Linguistic Analysis of the bioRxiv Preprint Landscape

Authors

- **David N. Nicholson**
  [0000-0003-0002-5761](#) · [danich1](#)
  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 · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (T32 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 (T32 HG010067)

- **Marvin Thielk**
  [0000-0002-0751-3664](#) · [MarvinT](#) · [TheNeuralCoder](#)
  Elsevier

- **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 · Funded by The Gordon and Betty Moore Foundation (GBMF4552); The National Institutes of Health (T32 HG010067)

Abstract

Preprints allow researchers to make their findings available to the scientific community before they have undergone peer review. This provides an opportunity to understand life sciences peer review and publishing practices. Studies of bioRxiv to date have been largely focused on article metadata, and have examined how preprints are downloaded, cited, published, and discussed online. A missing element has been examining the language contained within preprints. Comparing preprints with peer reviewed manuscripts provides an opportunity to examine how peer review changes these documents.

We sought to compare and contrast linguistic features within bioRxiv preprints to biomedical text as a whole and also their published counterparts. The most prevalent elements that changed appeared to be associated with typesetting and mentions of supplementary or additional files. We created document embeddings derived from a bioRxiv-trained word2vec model and found that they revealed different scientific approaches, linked unannotated preprint-peer reviewed article pairs, and were associated with the journal where a preprint would eventually be published. We examined factors associated with the time elapsed between the posting of a first preprint and the appearance of a peer reviewed publication and found that preprints with more versions posted and also those with more textual changes, assessed via document embeddings, took longer to publish. The full text landscape of bioRxiv provides an opportunity to explore scientific language and how peer review alters articles. We provide a web application with which users can explore how a bioRxiv or medRxiv preprint is positioned within the landscape and identify the journals and articles that are most similar.

Introduction

Preprints are scholarly works that are shared before they have been formally peer reviewed and published. The practice of sharing preprints has a long history [1]. The longest ongoing use started with physicists in the 1990s [2]. Preprints were used in the life sciences community during the 1960s before publisher pressure stopped the practice [1]. Over the past decade preprints have made a resurgence within the life sciences community [3,4]. Preprints are now becoming widely accepted and used within the life sciences and other communities [5,5,6,7,8,9,10,11]. Common preprint repositories include arXiv [12], bioRxiv [3] and medRxiv [13]; however, there are over 60 different repositories available [14].

The scientific community has begun to analyze the impact of preprints in the life sciences. Preprints are being posted at an increasing rate [15]. Preprints are also rapidly shared on social media, routinely downloaded, and cited [16]. Articles with matching preprint versions are cited and discussed more often than articles without them [17,18]. Certain categories of preprints seem to be read and shared differently by both scientists and non-scientists [19]. Across preprint servers, analyses suggest that between two-thirds to three-quarters of preprints are eventually published [4,20]. The time required for a preprint to be published can vary from preprint to preprint; however, preprints with a single version often take less time to publish than preprints with multiple versions [21], suggesting that authors may update their preprints between submissions for peer review.

Studies of life sciences preprints have primarily focused on the metadata associated with these articles; however, their textual content remains unexamined.

We sought to understand the language landscape of preprints by performing a linguistic analysis of the bioRxiv corpus. We examined textual differences between preprints and published literature by comparing the entire corpus of preprints with articles available in the open access PubMed Central repository. We also examined linguistic differences between preprints and their corresponding published pairs. Examining this shift will provide a unique opportunity to ascertain parts of the peer

review and publishing process and how it impacts the scholarly literature. Neural-network derived document embeddings provide a useful space for determining the textual similarity of preprints, which enables us to extend this work beyond word frequencies. Examining articles with particularly close proximity in this space reveals unannotated preprint-publication pairs that earlier analyses could not consider. In this space a preprint's nearest neighbors are also more likely than distant articles to share an eventual publishing venue with the preprint itself. We provide a webserver that will displays neighboring journals and articles for any preprint on bioRxiv or medRxiv, which can help authors identify similar papers or suitable journals. Our linguistic analysis, the first of the bioRxiv corpus, reveals the impact of the life sciences publishing process, introduces a method to identify matching preprint-published article pairs, demonstrates that the text content of preprints is related to their eventual publication venue, and provides a more complete picture of the fraction of preprints that are eventually published.

Materials and Methods

Corpora Examined

BioRxiv Corpus

BioRxiv [3] is a repository for life sciences preprints. We downloaded an xml snapshot of this repository on February 3, 2020 from bioRxiv's Amazon S3 bucket [22]. 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 select one of twenty-nine different categories. Researchers also select an article type, which can be a new result, confirmatory finding, or contradictory finding. Some preprints in this snapshot were withdrawn from bioRxiv: when this happens 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

PubMed Central (PMC) [23] is a repository that contains free-to-read articles. PMC articles can be closed access ones from research funded by the United States National Institutes of Health (NIH) appearing after an embargo period or those published under Gold Open Access [24] publishing schemes. Paper availability within PMC is largely dependent on the journal's participation level [25]. Individual journals can 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 [26], or not participate at all. As of September 2019, PMC had 5,725,819 articles available [27]. Out of these 5 million articles, about 3 million were open access and available for text processing systems [28,29]. We downloaded a snapshot of this open access subset on January 31, 2020. This snapshot contained many types of papers: 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 PMC Corpus.

The New York Times Annotated Corpus

The New York Times Annotated Corpus (NYTAC) is [30] is collection of newspaper articles from the New York Times dating from January 1, 1987 to June 19, 2007. This collection contains over 1.8 million articles where 1.5 million of those articles have undergone manual entity tagged by library scientists [30]. We downloaded this collection on August 3rd, 2020 from the Linguistic Data Consortium (see Software and Data Availability section) and used the entire collection for our corpora comparison analysis.

Comparing Corpora

We compared the bioRxiv, PMC, and NYTAC corpora to assess the similarities and differences between them. We use the NYTAC as an out-group to assess the similarity of two life sciences repositories when compared with non-life sciences text. The corpora contain both words and non-word symbols (e.g., $\pm$), which we refer to together as tokens to avoid confusion. We calculated the following statistics 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. Next, we used spaCy's "en_core_web_sm" model [31] (version 2.2.3) to preprocess all corpora and filtered out 326 spaCy-provided stopwords.

