number of stopwords, the average length of a document, the average length of a sentence, the number of negations, the number of coordinating conjunctions, the number of pronouns and the number of past tense verbs. Spacy is a lightweight and easy-to-use python package designed to preprocess and filter text [47]. We used spaCy's "en_core_web_sm" model [47] (version 2.2.3) to preprocess all corpora and filter out 326 spaCy-provided stopwords.

Following that cleaning process, we calculated the frequency of every token across all corpora. Because many tokens were unique to one set or the other and observed at low frequency, we focused on the union of the top 0.05% (~100) most frequently occurring tokens within each corpus. We generated a contingency table for each token in this union and calculated the odds ratio along with

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the 95% confidence interval [48]. We measured corpora similarity by calculating the Kullback–Leibler (KL) divergence across all corpora along with token enrichment analysis. This metric measures the extent to which two distributions differ. A low value of KL divergence implicates that two distributions are similar and vice versa for high values. The optimal number of tokens used to calculate the KL divergence is unknown, so we calculated this metric using a range of the 100 most frequently occurring tokens between two corpora to the 5000 most frequently occurring tokens.

Constructing a Document Representation for Life Sciences Text

We sought to build a language model to quantify linguistic similarities of biomedical preprint and articles. Word2vec is a suite of neural networks designed to model linguistic features of words based on their appearance in the text. These models are trained to either predict a word based on its sentence context, called a continuous bag of words (CBOW) model, or predict the context based on a given word, called a skipgram model [35]. Through these prediction tasks, both networks learn latent linguistic features that can be used for downstream tasks, such as identifying similar words. We used gensim [49] (version 3.8.1) to train a CBOW [35] model over all the main text within each preprint in the bioRxiv corpus. Determining the best number of dimensions for word embeddings can be a non-trivial task; however, it has been shown that optimal performance is between 100-1000 dimensions [50]. We chose to train the CBOW model using 300 hidden nodes, a batch size of 10000 words, and for 20 epochs. We set a fixed random seed and used gensim's default settings for all other hyperparameters. Once trained, every token present within the CBOW model is associated with a dense vector representing latent features captured by the network. We used these word vectors to generate a document representation for every article within the bioRxiv and PMCOA corpora. For each document, we used spaCy to lemmatize each token and then took the average of every lemmatized token present within the CBOW model and the individual document [36]. Any token present within the document but not in the CBOW model is ignored during this calculation process.

Visualizing and Characterizing Preprint Representations

We sought to visualize the landscape of preprints and determine the extent to which their representation as document vectors corresponded to author-supplied document labels. We used principal component analysis (PCA) [51] to project bioRxiv document vectors into a low-dimensional space. We trained this model using scikit-learn's [52] implementation of a randomized solver [53] with a random seed of 100, an output of 50 principal components (PCs), and default settings for all other hyperparameters. After training the model, every preprint within the bioRxiv corpus is assigned a score for each generated PC. We sought to uncover concepts captured the generated PCs and used the cosine similarity metric to examine these concepts. This metric takes two vectors as input and outputs a score between -1 (most dissimilar) and 1 (most similar). We used this metric to score the similarity between all generated PCs and every token within our CBOW model for our use case. We report the top 100 positive and negative scoring tokens as word clouds. The size of each word corresponds to the magnitude of similarity, and color represents positive (orange) or negative (blue) association.

Discovering Unannotated Preprint-Publication Relationships

The bioRxiv maintainers have automated procedures to link preprints to peer-reviewed versions, and many journals require authors to update preprints with a link to the published version. However, this automation is primarily based on the exact matching of specific preprint attributes. If authors change the title between a preprint and published version (e.g., [54] and [55]), then this change will prevent bioRxiv from automatically establishing a link. Furthermore, if the authors do not report the publication to bioRxiv, the preprint and its corresponding published version are treated as distinct entities despite representing the same underlying research. We hypothesize that close proximity in

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the document embedding space could match preprints with their corresponding published version. If this finding holds, we could use this embedding space to fill in links missed by existing automated processes. We used the subset of paper-preprint pairs annotated in CrossRef as described above to calculate the distribution of available preprint to published distances. This distribution was calculated by taking the Euclidean distance between the preprint's embedding coordinates and the coordinates of its corresponding published version. We also calculated a background distribution, which consisted of the distance between each preprint with an annotated publication and a randomly selected article from the same journal. We compared both distributions to determine if there was a difference between both groups as a significant difference would indicate that this embedding method can parse preprint-published pairs apart. Following the comparison of the two distributions, we calculated distances between preprints without a published version link with PMCOA articles that weren't matched with a corresponding preprint. We filtered any potential links with distances greater than the minimum value of the background distribution as we considered these pairs to be true negatives. Lastly, we binned the remaining pairs based on percentiles from the annotated pairs distribution at the [0,25th percentile), [25th percentile, 50th percentile), [50th percentile, 75th percentile), and [75th percentile, minimum background distance). We randomly sampled 50 articles from each bin and shuffled these four sets to produce a list of 200 potential preprint-published pairs with a randomized order. We supplied these pairs to two co-authors to manually determine if each link between a preprint and a putative matched version was correct or incorrect. After the curation process, we encountered eight disagreements between the reviewers. We supplied these pairs to a third scientist, who carefully reviewed each case and made a final determination. Using this curated set, we evaluated the extent to which distance in the embedding space revealed valid but unannotated links between preprints and their published versions.

Measuring Time Duration for Preprint Publication Process

Preprints that are published can take varying amounts of time to be published. We sought to measure the time required for preprints to be published in the peer-reviewed literature and compared this time measurement across author-selected preprint categories as well as individual preprints. First, we queried bioRxiv's application programming interface (API) to obtain the date a preprint was posted onto bioRxiv as well as the date a preprint was accepted for publication. We measured time elapsed as the difference between the date at which a preprint was first posted on bioRxiv and its publication date. Along with calculating the amount of time elapsed, we also recorded the number of different preprint versions posted onto bioRxiv.

Using this captured data, we used the Kaplan-Meier estimator [56] via the KaplanMeierFitter function from the lifelines [57] (version 0.25.6) python package to calculate the half-life of preprints across all preprint categories within bioRxiv. We considered survival events as preprints that have yet to be published. There were a limited number of cases in which authors appeared to post preprints after the publication date, which results in preprints receiving a negative time difference, as previously reported [58]. We removed these preprints for this analysis as they were incompatible with the rules of the bioRxiv repository.

Following our half-life calculation, we measured the textual difference between preprints and their corresponding published version by calculating the Euclidean distance for their respective embedding representation. This metric can be difficult to understand within the context of textual differences, so we sought to contextualize the meaning of a distance unit. We accomplish this by first randomly sampled with replacement a pair of preprints from the Bioinformatics topic area as this was well represented within bioRxiv and contains a diverse set of research articles. Next, we calculated the distance between two preprints 1000 times and reported the mean. We repeated the above procedure using every preprint within bioRxiv as a whole. These two means serve as normalized benchmarks to compare against as distance units are only meaningful when compared to other distances within the same space. Following our contextualization approach, we performed linear

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

regression to model the relationship between preprint version count with a preprint's time to publication. We also performed linear regression to measure the relationship between document embedding distance and a preprint's time to publication. For this analysis, we retained preprints with negative time within our linear regression model, and we observed that these preprints had minimal impact on results. We visualize our version count regression model as a violin plot and our document embeddings regression model as a square bin plot.

Building Classifiers to Detect Linguistically Similar Journal Venues and Published Articles

Preprints are more likely to be published in journals that contained similar content to work in question. We assessed this claim by building classifiers based on document and journal representations. First, we removed all journals that had fewer than 100 papers in the PMC corpus. We held our preprint-published subset (see above section 'Mapping bioRxiv preprints to their published counterparts') and treated it as a gold standard test set. We used the remainder of the PMCOA corpus for training and initial evaluation for our models.

Specific journals publish articles in a focused topic area, while others publish articles that cover many topics. Likewise, some journals have a publication rate of at most hundreds of papers per year, while others publish at a rate of at least ten thousand papers per year. Accounting for these characteristics, we designed two approaches - one centered on manuscripts and another centered on journals.

We identified manuscripts that were most similar to the preprint query for the manuscript-based approach and evaluated where these documents were published. We embedded each query article into the space defined by the word2vec model (see above section 'Constructing a Document Representation for Life Sciences Text'). We selected manuscripts close to the query via Euclidean distance in the embedding space. Once identified, we return the journal in which these articles were published. We also return the articles that led to each journal being reported as this approach allows for journals that frequently publish papers to engulf our results.

We constructed a journal-based approach to accompany the manuscript-based process to account for the overrepresentation of these high publishing frequency journals. We identified the most similar journals for this approach by constructing a journal representation in the same embedding space. We computed this representation by taking the average embedding of all published papers within a given journal. We then projected a query article into the same space and returned journals close to the query.

Both models were constructed using the scikit-learn k-Nearest Neighbors implementation [59] with the number of neighbors set to 10 as this is an appropriate number for our use case. We consider a prediction to be a true positive if the correct journal appears within our reported list of neighbors and evaluate our performance using 10-fold cross-validation on the training set along with test set evaluation.

Web Application for Discovering Similar Preprints and Journals

We developed a web application that places any bioRxiv or medRxiv preprint into the overall document landscape and identifies similar papers and journals. The application downloads a pdf version of any preprint hosted on the bioRxiv or medRxiv server uses PyMuPDF [60] to extract text from the downloaded pdf and feeds the extracted text into our CBOW model to construct a document embedding representation. We pass this representation onto our journal and manuscript search to identify journals based on the ten closest neighbors of individual papers and journal centroids. We implemented this search using the scikit-learn implementation of k-d trees. To run it more cost-

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

effectively in a cloud computing environment with limited available memory, we sharded the k-d trees into four trees.

The app provides a visualization of the article's position within our training data to illustrate the local publication landscape, We used SAUCIE [61], an autoencoder designed to cluster single-cell RNA-seq data, to build a two-dimensional embedding space that could be applied to newly generated preprints without retraining, a limitation of other approaches that we explored for visualizing entities expected to lie on a nonlinear manifold. We trained this model on document embeddings of PMC articles that did not contain a matching preprint version. We used the following parameters to train the model: a hidden size of 2, a learning rate of 0.001, lambda_b of 0, lambda_c of 0.001, and lambda_d of 0.001 for 5000 iterations. When a user requests a new document, we can then project that document onto our generated two-dimensional space; thereby, allowing the user to see where their preprint falls along the landscape. We illustrate our recommendations as a shortlist and provide access to our network visualization at our website (see Software and Data Availability).

Analysis of the Preprints in Motion Collection

Our manuscript describes the large-scale analysis of bioRxiv. Concurrent with our work, another set of authors performed a detailed curation and analysis of a subset of bioRxiv [25] that was focused on preprints posted during the initial stages of the COVID-19 pandemic. The curated analysis was designed to examine preprints at a time of increased readership [62] and includes certain preprints posted from January 1st, 2020 to April 30th, 2020 [25]. We sought to contextualize this subset, which we term "Preprints in Motion" after the title of the preprint [25], within our global picture of the bioRxiv preprint landscape. We extracted all preprints from the set reported in Preprints in Motion [25] and retained any entries in the bioRxiv repository. We manually downloaded the XML version of these preprints and mapped them to their published counterparts as described above. We used Pubmed Central's DOI converter [63] to map the published article DOIs with their respective PubMed Central IDs. We retained articles that were included in the PMCOA corpus and performed a token analysis as described to compare these preprints with their published versions. As above, we generated document embeddings for every obtained preprint and published article. We projected these preprint embeddings onto our publication landscape to visually observe the dispersion of this subset. Finally, we performed a time analysis that paralleled our approach for the full set of preprint-publication pairs to examine relationships between linguistic changes and the time to publication.

Results

Comparing bioRxiv to other corpora

bioRxiv Metadata Statistics

The preprint landscape is rapidly changing, and the number of bioRxiv preprints in our data download (71,118) was nearly double that of a recent study that reported on a snapshot with 37,648 preprints [13]. Because the rate of change is rapid, we first analyzed category data and compared our results with previous findings. As in previous reports [13], neuroscience remains the most common category of preprints, followed by bioinformatics (Supplemental Figure S1). Microbiology, which was fifth in the most recent report [13], has now surpassed evolutionary biology and genomics to move into third. When authors upload their preprints, they select from three result category types: new results, confirmatory results, or contradictory results. We found that nearly all preprints (97.5%) were categorized as new results, consistent with reports on a smaller set [64]. The results taken together suggest that while bioRxiv has experienced dramatic growth, how it is being used appears to have remained consistent in recent years.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Global analysis reveals similarities and differences between bioRxiv and PMC

Table 1: Summary statistics for the bioRxiv, PMC, and NYTAC corpora.

Metric	bioRxiv	PMC	NYTAC
document count	71,118	1,977,647	1,855,658
sentence count	22,195,739	480,489,811	72,171,037
token count	420,969,930	8,597,101,167	1,218,673,384
stopword count	158,429,441	3,153,077,263	559,391,073
avg. document length	312.10	242.96	38.89
avg. sentence length	22.71	21.46	19.89
negatives	1,148,382	24,928,801	7,272,401
coordinating conjunctions	14,295,736	307,082,313	38,730,053
coordinating conjunctions%	3.40%	3.57%	3.18%
pronouns	4,604,432	74,994,125	46,712,553
pronouns%	1.09%	0.87%	3.83%
passives	15,012,441	342,407,363	19,472,053
passive%	3.57%	3.98%	1.60%

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

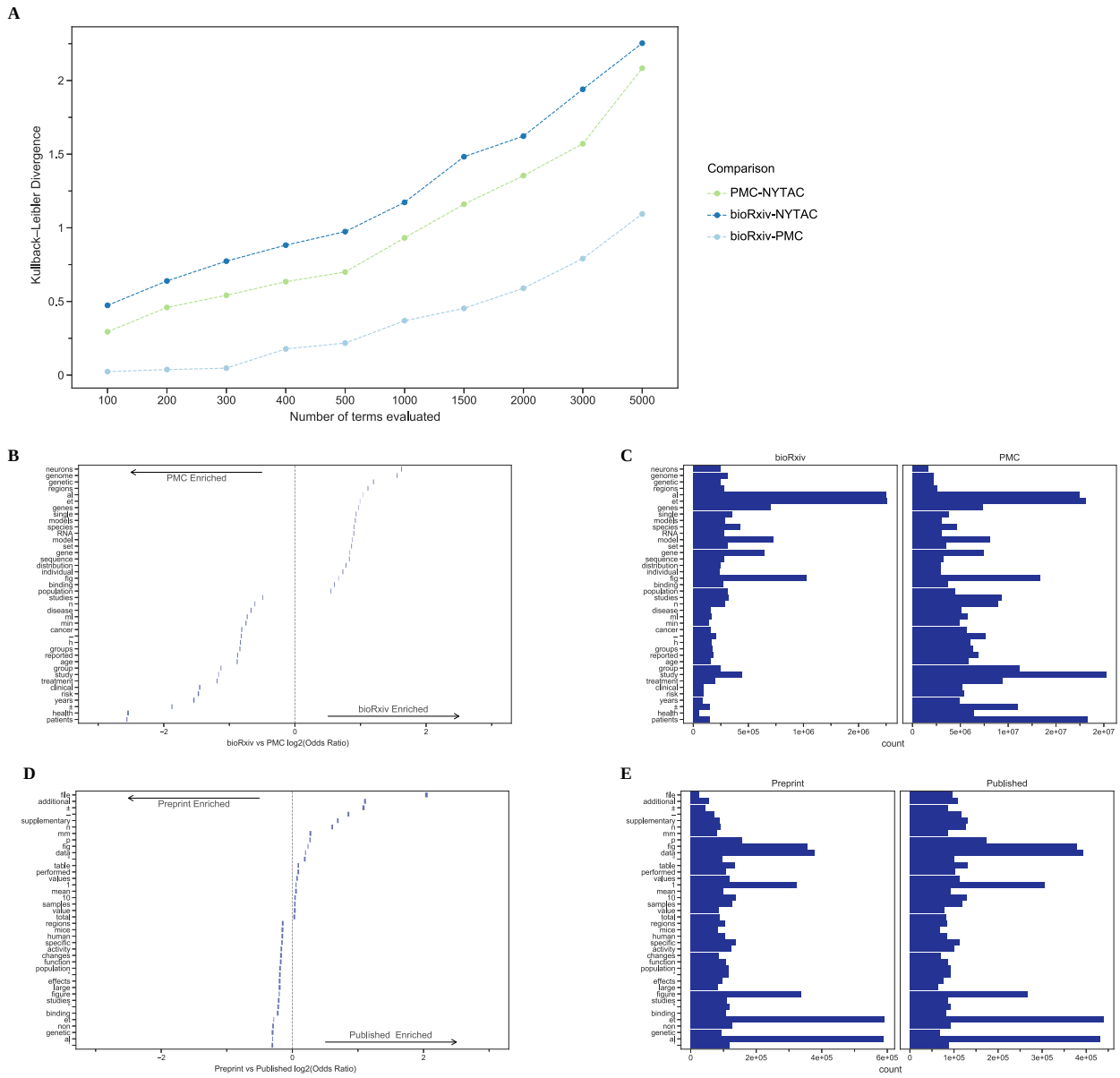


Figure 1: A. The Kullback–Leibler divergence measures the extent to which the distributions, not specific tokens, differ from each other. The token distribution of bioRxiv and PMC corpora is more similar than these biomedical corpora are to the NYTAC one. **B.** The significant differences in token frequencies for the corpora appear to be driven by the fields with the highest uptake of bioRxiv, as terms from neuroscience and genomics are relatively more abundant in bioRxiv. We plotted the 95% confidence interval for each reported token. **C.** Of the tokens that differ between bioRxiv and PMC, the most abundant in bioRxiv are "et" and "al" while the most abundant in PMC is "study." **D.** The significant differences in token frequencies for preprints and their corresponding published version often appear to be associated with typesetting and supplementary or additional materials. We plotted the 95% confidence interval for each reported token. **E.** The tokens with the largest absolute differences in abundance appear to be stylistic.

Documents within bioRxiv were slightly longer than those within PMCOA, but both were much longer than those from the control (NYTAC) (Table 1). The average sentence length, the fraction of pronouns, and the use of the passive voice were all more similar between bioRxiv and PMC than they were to NYTAC(Table 1). The Kullback–Leibler (KL) divergence of term frequency distributions between bioRxiv and PMCOA were low, especially among the top few hundred tokens (Figure 1A). As more tokens were incorporated, the KL divergence started to increase but remained much lower than the biomedical corpora compared against NYTAC. These findings support our notion that bioRxiv is linguistically similar to the PMCOA repository.

Terms like "neurons", "genome", and "genetic", which are common in genomics and neuroscience, were more common in bioRxiv than PMCOA while others associated with clinical research, such as "clinical" "patients" and "treatment" were more common in PMCOA (Figure 1B and 1C). When

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

controlling for the differences in the body of documents to identify textual changes associated with the publication process, we found that tokens such as “et” “al” were enriched for bioRxiv while “±”, “–” were enriched for PMCOA (Figure 1D and 1E). Furthermore, we found that specific changes appeared to be related to journal styles: “figure” was more common in bioRxiv while “fig” was relatively more common in PMCOA. Other changes appeared to be associated with an increasing reference to content external to the manuscript itself: the tokens “supplementary”, “additional” and “file” were all more common in PMCOA than bioRxiv, suggesting that journals are not simply replacing one token with another but that there are more mentions of such content after peer review.

These results taken together suggest that the structure of the text within preprints on bioRxiv are similar to published articles within PMCOA. The differences in uptake across fields are supported by differences in authors’ categorization of their articles and by the text within the articles themselves. At the level of individual manuscripts, the terms that change the most appear to be associated with typesetting, journal style, and an increasing reliance on additional materials after peer review.