Following cleaning, 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 used the union of the top 100 most frequent tokens from each pair of corpora to compare them. We generated a contingency table for each token in this union and calculated the odds ratio and 95% confidence interval [32]. We measured corpus similarity by calculating the Kullback–Leibler divergence across all three corpora, which focuses on token distribution differences as opposed to token-level differences.

Constructing a Document Representation for Life Sciences Text

We sought to build a model that would capture the linguistic similarity of articles. Word2vec is a suite of neural networks designed to model linguistic features of words based on their appearance in text. These models are trained to either predict a word based on its sentence context as a continuous bag of words (CBOW) or predict the context based on a given word in a skipgram model [33]. Through these prediction tasks the networks learn latent features that can be used for downstream tasks such as identifying similar words. We used gensim [34] (version 3.8.1) to train a word2vec continuous bag of words (CBOW) [33] model over the bioRxiv corpus. Our neural network architecture had 300 hidden nodes, and we trained this model for 20 epochs. We set a fixed random seed and used gensim's default settings for all other hyperparameters. Following training, we generated a document vector for every article within bioRxiv and PubMed Central. We calculated the document vector by taking the average of every token present within a given article [35]. Words absent from the word2vec model were ignored.

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) [36] to project bioRxiv document vectors into a low dimensional space. We trained this model using the scikit-learn [37] implementation of a randomized solver [38] with a random seed of 100, output of 50 principal components (PCs), and default settings for all other hyperparameters. After fitting, each preprint has a score for each PC. To visualize the tokens associated with each PC, we calculated the cosine similarity of each PC to all tokens in our word2vec model's vocabulary. We report the top 100 positive and negative scoring tokens in the form of word clouds, where 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 largely based on exact matching of certain attributes, and authors can forget to

establish a link after publication. Authors can change the title between a preprint and published version (e.g., [39] and [40]), which prevents bioRxiv from automatically establishing a link. If the authors do not report the publication to bioRxiv, the preprint and the published version are treated as distinct entities despite representing the same underlying research. We recognized that close proximity in the embedding space could reveal preprint to published version links that were missed by existing automated processes. First, we used CrossRef [41] to identify bioRxiv preprints that were linked to a corresponding published article. We ignored pairs that contained papers not in the PMC corpus. We calculated the distribution of known preprint to published distances 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. Next, we calculated distances between preprints without a published version link with PubMed Central articles that weren't matched with a corresponding preprint. We filtered any potential links with distances that were greater than the minimum value of the background distribution to reduce the curation burden. 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 for manual annotation. We 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 the preprint-publication pairs on which reviewers disagreed to a third scientist, who carefully reviewed each case and made a final determination. We used this curated set to evaluate the extent to which distance in the embedding space revealed true but unannotated links between preprints and their published versions.

Measuring Time Duration for Preprint Publication Process

We measured the time required for preprints to be published in the peer reviewed literature and compared this time within fields and as a function of the extent to which documents changed between the preprint and publication. 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 calculated the difference between the date at which a preprint was first posted and its publication date to provide a publication interval, and we also recorded the number of preprint versions posted onto bioRxiv. To measure the amount of textual difference, we calculated the Euclidean distance between the document representation of each preprint and the corresponding published version. We performed linear regression to model the relationship between preprint version count and a preprint's time to publication as well as the relationship between document representation distances and a preprint's time to publication. We visualized results as square bin plots. We observed a limited number of cases in which authors appeared to post preprints after the date of publication, which results in preprints receiving a negative time difference, as previously reported [42]. We did not remove preprints that had a negative time publication in our linear regression analysis as it was not strictly necessary, but we removed them in our survival curve analysis where they were incompatible with the analytical approach. In practice, the number with negative publication times and the short lead time between publication and preprint has a minimal impact on results.

Document distances can be difficult to understand, so we sought to contextualize the meaning of a distance unit. We selected preprints within the Bioinformatics topic area, which was well-represented on bioRxiv. For preprints submitted to the Bioinformatics topic area, we sampled a pair of preprints and calculated their differences 1000 times and reported the mean.

In addition to contextualizing the document distance, we also wanted to contextualize differences in the time to publication. We examined time to publication for each topic area using the Kaplan-Meier estimator [43] on preprints within bioRxiv, treating preprints not yet published as survival. We generated these curves using the KaplanMeierFitter function from the lifelines [44] (version 0.25.6) python package. We reported the half-life of each bioRxiv preprint category.

Building Journal Venue Classifiers

We hypothesized that preprints would be more likely to be published in journals that contained similar content to the work in question. To test this hypothesis, we designed an experiment examining document and journal representations. First, we removed all journals that had fewer than 100 papers in the PMC corpus. A subset of our PMC corpus was directly linked to papers in bioRxiv as they had been published as open access articles. We held out this subset and treated it as a gold standard test set. We used the remainder of the PMC corpus for training and initial evaluation via cross validation using the scikit-learn k-Nearest Neighbors implementation [45]. We imagined a use case of prioritizing relevant journals for preprint authors, and considered a list of ten journal suggestions to be an appropriate number and we considered a prediction to be a true positive if the correct journal appeared within the ten closest neighbors of the query article.

Certain 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.

For the manuscript-based approach, we identified the ten most similar published manuscripts and evaluated where the documents were published. We embedded each query article into the space defined by the word2vec model as described for preprints. We selected the ten manuscripts that were nearest by Euclidean distance in the embedding space and returned the journal in which they were published. The number of journals returned via this method could be less than ten as multiple papers in close proximity to query article may belong to the same journal. Because this approach was based on paper proximity, we could return the articles that led to each journal being returned. However, journals that publish more papers are more frequently recommended in this framing.

For the journal-based approach, we identified the ten most similar journals by constructing a journal representation in the same embedding space. We computed journal centroids as the average embedding of all published papers in the journal. We then project a query article into the same space and return the ten closest journal centroids by Euclidean distance. This technique guaranteed that at least ten distinct journals were returned and prevented journals that publish many papers from being heavily overrepresented.

In both cases, we set the number of neighbors for each model to be 10 and then evaluated both models via 10-fold cross validation. We evaluated performance of both classifiers on our gold standard test set of published preprints.