Document embeddings derived from bioRxiv reveal fields and subfields

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

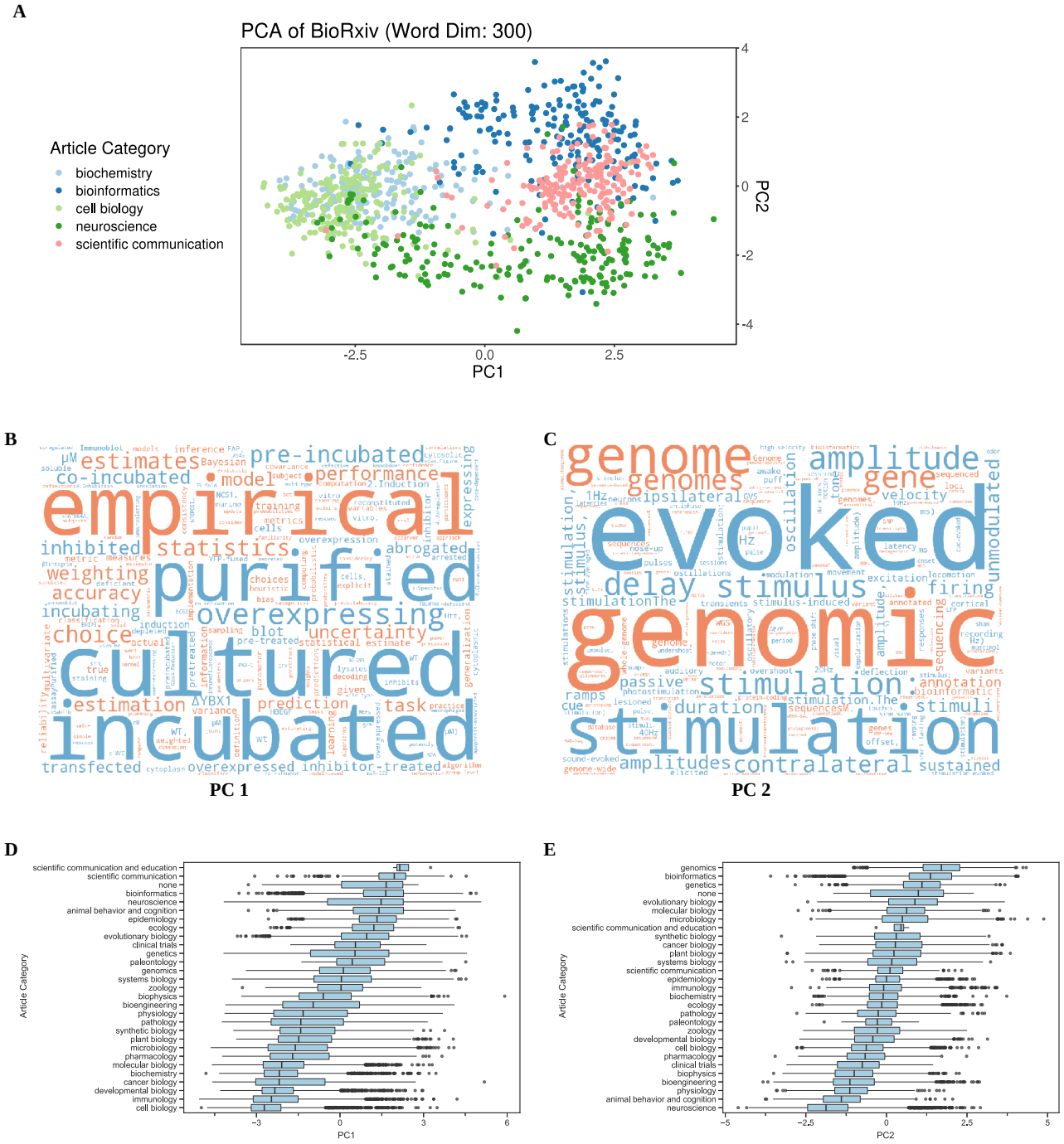


Figure 2: **A.** Principal components (PC) analysis of bioRxiv word2vec embeddings groups documents based on author-selected categories. We visualized documents from key categories on a scatterplot for the first two PCs. The first PC separated cell biology from informatics-related fields, and the second PC separated bioinformatics from neuroscience fields. **B.** A word cloud visualization of PC1. Each word cloud depicts the cosine similarity score between tokens and the first PC. Tokens in orange were most similar to the PC's positive direction, while tokens in blue were most similar to the PC's negative direction. The size of each token indicates the magnitude of the similarity. **C.** A word cloud visualization of PC2, which separated bioinformatics from neuroscience. Similar to the first PC, tokens in orange were most similar to the PC's positive direction, while tokens in blue were most similar to the PC's negative direction. The size of each token indicates the magnitude of the similarity. **D.** Examining PC1 values for each article by category created a continuum from informatics-related fields on the top through cell biology on the bottom. Specific article categories (neuroscience, genetics) were spread throughout PC1 values. **E.** Examining PC2 values for each article by category revealed fields like genomics, bioinformatics, and genetics on the top and neuroscience and behavior on the bottom.

Document embeddings provide a means to categorize the language of documents in a way that takes into account the similarities between terms [36,65,66]. We found that the first two PCs separated articles from different author-selected categories (Figure 2A). Certain neuroscience papers appeared

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

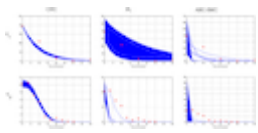
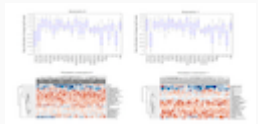
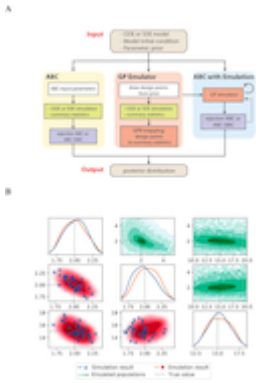
to be more associated with the cellular biology direction of PC1, while others seemed to be more associated with the informatics-related direction Figure 2A). This suggests that the concepts captured by PCs were not exclusively related to their field.

Visualizing token-PC similarity revealed tokens associated with certain research approaches (Figures 2B and 2C). Token association of PC1 shows the separation of cell biology and informatics-related fields through tokens: “empirical”, “estimates” and “statistics” depicted in orange and “cultured” and “overexpressing” shown in blue (Figure 2B). Association for PC2 shows the separation of bioinformatics and neuroscience via tokens: “genomic”, “genome” and “genomes” depicted in orange and “evoked”, “stimulus” and “stimulation” shown in blue (Figure 2C).

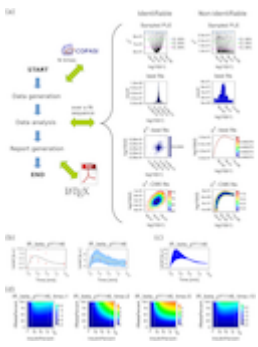
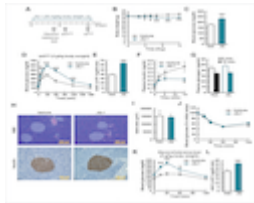
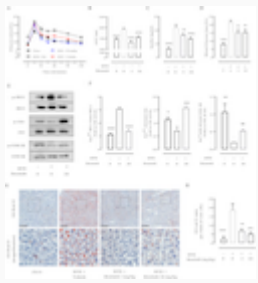
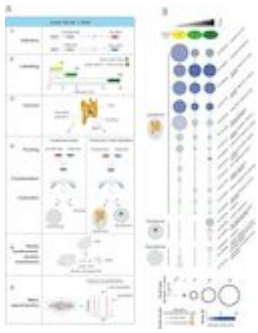
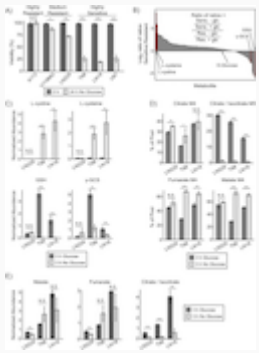
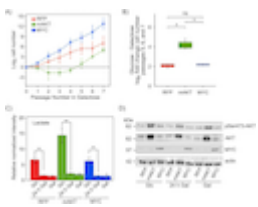
Examining the value for PC1 across all author-selected categories revealed an ordering of fields from cell biology to informatics-related disciplines (Figure 2D). These results suggest that a primary driver of the variability within the language used in bioRxiv could be the divide between informatics and cell biology approaches. A similar analysis for PC2 suggested that neuroscience and bioinformatics present a similar language continuum (Figure 2E). This result supports the notion that bioRxiv contains an influx of neuroscience and bioinformatics-related research results. For both of the top two PCs, the submitter-selected category of systems biology preprints was near the middle of the distribution and had a relatively large interquartile range when compared with other categories (Figure 2D and 2E), suggesting that systems biology is a broader subfield containing both informatics and cellular biology approaches.

Examining the top five and bottom five preprints within the systems biology field reinforces PC1’s dichotomous theme (Table 2). Preprints with the highest values [67,68,69,70,71] included software packages, machine learning analyses, and other computational biology manuscripts, while preprints with the lowest values [72,73,74,75,76] were focused on cellular signaling and protein activity. We provide the rest of our 50 generated PCs in our online repository (see Software and Data Availability).

Table 2: PC1 divided the author-selected category of systems biology preprints along an axis from computational to molecular approaches.

Title [citation]	PC1	License	Figure Thumbnail
Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis [67]	4.522818390064091	None	
FPtool a software tool to obtain in silico genotype-phenotype signatures and fingerprints based on massive model simulations [77]	4.348956760251298	CC-BY	
GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation [70]	4.259104249060651	CC-BY-NC-ND	
Notions of similarity for computational biology models [69]	4.079855550647664	CC-BY-NC-ND	

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Title [citation]	PC1	License	Figure Thumbnail
SBpipe: a collection of pipelines for automating repetitive simulation and analysis tasks [71]	4.022240241143516	CC-BY-NC-ND	
Bromodomain inhibition reveals FGF15/19 as a target of epigenetic regulation and metabolic control [78]	-3.4783803547922414	None	
Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease [75]	-3.6926161167521476	None	
Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection [72]	-3.728443135960558	CC-BY-ND	
NADPH consumption by L-cystine reduction creates a metabolic vulnerability upon glucose deprivation [74]	-3.7363965062637288	None	
AKT but not MYC promotes reactive oxygen species-mediated cell death in oxidative culture [76]	-3.8769231933681176	None	

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Document embedding similarities reveal unannotated preprint-publication pairs

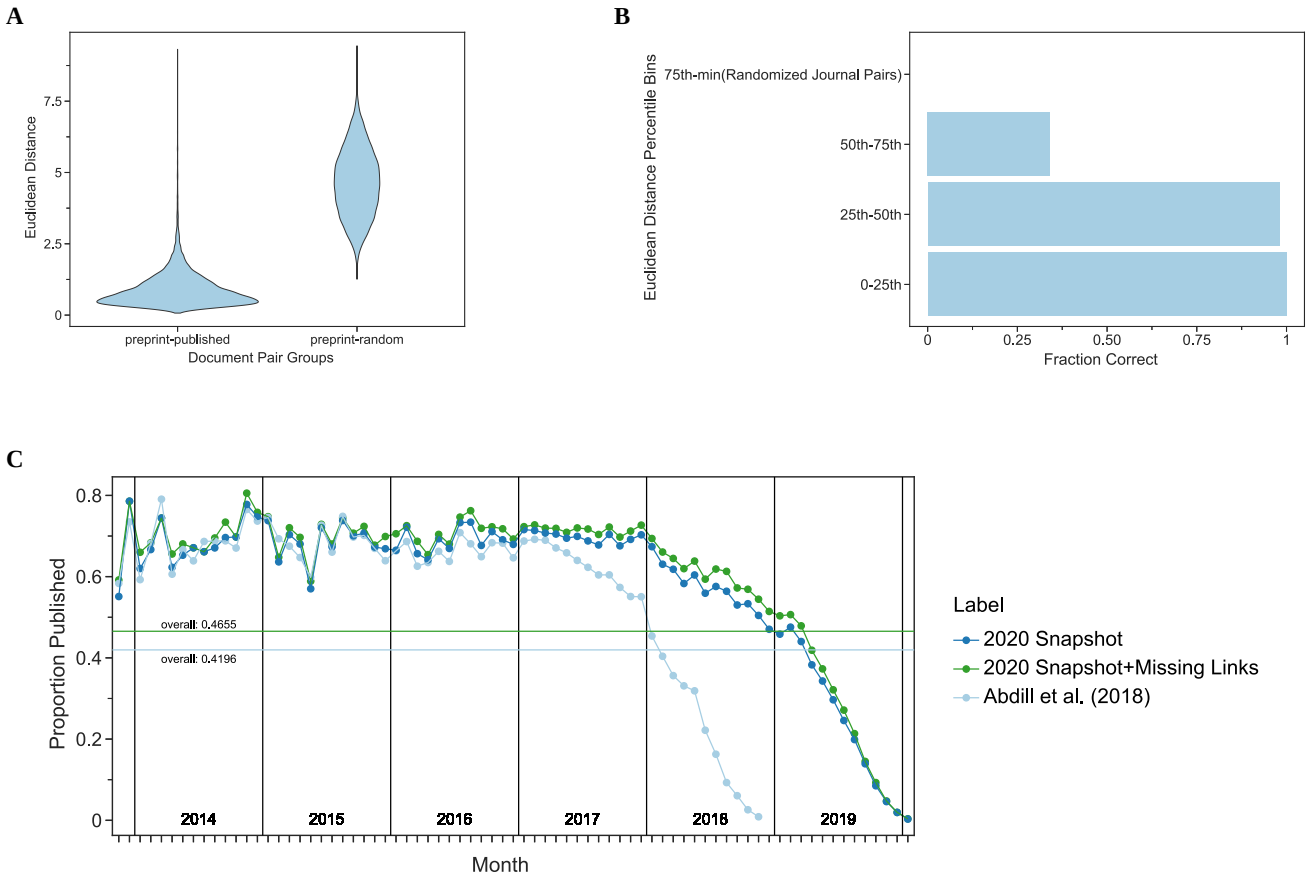


Figure 3: **A.** Preprints are closer in document embedding space to their corresponding peer-reviewed publication than they are to random papers published in the same journal. **B.** Potential preprint-publication pairs that are unannotated but within the 50th percentile of all preprint-publication pairs in the document embedding space are likely to represent true preprint-publication pairs. We depict the fraction of true positives over the total number of pairs in each bin. Accuracy is derived from the curation of a randomized list of 200 potential pairs (50 per quantile) performed in duplicate with a third rater used in the case of disagreement. **C.** Most preprints are eventually published. We show the publication rate of preprints since bioRxiv first started. The x-axis represents months since bioRxiv started, and the y-axis represents the proportion of preprints published given the month they were posted. The light blue line represents the publication rate previously estimated by Abdill et al. [13]. The dark blue line represents the updated publication rate using only CrossRef-derived annotations, while the dark green line includes annotations derived from our embedding space approach. The horizontal lines represent the overall proportion of preprints published as of the time of the annotation snapshot.

Distances between preprints and their corresponding published versions were nearly always lower than preprints paired with a random article published in the same journal (Figure 3A). This suggests that embedding distances can identify documents with similar textual content. Approximately 98% of our 200 pairs with an embedding distance in the 0-25th and 25th-50th percentile bins were scored as true matches (Figure 3B). These two bins contained 1,542 preprint-article pairs, suggesting that many preprints may have been published but not previously connected with their published versions. There is a particular enrichment for preprints published but unlinked within the 2017-2018 interval (Figure 3C). We expected a higher proportion of such preprints before 2019 (many of which may not have been published yet); however, observing relatively few missed annotations before 2017 was against our expectations. There are several possible explanations for this increasing fraction of missed annotations. As the number of preprints posted on bioRxiv grows, it may be harder for bioRxiv to establish a link between preprints and their published counterparts simply due to the scale of the challenge. It is possible that the set of authors participating in the preprint ecosystem is changing and that new participants may be less likely to report missed publications to bioRxiv. Finally, as familiarity

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with preprinting grows, it is possible that authors are posting preprints earlier in the process and that metadata fields that bioRxiv uses to establish a link may be less stable.

Preprints with more versions or more text changes took longer to publish

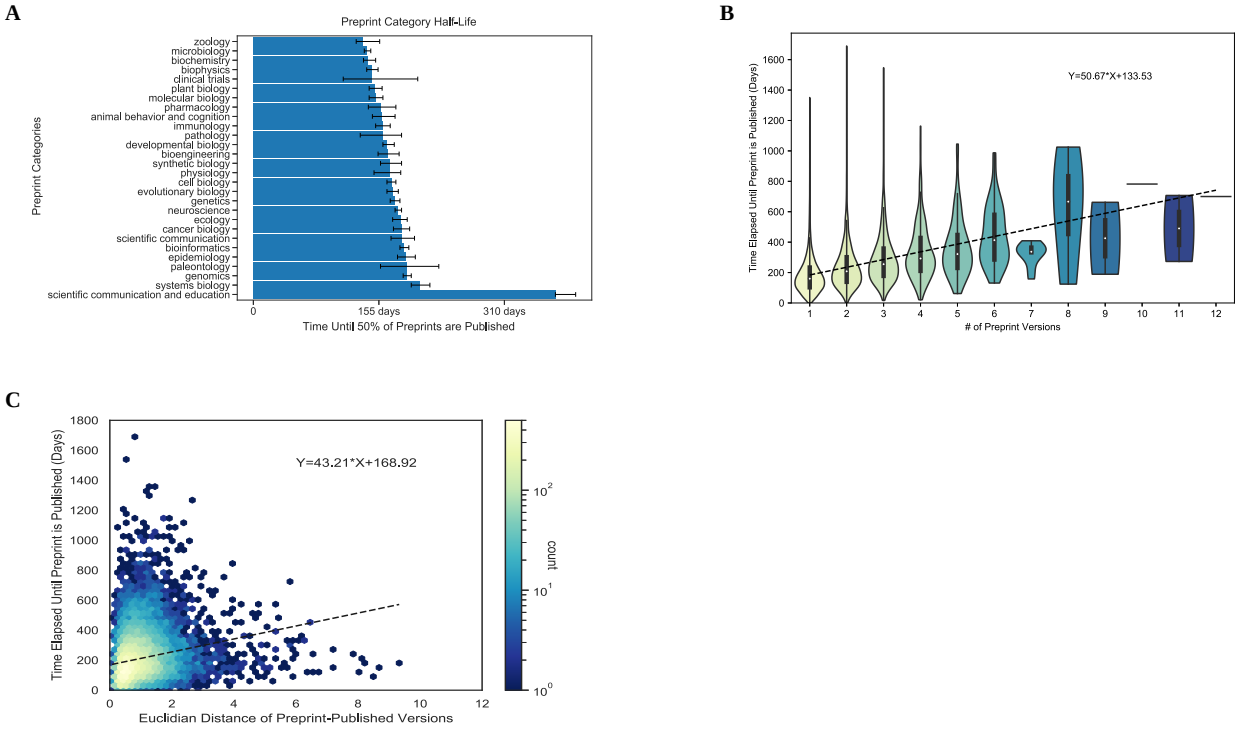


Figure 4: **A.** Author-selected categories were associated with modest differences in respect to publication half-life. Author-selected preprint categories are shown on the y-axis, while the x-axis shows the median time-to-publish for each category. Error bars represent 95% confidence intervals for each median measurement. **B.** Preprints with more versions were associated with a longer time to publish. The x-axis shows the number of versions of a preprint posted on bioRxiv. The y-axis indicates the number of days that elapsed between the first version of a preprint posted on bioRxiv and the date at which the peer-reviewed publication appeared. The density of observations is depicted in the violin plot with an embedded boxplot. **C.** Preprints with more substantial text changes took longer to be published. The x-axis shows the Euclidean distance between document representations of the first version of a preprint and its peer-reviewed form. The y-axis shows the number of days elapsed between the first version of a preprint posted on bioRxiv and when a preprint is published. The color bar on the right represents the density of each hexbin in this plot, where more dense regions are shown in a brighter color.

The process of peer review includes several steps, which take variable amounts of time [79], and we sought to measure if there is a difference in publication time between author-selected categories of preprints (Figure 4A). Of the most abundant preprint categories microbiology was the fastest to publish (140 days, (137, 145 days) [95% CI]) and genomics was the slowest (190 days, (185, 195 days) [95% CI]) (Figure 4A). We did observe category-specific differences; however, these differences were generally modest, suggesting that the peer review process did not differ dramatically between preprint categories. One exception was the Scientific Communication and Education category, which took substantially longer to be peer-reviewed and published (373 days, (373, 398 days) [95% CI]). This hints that there may be differences in the publication or peer review process or culture that apply to preprints in this category.

Examining peer review's effect on individual preprints, we found a positive correlation between preprints with multiple versions and the time elapsed until publication (Figure 4B). Each new version adds additional 51 days before a preprint is published. This time duration seems broadly compatible with the amount of time it would take to receive reviews and revise a manuscript, suggesting that many authors may be updating their preprints in response to peer reviews or other external

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

feedback. The embedding space allows us to compare preprint and published documents to determine if the level of change that documents undergo relates to the time it takes them to be published. Distances in this space are arbitrary and must be compared to reference distances. We found that the average distance of two randomly selected papers from the bioinformatics category was 4.470, while the average distance of two randomly selected papers from bioRxiv was 5.343. Preprints with large embedding space distances from their corresponding peer-reviewed publication took longer to publish (Figure 4C): each additional unit of distance corresponded to roughly forty-three additional days.

Overall, our findings support a model where preprints are reviewed multiple times or require more extensive revisions take longer to publish.

Preprints with similar document embeddings share publication venues

We developed an online application that returns a listing of published papers and journals closest to a query preprint in document embedding space. This application uses two k-nearest neighbor classifiers that achieved better performance than our baseline model (Supplemental Figure S2) to identify these entities. Users supply our app with digital object identifiers (DOIs) from bioRxiv or medRxiv, and the corresponding preprint is downloaded from the repository. Next, the preprint's PDF is converted to text, and this text is used to construct a document embedding representation. This representation is supplied to our classifiers to generate a listing of the ten papers and journals with the most similar representations in the embedding space (Figures 5A, 5B and 5C). Furthermore, the user-requested preprint's location in this embedding space is then displayed on our interactive map, and users can select regions to identify the terms most associated with those regions (Figures 5D and 5E). Users can also explore the terms associated with the top 50 PCs derived from the document embeddings, and those PCs vary across the document landscape.