Web Application for Discovering Similar Preprints and Journals

We developed a web application that identifies similar papers and journals for any bioRxiv and medRxiv preprint and that places the preprint into the overall document landscape. Our web application downloads a pdf version of a preprint hosted on the bioRxiv or medRxiv server. We use pdfminer [46] to extract text from the downloaded pdf. The extracted text is then fed into our word2vec 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 as well as journal centroids. We implemented this search using the scikit-learn implementation of k-d trees. To run cost effectively on Amazon Web Services, we sharded the k-d trees into four trees.

Accompanying these recommendations, we also provide a neural network derived visualization of our training set and the article's position within it. We used SAUCIE [47], 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 2000 iterations. When a user requests a new document, we can then project the document on the pretrained model to generate a visualization in two-dimensional space. We illustrate our recommendations as a short list and provide access to our network visualization at https://greenelab.github.io/annorxiver-journal-recommender/.

We used the fully trained model to project user-requested bioRxiv preprints onto the generated landscape to enable users to see where their preprint falls along the landscape.

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 [48]. Because the rate of change is rapid, we first analyzed category data and compared our results with previous findings. As in previous reports [48], neuroscience remains the most common category of preprint followed by bioinformatics (Supplemental Figure S1). Microbiology, which was fifth in the most recent report [48], 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, which is consistent with reports on a smaller set [49]. Taken together, the results suggest that while bioRxiv has experienced dramatic growth, the way in which it is being used appears to have remained consistent in recent years.

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

Metric	bioRxiv	PMC	NYTAC
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%
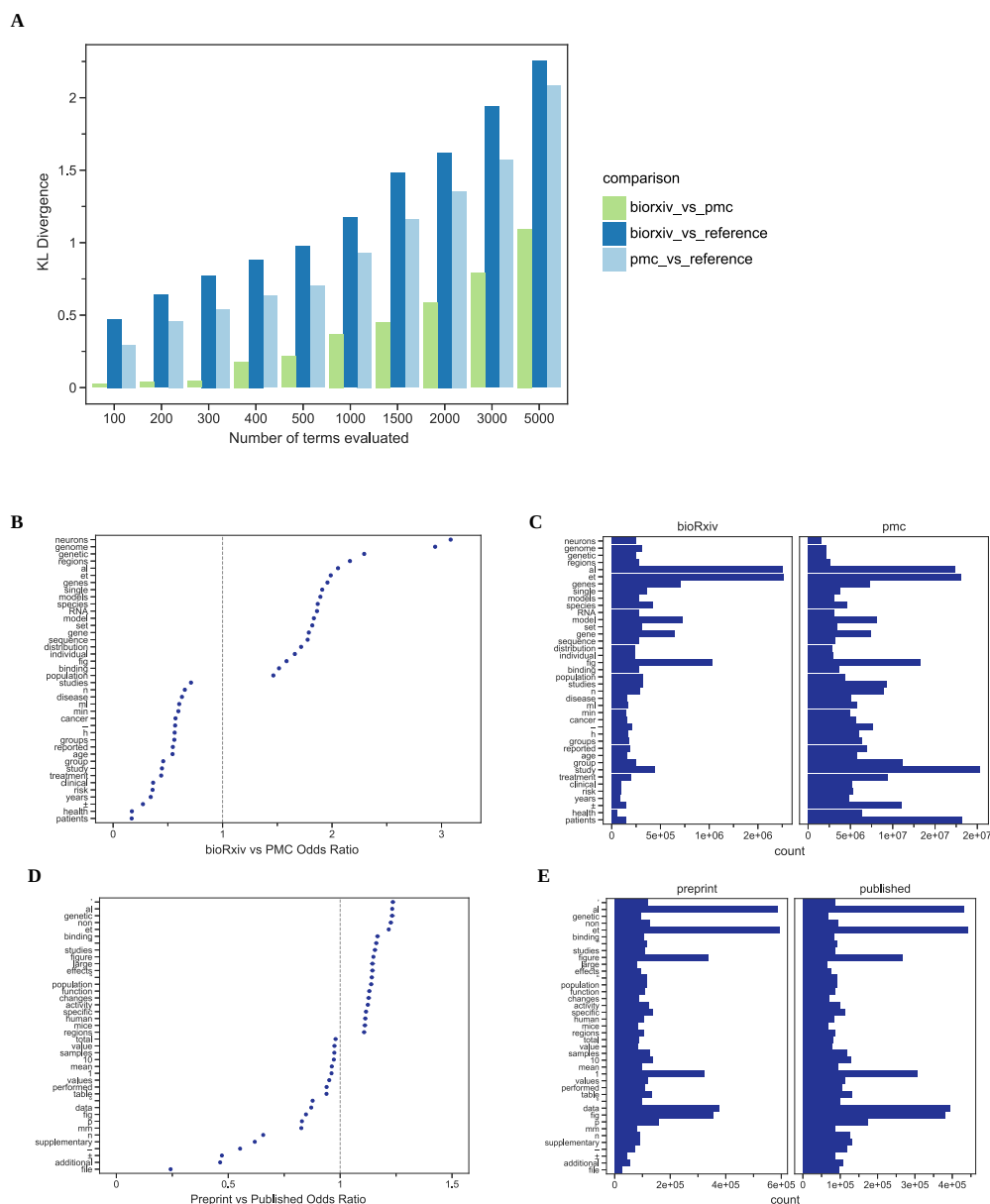
Figure 1: A. The Kullback–Leibler divergence measures the extent to which the distributions, not specific tokens, differ from each other. The token distribution of bioRxiv and PMC corpora is more similar than these biomedical corpora are to the NYTAC one. B. The major differences in token frequencies for the corpora appear to be driven by the fields that have had the highest uptake of bioRxiv, as terms from neuroscience and genomics are relatively more abundant in bioRxiv. Points indicate the log_2(OR) for each token and error bars indicate the 95% confidence intervals. C. Of the terms 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 major differences in token frequencies for preprints and their corresponding published version often appear to be associated with typesetting and supplementary or additional materials. Points indicate the log_2(OR) for each token and error bars indicate the 95% confidence intervals. E. The tokens with the largest absolute differences in abundance appear to be stylistic.

We first compared the overall text of bioRxiv with PMC, adding a corpus of professionally written but non-biomedical text (NYTAC) as a control. Documents on bioRxiv were slightly longer than those on PMC, but both were much longer than those from NYTAC (Table 1). Other than length, both corpora were otherwise quite similar. The average sentence length, 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 divergence measures the extent to which two distributions, but not the specific entities that comprise those distributions, differ. The distribution of term frequencies in bioRxiv and PMC were low, especially among the top few hundred tokens (Figure ??A). Differences began to emerge over more terms, but remained much lower than when the biomedical corpora are compared with NYTAC.

Examining the frequencies of individual terms revealed differences between the biomedical corpora. Previous work examining author-selected categories has reported that fields appear to have preprinted unevenly, with certain life sciences research fields having more uptake than others [48]. However, it was possible that authors simply selected certain fields preferentially but that the content was similar to the broader corpus of life sciences text. We directly examined this by comparing term frequencies between bioRxiv and PMC. We found that among the terms that differed the most many were associated with certain life sciences research fields. Terms like "neurons" "genome" and "genetic", which are common in genomics and neuroscience, were more common in bioRxiv than PMC while others associated with clinical research, such as "clinical" "patients" and "treatment" were more common in PMC (Figure ??B and ??C).

We next controlled for differences in the body of documents to identify term-level changes associated with the publication process itself by examining only pairs of preprints and their corresponding publication (Figure ??D and ??E). The tokens that differed included "et" "al", "±", "–" and others that appeared to be typesetting related. Certain changes appeared to be related to journal styles: "figure" was more common in bioRxiv while "fig" was relatively more common in PMC. 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 PMC 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.

Taken together these results suggested that the structure of the text in documents on bioRxiv was similar to that on PMC. The differences in uptake across fields is supported not only by differences in authors' categorization of their articles but also in the text of 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.

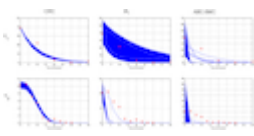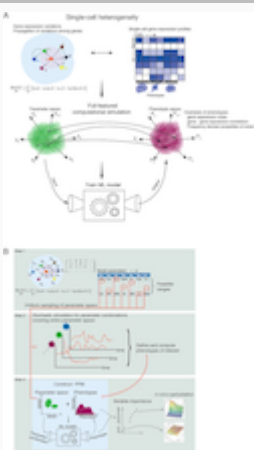## Document embeddings derived from bioRxiv reveal fields and subfields

Figure 2: A. Principal components (PC) analysis of bioRxiv word2vec embeddings groups documents by 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. The second PC separated bioinformatics from neuroscience. Certain neuroscience papers appeared to be more associated with the cellular biology direction of PC1, while others appeared to be more associated with the informatics-related direction, which suggested that the concepts captured by PCs were not exclusively related to field. B. A word cloud visualization of PC1, which separated informatics disciplines (positive direction) from cell biology (negative direction) showed that tokens "empirical" "estimates" and "statistics" characterized the positive direction while "cultured" and "overexpressing" characterized the negative one. Each word cloud depicts the cosine similarity score between tokens and the second 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, showed that tokens "genomic" "genome" and "genomes" characterized the positive direction while "evoked" "stimulus" and "stimulation" characterized the negative one. 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. Certain 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 [35,50,51]. We first trained word embeddings using a 300-dimensional word2vec continuous bag of words model. We combined word embeddings to produce an embedding for each bioRxiv or PMC document by calculating the average of all words present in each respective document. This placed each document in a 300-dimensional space where each individual dimension is arbitrary. To provide more structure to the dataset, we examined the

predominant patterns in these embeddings by performing principal components analysis of bioRxiv. The principal components (PCs) are ordered by the proportion of the variance explained. We found that the first two PCs separated articles from different author-selected categories (Figure 2A).

We sought to understand the token patterns that drove these overall differences between documents. We identified the tokens most similar to a PC by calculating the cosine similarity score between tokens' embeddings in the word2vec space and each PC. Visualizing token-PC similarity revealed tokens associated with certain research approaches (Figures 2B and 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).
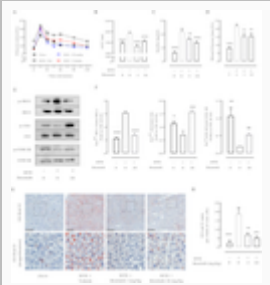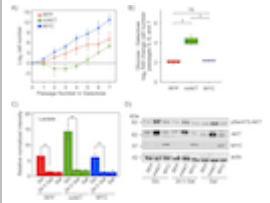
While the PC1 value range for each author-selected category was high, these results suggested that a primary driver in the variability of language use on bioRxiv could be the divide between data science approaches and cell biology ones. A similar analysis for PC2 suggested that neuroscience and genomics present a similar language continuum (Figure 2E). 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).

We examined the preprints that had the highest and lowest values for PC1 within systems biology (Table {#tbl:five_pc1_table}). The preprints with the highest five PC values [52,53,54,55] included software packages, machine learning analyses, and other computational biology manuscripts. The preprints with the lowest five PC values [56,57,58,59,60] were focused on signaling. We provide the top 50 PCs of bioRxiv embeddings within 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	Figure Link
Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis [52]	4.700554908074704	None		https://www.biorxiv.org/content/biorxiv/early/2017/10/02/197400/F1.large.jpg
Machine learning of stochastic gene network phenotypes [53]	4.410660604449826	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2019/10/31/825943/F5.large.jpg
Notions of similarity for computational biology models [61]	4.355295926618207	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2016/03/21/044818/F1.large.jpg

Title [citation]	PC1	License	Figure Thumbnail	Figure Link
GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation [54]	4.351517618262304	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2019/09/18/769299/F1.large.jpg
SBpipe: a collection of pipelines for automating repetitive simulation and analysis tasks [55]	4.321847854182741	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2017/02/09/107250/F1.large.jpg
Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection [56]	-4.263964235099807	CC-BY-ND		https://www.biorxiv.org/content/biorxiv/early/2018/11/07/455048/F1.large.jpg
Systems analysis by mass cytometry identifies susceptibility of latent HIV-infected T cells to targeting of p38 and mTOR pathways [57]	-4.279016673409032	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2018/07/19/371922/F1.large.jpg
NADPH consumption by L-cystine reduction creates a metabolic vulnerability upon glucose deprivation [58]	-4.592209988884236	None		https://www.biorxiv.org/content/biorxiv/early/2019/08/13/733162/F1.large.jpg