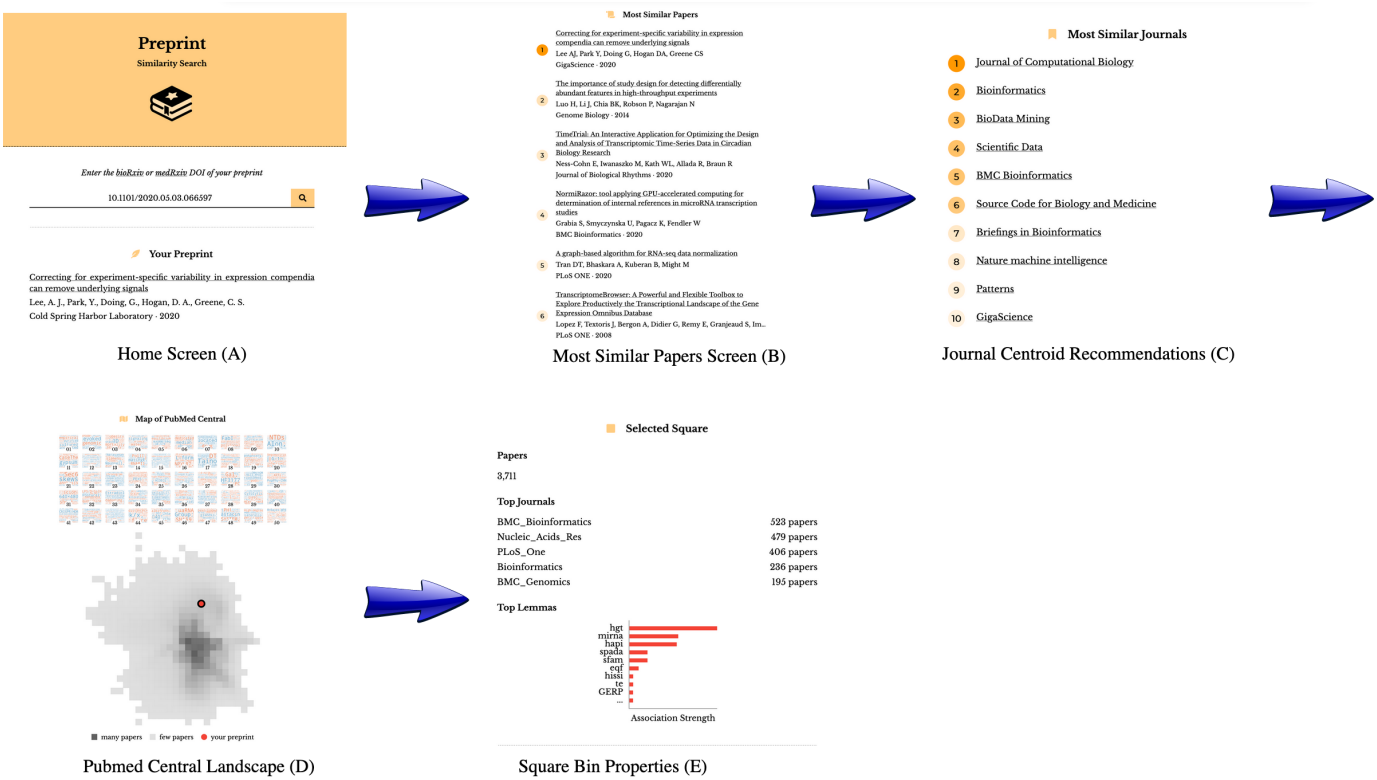


Figure 5: The preprint-similarity-search app workflow allows users to examine where an individual preprint falls in the overall document landscape. **A.** Starting with the home screen, users can paste in a bioRxiv or medRxiv DOI, which sends a request to bioRxiv or medRxiv. Next, the app preprocesses the requested preprint and returns a listing of (**B**) the top ten most similar papers and (**C**) the ten closest journals. **D.** The app also displays the location of the query

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

preprint in PMC. **E.** Users can select a square within the landscape to examine statistics associated with the square, including the top journals by article count in that square and the odds ratio of tokens.

Contextualizing the Preprints in Motion Collection

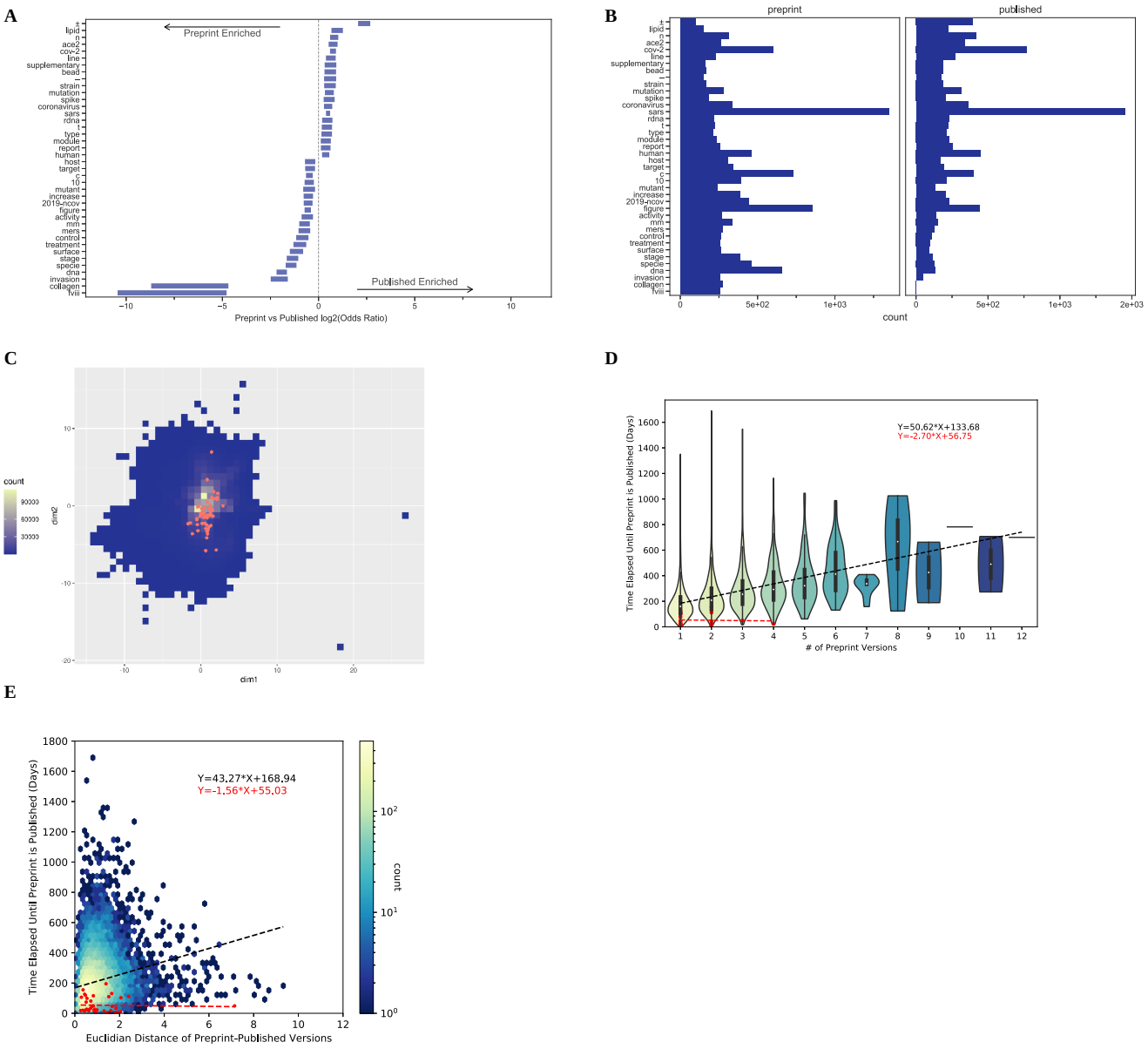


Figure 6: The Preprints in Motion Collection results are similar to all preprint results, except that their time to publication was independent of the number of preprint versions and amount of linguistic change. **A.** Tokens that differed included those associated with typesetting and those related to the nomenclature of the virus that causes COVID-19. Error bars show 95% confidence intervals for each token. **B.** Of the tokens that differ between Preprints in Motion and their published counterparts, the most abundant were associated with the nomenclature of the virus. **C.** The Preprints in Motion fall across the landscape of PMCOA with respect to linguistic properties. This square bin plot depicts the binning of all published papers within the PMCOA corpus. High-density regions are depicted in yellow, while low-density regions are in dark blue. Red dots represent the Preprints in Motion Collection. **D.** The Preprints in Motion were published faster than other bioRxiv preprints, and the number of versions was not associated with an increase in time to publication. The x-axis shows the number of versions of a preprint posted on bioRxiv. The y-axis indicates the number of days that elapsed between the first version of a preprint posted on bioRxiv and the date at which the peer-reviewed publication appeared. The density of observations is depicted in the violin plot with an embedded boxplot. The red dots and red regression line represent Preprints in Motion. **D.** The Preprints in Motion were published faster than other bioRxiv preprints, and no dependence between the amount of linguistic change and time to publish was observed. The x-axis shows the Euclidean distance between document representations of the first version of a preprint and its peer-reviewed form. The y-axis shows the number of days elapsed between the first version of a preprint posted on bioRxiv and when a preprint is published. The color bar on the right represents the density of each hexbin in this

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

plot, where more dense regions are shown in a brighter color. The red dots and red regression line represent Preprints in Motion.

The Preprints in Motion collection included a set of preprints posted during the first four months of 2020. We examined the extent to which preprints in this set were representative of the patterns that we identified from our analysis on all of bioRxiv. As with all of bioRxiv, typesetting tokens changed between preprints and their paired publications. Our token-level analysis identified certain patterns consistent with our findings across bioRxiv (Figure 6A and 6B). However, in this set, we also observe changes likely associated with the fast-moving nature of COVID-19 research: the token "2019-ncov" became less frequently represented while "sars" and "cov-2" became more represented, likely due to a shift in nomenclature from "2019-nCoV" to "SARS-CoV-2". The Preprints in Motion were not strongly colocalized in the linguistic landscape, suggesting that the collection covers a diverse set of research approaches (Figure 6C). Preprints in this collection were published faster than the broader set of bioRxiv preprints (Figure 6D and 6E). The relationship between time to publication and the number of versions (Figure 6D) and the relationship between time to publication and the amount of linguistic change (Figure 6E) were both lost in the Preprints in Motion set. Our findings suggest that Preprints in Motion changed during publication in ways aligned with changes in the full preprint set but that peer review was accelerated in ways that broke the time dependences observed with the full bioRxiv set.