Title [citation]	PC1	License	Figure Thumbnail	Figure Link
Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease [59]	-4.736613689905791	None		https://www.biorxiv.org/content/biorxiv/early/2019/08/28/745943/F1.large.jpg
AKT but not MYC promotes reactive oxygen species-mediated cell death in oxidative culture [60]	-4.826793742506695	None		https://www.biorxiv.org/content/biorxiv/early/2019/09/01/754572/F1.large.jpg

Document embedding similarities reveal unannotated preprint-publication pairs

Metaresearch into bioRxiv, including our own, relies on annotations of preprints that have been published to their corresponding peer reviewed publication. Many journals require that authors update preprints with links to the published version of their article. This is accomplished in two ways: bioRxiv may detect the link and automatically add it or authors may notify bioRxiv that their preprint was published. However, bioRxiv establishes these links based on consistency in metadata (i.e., title, author names, etc). Article titles, author lists, and other elements may change as a result of the peer review process. In these cases, if authors do not notify bioRxiv that the preprint has since been published then the scientific record remains incomplete.

Based on our finding that document embeddings captured fields and subfields, we expected that preprint-publication pairs with similar embedding values may represent the same document. We examined the extent to which annotated preprint-publication pairs were closer in this space that preprints were to a random paper published in the same journal in which the preprint was eventually published using already annotated pairs. We found that 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).
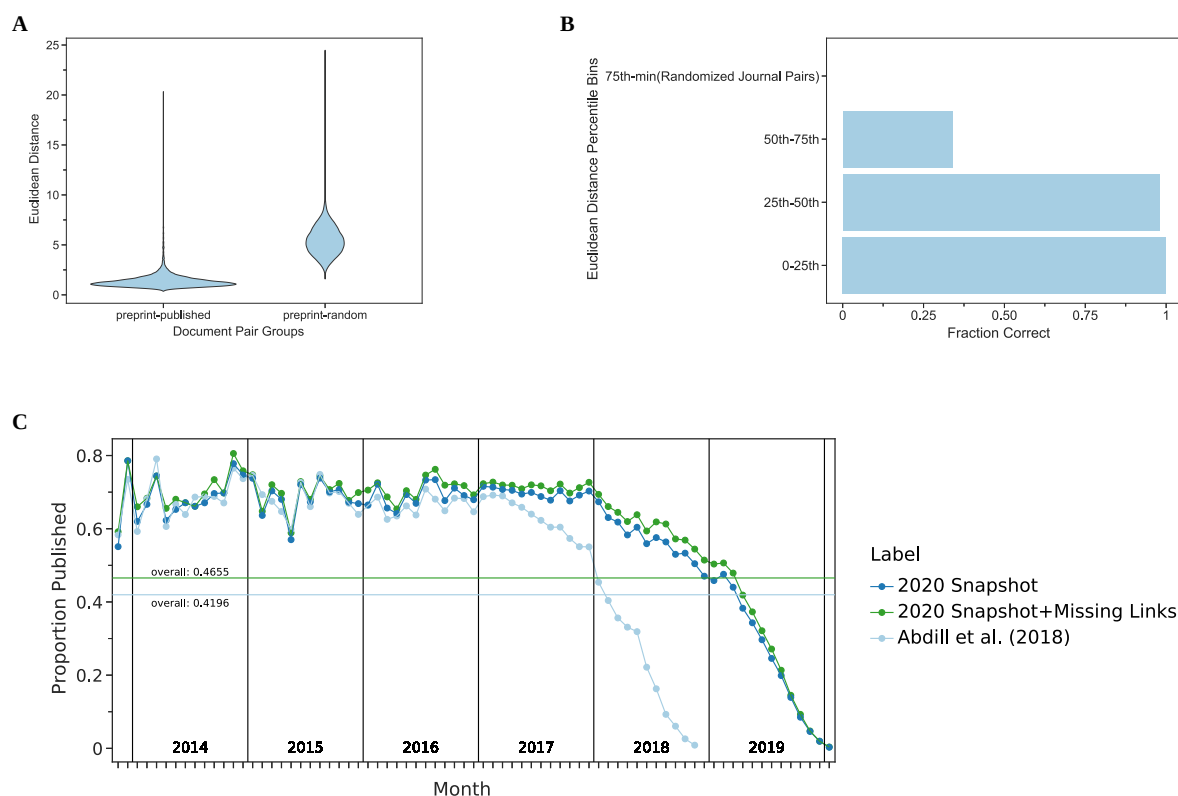
Figure 3: A. Preprints are closer in document embedding space to their corresponding peer reviewed publication than they are to random papers published in the same journal. B. Potential preprint-publication pairs that are unannotated but within the 50th percentile of all preprint-publication pairs in the document embedding space are likely 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 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. The light blue line represents the publication rate estimated by Abdill et al. [48]. The dark blue line represents the updated publication rate using only CrossRef-derived annotations while the dark green line includes annotations derived from similarity in the embedding space. The horizontal lines represent the overall proportion of preprints that are were published as of the time of the annotation snapshot.

Based on this finding, we analyzed preprint-publication pairs that were close in document space but not annotated as such. We separated these pairs into four quantiles with the first three based on the distribution of preprint-publication distances and the fourth going from the 75th percentile in the preprint-publication pair space to the smallest value observed for the preprint-random set. We then selected 50 preprint-publication pairs from each of these sets and shuffled them to create a random list of 200 possible pairs. Two scientists then examined the pairs in these randomized lists determining if an preprint-publication pair represented the peer reviewed publication of the primary content described in the preprint. Across the entire list we found a high inter-rater reliability of with a Cohen's Kappa [62] of 0.92. In the case of disagreements, a third scientist more carefully examined the pairs and made a final determination. Of the 200 pairs that we examined, approximately 98% of 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 1,720 preprint-article pairs, suggesting that many preprints have been published but not previously connected with their published versions.