Discussion and Conclusions

BioRxiv is a constantly growing repository that contains life science preprints. The majority of research involving bioRxiv focuses on the metadata of preprints; however, the language contained within these preprints has not previously been systematically examined. Throughout this work, we sought to analyze the language within these preprints and understand how it changes in response to peer review. Our global corpora analysis found that writing within bioRxiv is consistent with the biomedical literature in the PMCOA repository, suggesting that bioRxiv is linguistically similar to PMCOA. Token-level analyses between bioRxiv and PMCOA suggested that research fields drive significant differences; e.g., more patient-related research is prevalent in PMCOA than bioRxiv. This observation is expected as preprints focused on medicine are supported by the complementary medRxiv repository [8]. Token-level analyses for preprints and their corresponding published version suggest that peer review may focus on data availability and incorporating extra sections for published papers; however, future studies are needed to ascertain individual token level changes as preprints venture through the publication process.

Document embeddings are a versatile way to examine language contained within preprints, understanding peer review's effect on preprints, and provide extra functionality for preprint repositories. Examining linguistic variance within document embeddings of life science preprints revealed that the largest source of variability was informatics. This observation bisects the majority of life science research categories that have integrated preprints within their publication workflow. Preprints are typically linked with their published articles via bioRxiv manually establishing links or authors self-reporting that their preprint has been published; however, gaps can occur as preprints change their appearance through multiple versions or authors do not notify bioRxiv. Our work suggests that document embeddings can help fill in missing links within bioRxiv. Furthermore, our analysis reveals that the publication rate for preprints is higher than previously estimated, even though our analysis can only account for published open access papers. Our results raise the lower bound of the total preprint publication fraction; however, the true fraction is necessarily higher. Future work, especially that which aims to assess the fraction of preprints that are eventually published, should account for the possibility of missed annotations.

Preprints take a variable amount of time to become published, and we examined factors that influence a preprint's time to publication. Our half-life analysis on preprint categories revealed that preprints in most bioRxiv categories take similar amounts of time to be published. An apparent

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

exception is the scientific communication and education category, which contained preprints that took much longer to publish. Regarding individual preprints, each new version adds several weeks to a preprints time to publication, which is roughly aligned with authors making changes after a round of peer review; furthermore, preprints that undergo substantial changes take longer to publish. Overall, these results illustrate that bioRxiv is a practical resource for obtaining insight into the peer-review process.

Lastly, we found that document embeddings were associated with the eventual journal at which the work was published. We trained two machine learning models to identify which journals publish linguistically similar papers towards a query preprint. Our models achieved a considerably higher fold change over the baseline model, so we constructed a web application that makes our models available to the public and returns a list of the papers and journals that are linguistically similar to a bioRxiv or medRxiv preprint.

Software and Data Availability

An online version of this manuscript is available under a Creative Commons Attribution License at https://greenelab.github.io/annorxiver_manuscript/. Source for the research portions of this project is dual licensed under the BSD 3-Clause and Creative Commons Public Domain Dedication Licenses at https://github.com/greenelab/annorxiver. The preprint similarity search website can be found at https://greenelab.github.io/preprint-similarity-search/, and code for the website is available under a BSD-2-Clause Plus Patent License at https://github.com/greenelab/preprint-similarity-search. Full text access for the bioRxiv repository is available at https://www.biorxiv.org/tdm. Access to PubMed Central's Open Access subset is available on NCBI's FTP server at https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/. Access to the New York Times Annotated Corpus (NYTAC) is available upon request with the Linguistic Data Consortium at https://catalog.ldc.upenn.edu/LDC2008T19.

Acknowledgments

The authors would like to thank Ariel Hippen Anderson for evaluating potential missing preprint to published version links. We also would like to thank Richard Sever and the bioRxiv team for their assistance with access to and support with questions about preprint full text downloaded from bioRxiv.

Funding

This work was supported by grants from the Gordon Betty Moore Foundation (GBMF4552) and the National Institutes of Health's National Human Genome Research Institute (NHGRI) under awards T32 HG00046 and R01 HG010067.

Competing Interests

Marvin Thielk receives a salary from Elsevier Inc. where he contributes NLP expertise to health content operations. Elsevier did not restrict the results or interpretations that could be published in this manuscript. The opinions expressed here do not reflect the official policy or positions of Elsevier Inc.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. **Scientific communication pathways: an overview and introduction to a symposium**
David F. Zaye, W. V. Metanomski
Journal of Chemical Information and Computer Sciences (2002-05-01) https://doi.org/bwsxhg
DOI: 10.1021/ci00050a001

2. **The trouble with medical journals**
Richard Smith
Journal of the Royal Society of Medicine (2006)

3. **The Transition from Paper to Electronic Journals**
Hak Joon Kim
The Serials Librarian (2001-11-19) https://doi.org/d7rnh2
DOI: 10.1300/j123v41n01_04

4. **Preprints: What Role Do These Have in Communicating Scientific Results?**
Susan A. Elmore
Toxicologic Pathology (2018-04-08) https://doi.org/ghdd7c
DOI: 10.1177/0192623318767322 · PMID: 29628000 · PMCID: PMC5999550

5. **The prehistory of biology preprints: A forgotten experiment from the 1960s**
Matthew Cobb
PLOS Biology (2017-11-16) https://doi.org/c6wv
DOI: 10.1371/journal.pbio.2003995 · PMID: 29145518 · PMCID: PMC5690419

6. **arXiv.org: the Los Alamos National Laboratory e-print server**
Gerry McKiernan
International Journal on Grey Literature (2000-09) https://doi.org/fg8pw7
DOI: 10.1108/14666180010345564

7. **bioRxiv: the preprint server for biology**
Richard Sever, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, John R. Inglis
Cold Spring Harbor Laboratory (2019-11-06) https://doi.org/ggc46z
DOI: 10.1101/833400

8. **medRxiv.org - the preprint server for Health Sciences** https://www.medrxiv.org/

9. **The Second Wave of Preprint Servers: How Can Publishers Keep Afloat?**
By
The Scholarly Kitchen (2019-10-16) https://scholarlykitchen.sspnet.org/2019/10/16/the-second-wave-of-preprint-servers-how-can-publishers-keep-afloat/

10. **Rxivist.org: Sorting biology preprints using social media and readership metrics**
Richard J. Abdill, Ran Blekhman
PLOS Biology (2019-05-21) https://doi.org/dm27
DOI: 10.1371/journal.pbio.3000269 · PMID: 31112533 · PMCID: PMC6546241

11. **How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations**

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Xin Shuai, Alberto Pepe, Johan Bollen
PLoS ONE (2012-11-01) https://doi.org/f4cw6r
DOI: 10.1371/journal.pone.0047523 · PMID: 23133597 · PMCID: PMC3486871

12. **Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation**
Jedidiah Carlson, Kelley Harris
Cold Spring Harbor Laboratory (2020-03-10) https://doi.org/ghdd66
DOI: 10.1101/2020.03.06.981589

13. **Tracking the popularity and outcomes of all bioRxiv preprints**
Richard J Abdill, Ran Blekhman
eLife (2019-04-24) https://doi.org/gf2str
DOI: 10.7554/elife.45133 · PMID: 31017570 · PMCID: PMC6510536

14. **An analysis of published journals for papers posted on bioR $_{X}$ iv**
Hiroyuki Tsunoda, Yuan Sun, Masaki Nishizawa, Xiaomin Liu, Kou Amano
Proceedings of the Association for Information Science and Technology (2019-10-18) https://doi.org/ggz7f9
DOI: 10.1002/pra2.175

15. **The relationship between bioRxiv preprints, citations and altmetrics**
Nicholas Fraser, Fakhri Momeni, Philipp Mayr, Isabella Peters
Quantitative Science Studies (2020-04-01) https://doi.org/gg2cz3
DOI: 10.1162/qss_a_00043

16. **Releasing a preprint is associated with more attention and citations for the peer-reviewed article**
Darwin Y Fu, Jacob J Hughey
eLife (2019-12-06) https://doi.org/ghd3mv
DOI: 10.7554/elife.52646 · PMID: 31808742 · PMCID: PMC6914335

17. **Preprints and Scholarly Communication: An Exploratory Qualitative Study of Adoption, Practices, Drivers and Barriers**
Andrea Chiarelli, Rob Johnson, Stephen Pinfield, Emma Richens
F1000Research (2019-11-25) https://doi.org/ghp38z
DOI: 10.12688/f1000research.19619.2 · PMID: 32055396 · PMCID: PMC6961415

18. **The Need for Speed: How Quickly Do Preprints Become Published Articles?**
Rachel Herbert, Kate Gasson, Alex Ponsford
SSRN Electronic Journal (2019) https://doi.org/ghd3mt
DOI: 10.2139/ssrn.3455146

19. **Technical and social issues influencing the adoption of preprints in the life sciences**
Naomi C. Penfold, Jessica K. Polka
PLOS Genetics (2020-04-20) https://doi.org/dtt2
DOI: 10.1371/journal.pgen.1008565 · PMID: 32310942 · PMCID: PMC7170218

20. **Biologists urged to hug a preprint**
Ewen Callaway, Kendall Powell
Nature (2016-02-16) https://doi.org/ghdd62
DOI: 10.1038/530265a · PMID: 26887471

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

21. **Day-to-day discovery of preprint–publication links**
Guillaume Cabanac, Theodora Oikonomidi, Isabelle Boutron
Scientometrics (2021-04-18) https://doi.org/gjr9k4
DOI: 10.1007/s11192-021-03900-7 · PMCID: PMC8053368

22. **On the value of preprints: An early career researcher perspective**
Sarvenaz Sarabipour, Humberto J. Debat, Edward Emmott, Steven J. Burgess, Benjamin Schwessinger, Zach Hensel
PLOS Biology (2019-02-21) https://doi.org/gfw8hd
DOI: 10.1371/journal.pbio.3000151 · PMID: 30789895 · PMCID: PMC6400415

23. **Prepublication Communication of Research Results**
Michael J. Adams, Reid N. Harris, Evan H. C. Grant, Matthew J. Gray, M. Camille Hopkins, Samuel A. Iverson, Robert Likens, Mark Mandica, Deanna H. Olson, Alex Shepack, Hardin Waddle
EcoHealth (2018-08-07) https://doi.org/ghn66s
DOI: 10.1007/s10393-018-1352-3 · PMID: 30088185 · PMCID: PMC6245104

24. **Peer Review and bioRxiv**
Leslie M. Loew
Biophysical Journal (2016-08) https://doi.org/ghdd6x
DOI: 10.1016/j.bpj.2016.06.035 · PMID: 27508451 · PMCID: PMC4982934

25. **Preprints in motion: tracking changes between posting and journal publication**
Jessica K Polka, Gautam Dey, Máté Pálfy, Federico Nanni, Liam Brierley, Nicholas Fraser, Jonathon Alexis Coates
Cold Spring Harbor Laboratory (2021-04-04) https://doi.org/gh5mhm
DOI: 10.1101/2021.02.20.432090

26. **Textual Analysis in Accounting and Finance: A Survey**
TIM LOUGHRAN, BILL MCDONALD
Journal of Accounting Research (2016-09) https://doi.org/gc3hf7
DOI: 10.1111/1475-679x.12123

27. **The textual characteristics of traditional and Open Access scientific journals are similar**
Karin Verspoor, K Bretonnel Cohen, Lawrence Hunter
BMC Bioinformatics (2009-06-15) https://doi.org/b973tn
DOI: 10.1186/1471-2105-10-183 · PMID: 19527520 · PMCID: PMC2714574

28. **Current findings from research on structured abstracts**
James Hartley
Journal of the Medical Library Association : JMLA (2004-07)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC442180/
PMID: 15243644 · PMCID: PMC442180

29. **A survey on annotation tools for the biomedical literature**
M. Neves, U. Leser
Briefings in Bioinformatics (2012-12-18) https://doi.org/f5vzsj
DOI: 10.1093/bib/bbs084 · PMID: 23255168

30. **PubTator central: automated concept annotation for biomedical full text articles**
Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu
Nucleic Acids Research (2019-07-02) https://doi.org/ggzfsc
DOI: 10.1093/nar/gkz389 · PMID: 31114887 · PMCID: PMC6602571

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

31. **Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles**
K. Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, Lawrence E. Hunter
BMC Bioinformatics (2017-08-17) https://doi.org/ghmbw2
DOI: 10.1186/s12859-017-1775-9 · PMID: 28818042 · PMCID: PMC5561560

32. **The structural and content aspects of abstracts versus bodies of full text journal articles are different**
K Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, Lawrence E Hunter
BMC Bioinformatics (2010-09-29) https://doi.org/b9f6rn
DOI: 10.1186/1471-2105-11-492 · PMID: 20920264 · PMCID: PMC3098079

33. **A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools**
Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L Johnson, Christophe Roeder, Jinho D Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, … Lawrence E Hunter
BMC Bioinformatics (2012-08-17) https://doi.org/gb8t7v
DOI: 10.1186/1471-2105-13-207 · PMID: 22901054 · PMCID: PMC3483229

34. **From POS tagging to dependency parsing for biomedical event extraction**
Dat Quoc Nguyen, Karin Verspoor
BMC Bioinformatics (2019-02-12) https://doi.org/ggsrkw
DOI: 10.1186/s12859-019-2604-0 · PMID: 30755172 · PMCID: PMC6373122

35. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-09-10) https://arxiv.org/abs/1301.3781

36. **Distributed Representations of Sentences and Documents**
Quoc V. Le, Tomas Mikolov
arXiv (2014-05-26) https://arxiv.org/abs/1405.4053

37. https://www.biorxiv.org/tdm

38. **PubMed Central: The GenBank of the published literature**
R. J. Roberts
Proceedings of the National Academy of Sciences (2001-01-16) https://doi.org/bbn9k8
DOI: 10.1073/pnas.98.2.381 · PMID: 11209037 · PMCID: PMC33354

39. **How Papers Get Into PMC** https://www.ncbi.nlm.nih.gov/pmc/about/submission-methods/

40. **Gold open access: the best of both worlds**
M. A. G. van der Heyden, T. A. B. van Veen
Netherlands Heart Journal (2017-12-01) https://doi.org/ggzfr9
DOI: 10.1007/s12471-017-1064-2 · PMID: 29196877 · PMCID: PMC5758455

41. **8.2.2 NIH Public Access Policy**
https://grants.nih.gov/grants/policy/nihgps/html5/section_8/8.2.2_nih_public_access_policy.htm

42. **PMC Overview** https://www.ncbi.nlm.nih.gov/pmc/about/intro/

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

43. **PMC text mining subset in BioC: about three million full-text articles and growing**
Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, Zhiyong Lu
Bioinformatics (2019-09-15) https://doi.org/ggzfsb
DOI: 10.1093/bioinformatics/btz070 · PMID: 30715220 · PMCID: PMC6748740

44. **Author Manuscripts in PMC** https://www.ncbi.nlm.nih.gov/pmc/about/authorms/

45. **The new york times annotated corpus**
Evan Sandhaus
Linguistic Data Consortium, Philadelphia (2008)

46. **CrossRef Text and Data Mining Services**
Rachael Lammey
Insights the UKSG journal (2015-07-07) https://doi.org/gg4hp9
DOI: 10.1629/uksg.233

47. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
(2017)

48. **Odds Ratio**
Steven Tenny, Mary R. Hoffman
StatPearls (2021) http://www.ncbi.nlm.nih.gov/books/NBK431098/

49. **Software Framework for Topic Modelling with Large Corpora**
Radim Řehůřek, Petr Sojka
Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (2010-05-22)

50. **On the Dimensionality of Word Embedding**
Zi Yin, Yuanyuan Shen
arXiv (2018-12-12) https://arxiv.org/abs/1812.04224

51. **Probabilistic Principal Component Analysis**
Michael E. Tipping, Christopher M. Bishop
Journal of the Royal Statistical Society: Series B (Statistical Methodology) (1999-08)
https://doi.org/b3hjwt
DOI: 10.1111/1467-9868.00196

52. **Scikit-learn: Machine learning in Python**
F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, ... E. Duchesnay
Journal of Machine Learning Research (2011)

53. **Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions**
Nathan Halko, Per-Gunnar Martinsson, Joel A. Tropp
arXiv (2014-04-29) https://arxiv.org/abs/0909.4061

54. **The *Drosophila* Cortactin Binding Protein 2 homolog, Nausicaa, regulates lamellipodial actin dynamics in a Cortactin-dependent manner**
Meghan E. O'Connell, Divya Sridharan, Tristan Driscoll, Ipsita Krishnamurthy, Wick G. Perry, Derek A. Applewhite

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cold Spring Harbor Laboratory (2018-07-24) https://doi.org/gg4hp7
DOI: 10.1101/376665

55. **The *Drosophila* protein, Nausicaa, regulates lamellipodial actin dynamics in a Cortactin-dependent manner**
Meghan E. O'Connell, Divya Sridharan, Tristan Driscoll, Ipsita Krishnamurthy, Wick G. Perry, Derek A. Applewhite
Biology Open (2019-06-15) https://doi.org/gg4hp8
DOI: 10.1242/bio.038232 · PMID: 31164339 · PMCID: PMC6602326

56. **Understanding survival analysis: Kaplan-Meier estimate**
Jugal Kishore, ManishKumar Goel, Pardeep Khanna
International Journal of Ayurveda Research (2010) https://doi.org/fdft75
DOI: 10.4103/0974-7788.76794 · PMID: 21455458 · PMCID: PMC3059453

57. **CamDavidsonPilon/lifelines: v0.25.6**
Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Sean-Reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, … Jlim13
Zenodo (2020-10-26) https://doi.org/ghh2d3
DOI: 10.5281/zenodo.4136578

58. **Medium – Where good ideas find you.**
Medium
https://medium.com

59. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, … Édouard Duchesnay
arXiv (2018-06-06) https://arxiv.org/abs/1201.0490

60. **Introduction — PyMuPDF 1.18.13 documentation**
https://pymupdf.readthedocs.io/en/latest/intro.html

61. **Assessing the Heterogeneity of Cardiac Non-myocytes and the Effect of Cell Culture with Integrative Single Cell Analysis**
Brian S. Iskra, Logan Davis, Henry E. Miller, Yu-Chiao Chiu, Alexander R. Bishop, Yidong Chen, Gregory J. Aune
Cold Spring Harbor Laboratory (2020-03-05) https://doi.org/gg9353
DOI: 10.1101/2020.03.04.975177

62. **Preprinting the COVID-19 pandemic**
Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K Polka, Máté Pálfy, Federico Nanni, Jonathon Alexis Coates
Cold Spring Harbor Laboratory (2021-02-05) https://doi.org/dxdb
DOI: 10.1101/2020.05.22.111294

63. **PMCID - PMID - Manuscript ID - DOI Converter** https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/

64. **Altmetric Scores, Citations, and Publication of Studies Posted as Preprints**
Stylianos Serghiou, John P. A. Ioannidis
JAMA (2018-01-23) https://doi.org/gftc69
DOI: 10.1001/jama.2017.21168 · PMID: 29362788 · PMCID: PMC5833561

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

65. **Efficient Vector Representation for Documents through Corruption**
Minmin Chen
arXiv (2017-07-11) https://arxiv.org/abs/1707.02377

66. **Document Network Projection in Pretrained Word Embedding Space**
Antoine Gourru, Adrien Guille, Julien Velcin, Julien Jacques
arXiv (2020-01-17) https://arxiv.org/abs/2001.05727

67. **Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis**
Fortunato Bianconi, Chiara Antonini, Lorenzo Tomassoni, Paolo Valigi
Cold Spring Harbor Laboratory (2017-10-02) https://doi.org/gg9393
DOI: 10.1101/197400

68. **Machine learning of stochastic gene network phenotypes**
Kyemyung Park, Thorsten Prüstel, Yong Lu, John S. Tsang
Cold Spring Harbor Laboratory (2019-10-31) https://doi.org/gg94bm
DOI: 10.1101/825943

69. **Notions of similarity for computational biology models**
Ron Henkel, Robert Hoehndorf, Tim Kacprowski, Christian Knüpfer, Wolfram Liebermeister, Dagmar Waltemath
Cold Spring Harbor Laboratory (2016-03-21) https://doi.org/gg939z
DOI: 10.1101/044818

70. **GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation**
Evgeny Tankhilevich, Jonathan Ish-Horowicz, Tara Hameed, Elisabeth Roesch, Istvan Kleijn, Michael PH Stumpf, Fei He
Cold Spring Harbor Laboratory (2019-09-18) https://doi.org/gg94bj
DOI: 10.1101/769299