We overlaid these new annotations onto existing annotations to reassess the overall preprint publication rate reported by Abdill et al. [48]. Our filtering criteria were intentionally stringent, so the increased estimate of publication rate amounts to a few percent (Figure 3C). We noticed that there was a particular enrichment of unannotated but published preprints in the 2017-2018 interval. We would expect a higher proportion of such preprints before the year 2019 (many of which may not

have been published yet); however, we did not expect to observe relatively few missed annotations before 2017. It is possible that as the number of preprints grows, it has become harder to establish links. Alternatively, authors now adopting preprinting practices may be less likely to notify the preprint server upon publication if the links are not automatically detected. In any case, future work, especially that which aims to assess the fraction of preprints that are eventually published, should account for the possibility of missed annotations. We supplied our set of 1720 high-confidence annotations to the bioRxiv staff.

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

A

B

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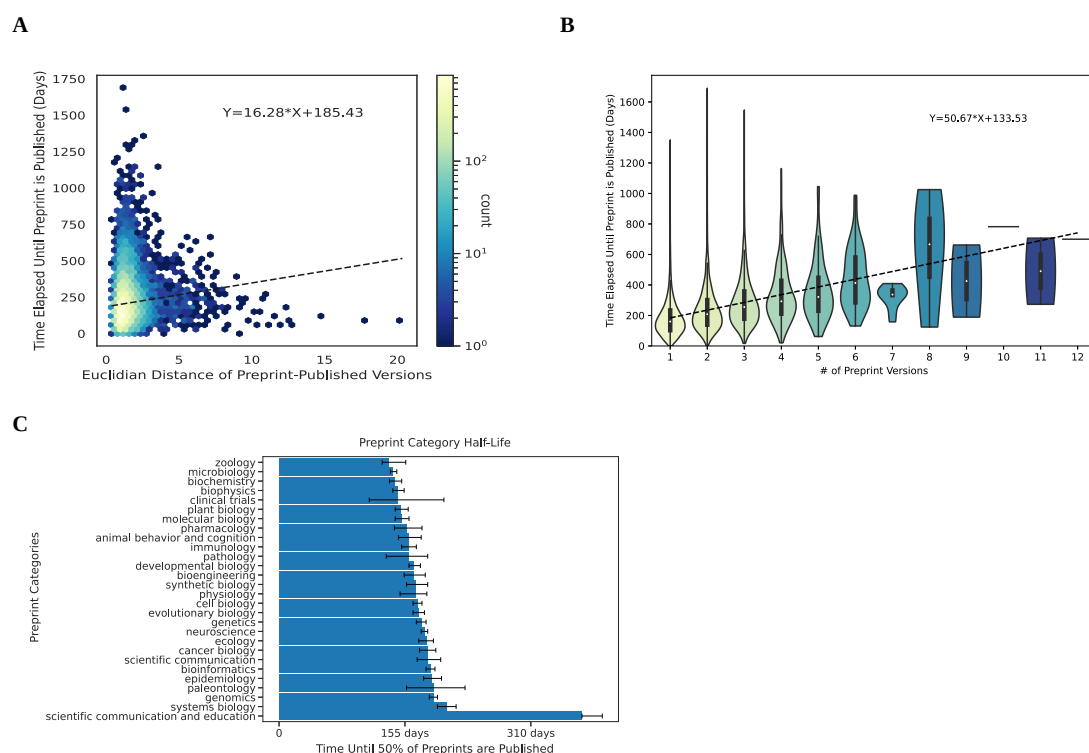
Figure 4: A. Preprints with more substantial text changes took longer to be published. The x-axis shows the Euclidean distance between document representations of the first version of a preprint and it's peer reviewed form. The y-axis shows the number of days elapsed between when the first version of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

The process of peer review includes a number of steps which take variable amounts of time [63]. Comparing bioRxiv preprints with their corresponding published version provides an opportunity to better understand peer review and publishing. We examined how long it took to publish preprints in their peer reviewed form for each author-selected category (Figure 4C). We considered a preprint's endpoint to be publication and used the Kaplan-Meier estimator to estimate how much time has elapsed. 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]). Though we did observe category-specific differences, these differences were generally modest. 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]).

Preprint authors have the opportunity to update their preprint at any point until the peer reviewed publication appears. We examined whether or not the number of versions of a preprint was associated with a change in the time to publish. We found that preprints with more versions generally took longer to publish (Figure 4B). Linear regression comparing the time to publish with the number of versions revealed an increasing relationship, with each version being associated with an additional 51 days before publication. This time period 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 feedback.

We next used the full text content to examine the extent to which the magnitude of change to the text between the preprint and peer reviewed publication was associated with a change in the time to publication. We used distance in the document embedding space to quantify the amount of change to the text. We first established a baseline by selecting pairs of random preprints in the author-selected bioinformatics category and found a mean embedding distance between random pairs of 5.068. Preprints that were further in the embedding space from their corresponding peer reviewed publication took longer to publish, with each unit in the embedding space corresponding to approximately sixteen additional days (Figure 4A). The number of version and document distance effects, taken together, support a model where preprints that are reviewed more or that require larger revisions take longer to publish.

Preprints with similar document embeddings share publication venues

We hypothesized that document embedding might capture information associated with the eventual publication venue. We used the annotated preprint-publication pairs as a fully held out test set. We embedded all PMC papers except those annotated to have a corresponding preprint into the bioRxiv embedding space. For journals, we averaged the document representation of all papers in the journal to produce a journal centroid. We used a k-nearest neighbor classifier trained on the remaining PMC documents to identify the most similar published papers and assessed performance - the extent to which the journal a paper appeared in the returned list of ten neighbors - through cross validation and the held out test set (Supplemental Figure S4). We found an enrichment of papers published within the same journal among the neighbors of the target paper.

We developed an online app that returns the nearest published papers and journals that are nearest to a preprint. Users supply digital object identifiers (DOIs) from bioRxiv or medRxiv. The application downloads the article from the preprint repository, converts the PDF to text, calculates a document embedding score, and returns the ten papers and journals with the most similar representations in the embedding space (Figure ??). We also sought to display the position of each preprint in the overall landscape. We required an embedding that could be rapidly calculated for new documents, so we used the Sparse Autoencoder for Unsupervised Clustering, Imputation, and Embedding (SAUCIE) approach previously described for the analysis of single-cell gene expression data [64]. The user-requested preprint's location in this space is then displayed and users can select regions to identify the terms most associated with those regions. 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.

![The journal 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 query 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.] (images/journal_recommender_workflow.png){#fig:journal_rec_workflow width="100%"}

Conclusions

We analyzed the full text of bioRxiv, comparing it with the open access portion of PMC, and found that the overall manner of writing is consistent with the biomedical literature. A benefit of analyzing bioRxiv text is that we can compare preprint and published versions to examine the influence of the publication process. Token-level analyses suggest that differences between corpora are driven by fields, while comparisons of preprints with their corresponding publication reveals differences in typesetting and supplementary materials.

Previous analyses have focused on article metadata, including which papers are heavily downloaded and discussed [48] and by which communities discuss them [65]. We found that some preprints are high similar to published articles within the PMC open access corpus, and a detailed examination revealed that many preprints were published but not previously annotated. Using the full text of documents to correct these missing annotations provides a comprehensive understanding of the extent to which preprints are published and this correction resulted in a publication rate that is higher than previously estimated. Importantly, this only accounts for papers that are published open access, so our analysis should be considered to raise the lower bound but the truly published fraction is likely to be higher.

Our work presents a first step towards understanding the process by which peer review alters life sciences papers through an analysis of the full text of preprints and their corresponding publication. Our broad-based examination suggests certain changes, but the scale of changes appears modest. We lay the groundwork for future work that aims to identify sentences and claims that are altered during review. Our finding that document embeddings reveal manuscripts with similar outcomes is likely to be important to new tools that accelerate publishing, including those that automatically perform integrity checks and other critically important contributions. We also found that document embeddings captured many elements of how authors use language in preprints. Fields were separated first on an informatics / molecular axis, and author-selected categories could be distinguished based on embeddings. Document embeddings were also associated with the eventual journal at which the work was published. Based on this observation, we supply a web application that returns the papers and journals that are most 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 journal recommendation website can be found at https://greenelab.github.io/annorxiver-journal-recommender/, and code for the website is available under a BSD-2-Clause Plus Patent License at https://github.com/greenelab/annorxiver-journal-recommender. 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

Competing Interests

Need to note Marvin's employer here.

References

1. **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

2. **Preprint Déjà Vu**
   Paul Ginsparg
   The EMBO Journal (2016-10-19) https://doi.org/f3r9vf
   DOI: 10.15252/embj.201695531 · PMID: 27760783 · PMCID: PMC5167339

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

4. **Abstract**
   eLife Sciences Publications, Ltd
   (2019-05-09) https://doi.org/gf5cqt
   DOI: 10.7554/elife.45133.001

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

6. **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

7. **Preprints for the life sciences**
   J. M. Berg, N. Bhalla, P. E. Bourne, M. Chalfie, D. G. Drubin, J. S. Fraser, C. W. Greider, M. Hendricks,
   C. Jones, R. Kiley, … C. Wolberger
   Science (2016-05-19) https://doi.org/bmp7
   DOI: 10.1126/science.aaf9133 · PMID: 27199406

8. **The rise of preprints in chemistry**
   François-Xavier Coudert
   Nature Chemistry (2020-05-18) https://doi.org/ghdd64
   DOI: 10.1038/s41557-020-0477-5 · PMID: 32424256

9. **Preprints: An underutilized mechanism to accelerate outbreak science**
   Michael A. Johansson, Nicholas G. Reich, Lauren Ancel Meyers, Marc Lipsitch
   PLOS Medicine (2018-04-03) https://doi.org/gg922h
   DOI: 10.1371/journal.pmed.1002549 · PMID: 29614073 · PMCID: PMC5882117

10. **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

11. **In praise of preprints**
Norman K. Fry, Helina Marshall, Tasha Mellins-Cohen
Microbial Genomics (2019-04-01) https://doi.org/gg3bxc
DOI: 10.1099/mgen.0.000259 · PMID: 30938670 · PMCID: PMC6521583

12. **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

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

14. **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/

15. **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

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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **Machine access and text/data mining resources | bioRxiv** https://www.biorxiv.org/tdm

23. **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

24. **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

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

26. **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

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

28. **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

29. **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

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

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

32. **Odds Ratio**
Steven Tenny, Mary R. Hoffman
StatPearls (2020) http://www.ncbi.nlm.nih.gov/books/NBK431098/

33. **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

34. **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)

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

36. **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

37. **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)

38. **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

39. **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 (2018-07-24) https://doi.org/gg4hp7
DOI: 10.1101/376665

40. **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

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

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

43. **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

44. **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

45. **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

46. **Welcome to pdfminer.six's documentation! — pdfminer.six 20201018 documentation**
https://pdfminersix.readthedocs.io/en/latest/index.html

47. **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

48. **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

49. **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

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

51. **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

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

53. **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

54. **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

55. **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

56. **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
bioRxiv (2018-11-07) https://doi.org/gg94bc
DOI: 10.1101/455048

57. **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

58. **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 (2019-08-13) https://doi.org/gg94bf
DOI: 10.1101/733162

59. **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
bioRxiv (2019-08-28) https://doi.org/gg94bg
DOI: 10.1101/745943

60. **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
bioRxiv (2019-09-01) https://doi.org/gg94bh
DOI: 10.1101/754572

61. **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

62. **A Coefficient of Agreement for Nominal Scales**
Jacob Cohen
Educational and Psychological Measurement (2016-07-02) https://doi.org/dghsrr
DOI: 10.1177/001316446002000104

63. **Peer review and the publication process**
    Parveen Azam Ali, Roger Watson
    Nursing Open (2016-10) https://doi.org/c4g8
    DOI: 10.1002/nop2.51 · PMID: 27708830 · PMCID: PMC5050543

64. **Exploring single-cell data with deep multitasking neural networks**
    Matthew Amodio, David van Dijk, Krishnan Srinivasan, William S. Chen, Hussein Mohsen, Kevin R. Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, … Smita Krishnaswamy
    Nature Methods (2019-10-07) https://doi.org/gf9rsg
    DOI: 10.1038/s41592-019-0576-7 · PMID: 31591579