71. **SBpipe: a collection of pipelines for automating repetitive simulation and analysis tasks**
Piero Dalle Pezze, Nicolas Le Novère
Cold Spring Harbor Laboratory (2017-02-09) https://doi.org/gg9392
DOI: 10.1101/107250

72. **Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection**
Joel Selkrig, Nan Li, Jacob Bobonis, Annika Hausmann, Anna Sueki, Haruna Imamura, Bachir El Debs, Gianluca Sigismondo, Bogdan I. Florea, Herman S. Overkleeft, ... Athanasios Typas
Cold Spring Harbor Laboratory (2018-11-07) https://doi.org/gg94bc
DOI: 10.1101/455048

73. **Systems analysis by mass cytometry identifies susceptibility of latent HIV-infected T cells to targeting of p38 and mTOR pathways**
Linda E. Fong, Victor L. Bass, Serena Spudich, Kathryn Miller-Jensen
Cold Spring Harbor Laboratory (2018-07-19) https://doi.org/gg9398
DOI: 10.1101/371922

74. **NADPH consumption by L-cystine reduction creates a metabolic vulnerability upon glucose deprivation**
James H. Joly, Alireza Delfarah, Philip S. Phung, Sydney Parrish, Nicholas A. Graham

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cold Spring Harbor Laboratory (2019-08-13) https://doi.org/gg94bf
DOI: 10.1101/733162

75. **Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease**
Gareth S. D. Purvis, Massimo Collino, Haidee M. A. Tavio, Fausto Chiazza, Caroline E. O'Riodan, Lynda Zeboudj, Nick Guisot, Peter Bunyard, David R. Greaves, Christoph Thiemermann
Cold Spring Harbor Laboratory (2019-08-28) https://doi.org/gg94bg
DOI: 10.1101/745943

76. **AKT but not MYC promotes reactive oxygen species-mediated cell death in oxidative culture**
Dongqing Zheng, Jonathan H. Sussman, Matthew P. Jeon, Sydney T. Parrish, Alireza Delfarah, Nicholas A. Graham
Cold Spring Harbor Laboratory (2019-09-01) https://doi.org/gg94bh
DOI: 10.1101/754572

77. ***FPtool* a software tool to obtain *in silico* genotype-phenotype signatures and fingerprints based on massive model simulations**
Guido Santos, Julio Vera
Cold Spring Harbor Laboratory (2018-02-18) https://doi.org/gjr9m9
DOI: 10.1101/266775

78. **Bromodomain inhibition reveals FGF15/19 as a target of epigenetic regulation and metabolic control**
Chisayo Kozuka, Vicencia Sales, Soravis Osataphan, Yixing Yuchi, Jeremy Chimene-Weiss, Christopher Mulla, Elvira Isganaitis, Jessica Desmond, Suzuka Sanechika, Joji Kusuyama, … Mary-Elizabeth Patti
Cold Spring Harbor Laboratory (2019-12-12) https://doi.org/gjr9m8
DOI: 10.1101/2019.12.11.872887

79. **Peer review and the publication process**
Parveen Azam Ali, Roger Watson
Nursing Open (2016-03-16) https://doi.org/c4g8
DOI: 10.1002/nop2.51 · PMID: 27708830 · PMCID: PMC5050543

Supplemental Figures

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.04.433874; this version posted May 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

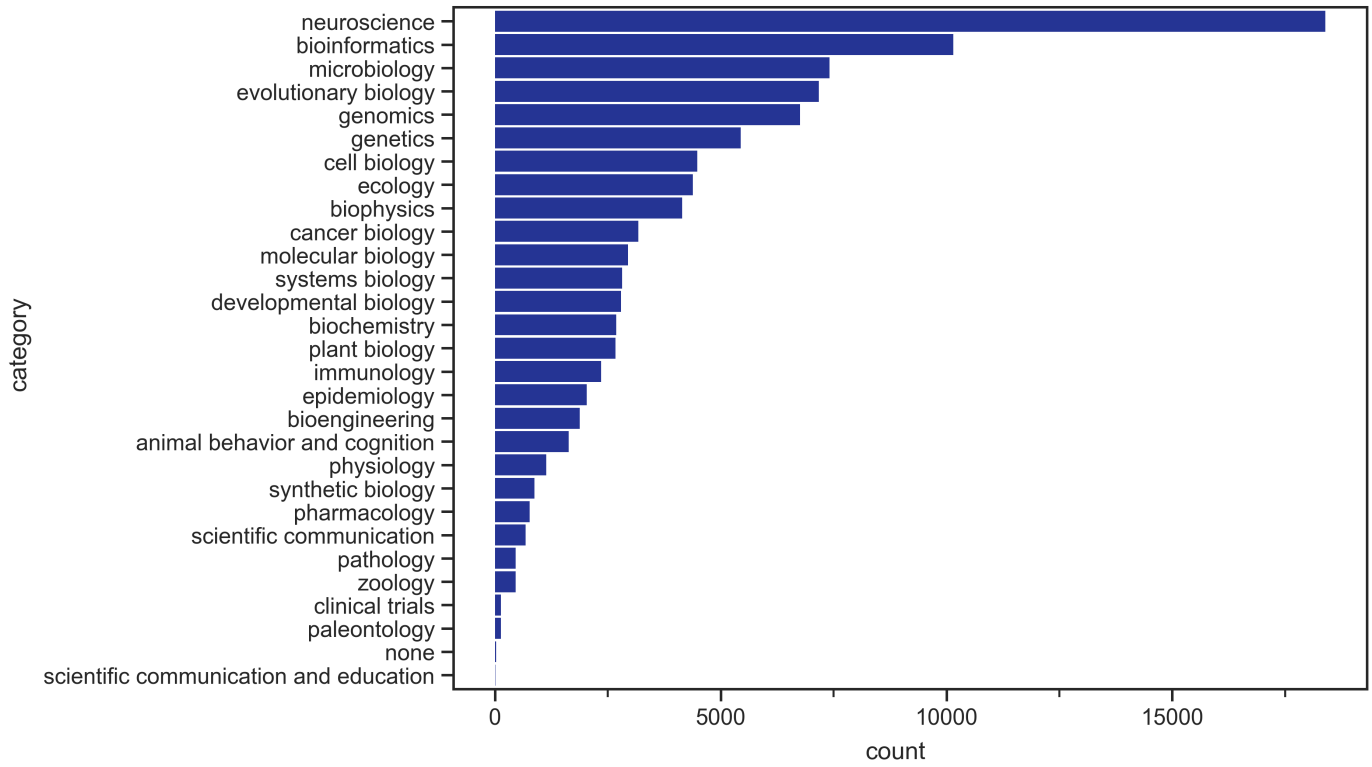


Figure S1: Neuroscience and bioinformatics are the two most common author-selected topics for bioRxiv preprints.

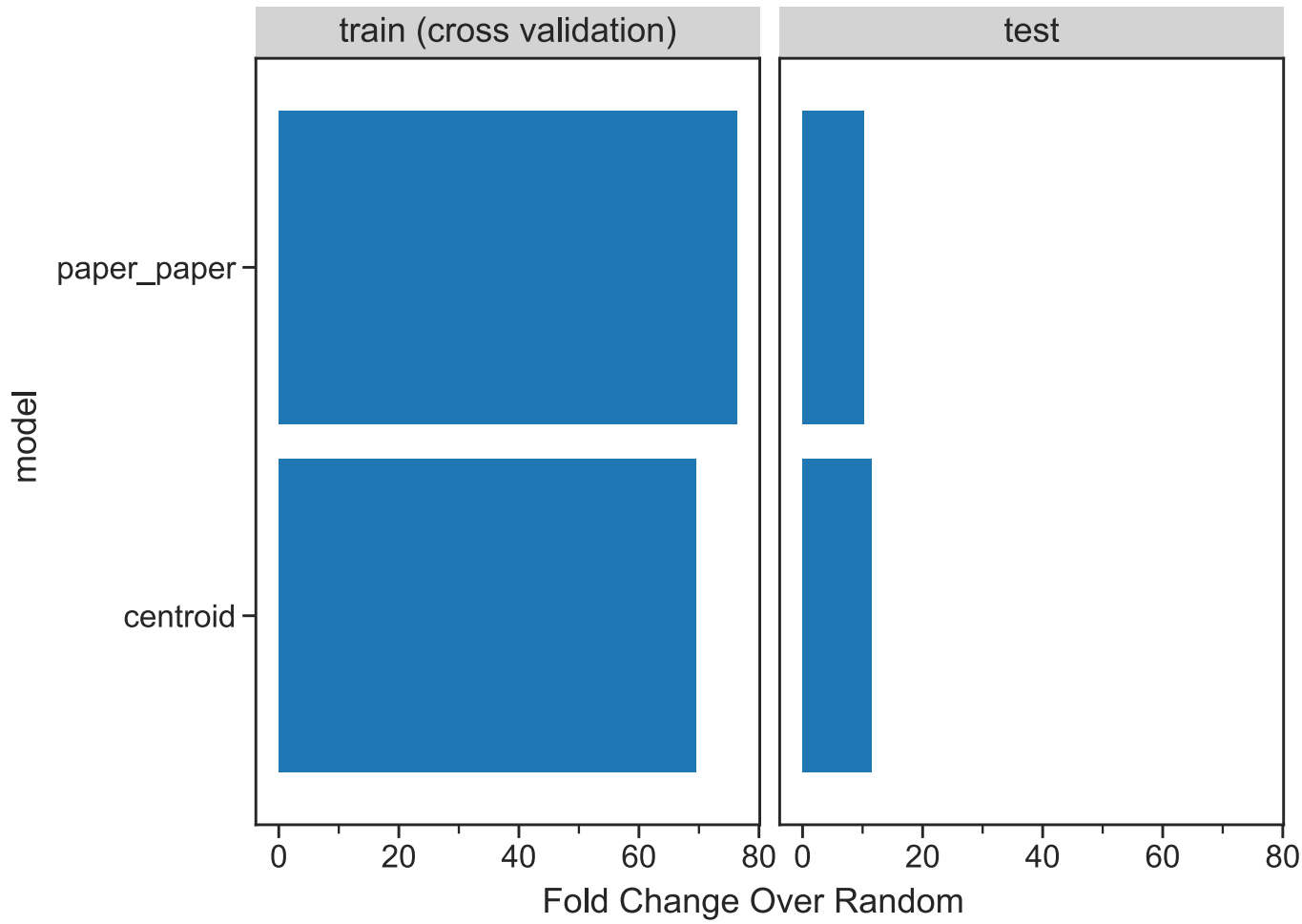


Figure S2: Both classifiers outperform the randomized baseline when predicting a paper's journal endpoint. This bargraph shows each model's accuracy in respect to predicting the training and test set.