65. **Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation**
    Jedidiah Carlson, Kelley Harris
    PLOS Biology (2020-09-22) https://doi.org/ghk53x
    DOI: 10.1371/journal.pbio.3000860 · PMID: 32960891 · PMCID: PMC7508356
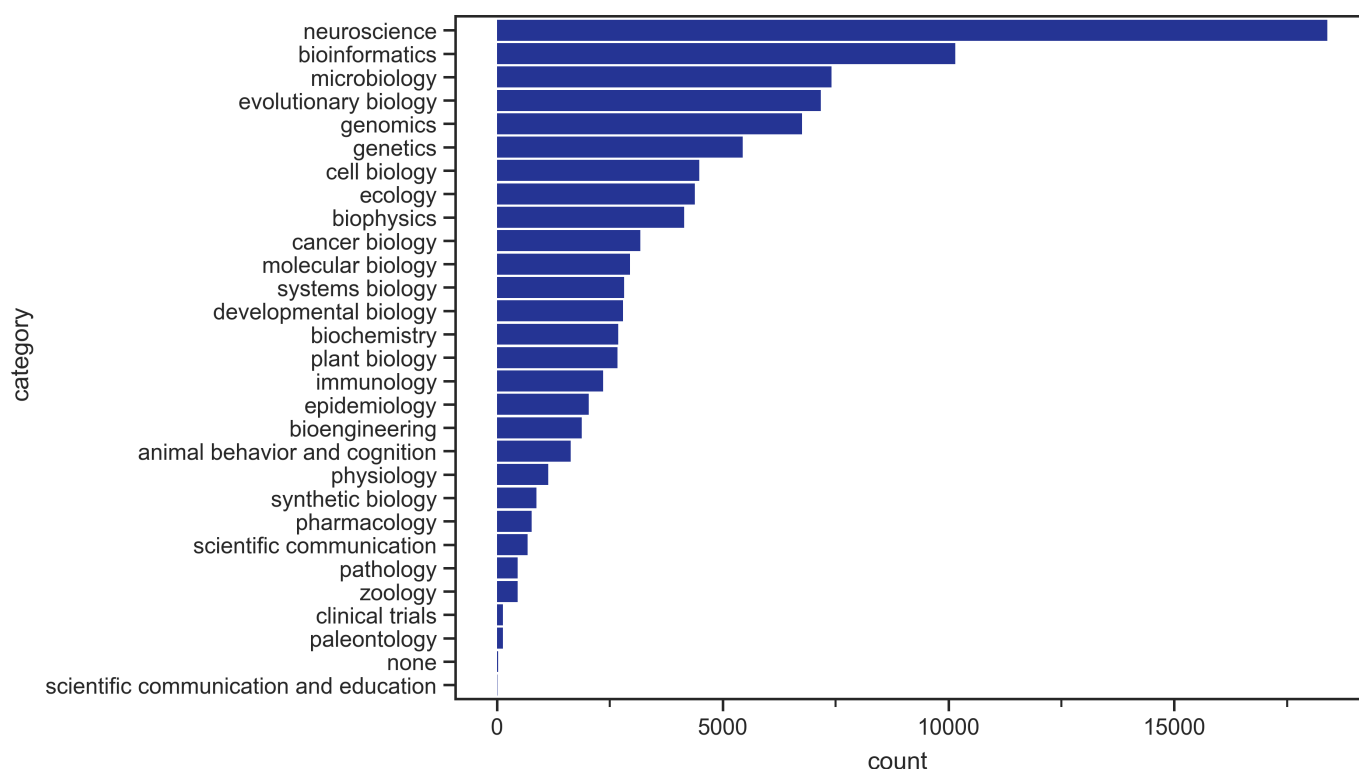
Supplemental Figures

Figure S1: Neuroscience and bioinformatics are the two most common author-selected topics for bioRxiv preprints.
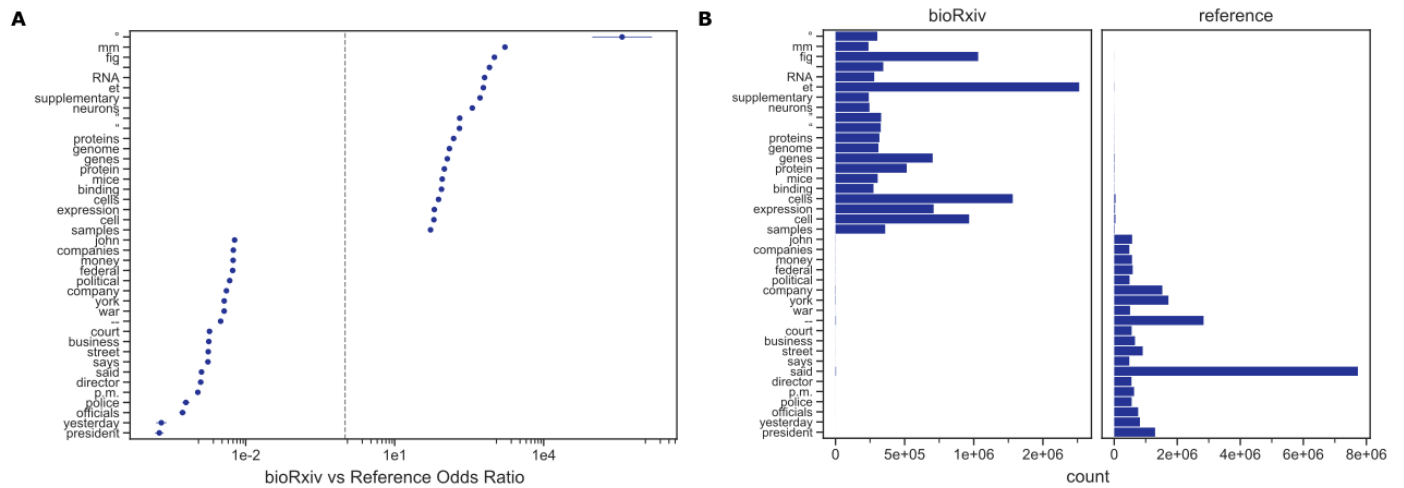
Figure S2: Topic associated tokens are highly enriched when comparing bioRxiv to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to bioRxiv. Values greater than one indicate a high association with bioRxiv whereas values less than one indicate high association with the New York Times. The dotted line provides a breaking point between both categories. The plot on the right (B) is a bar chart of token frequency appearing in bioRxiv and New York Times respectively.
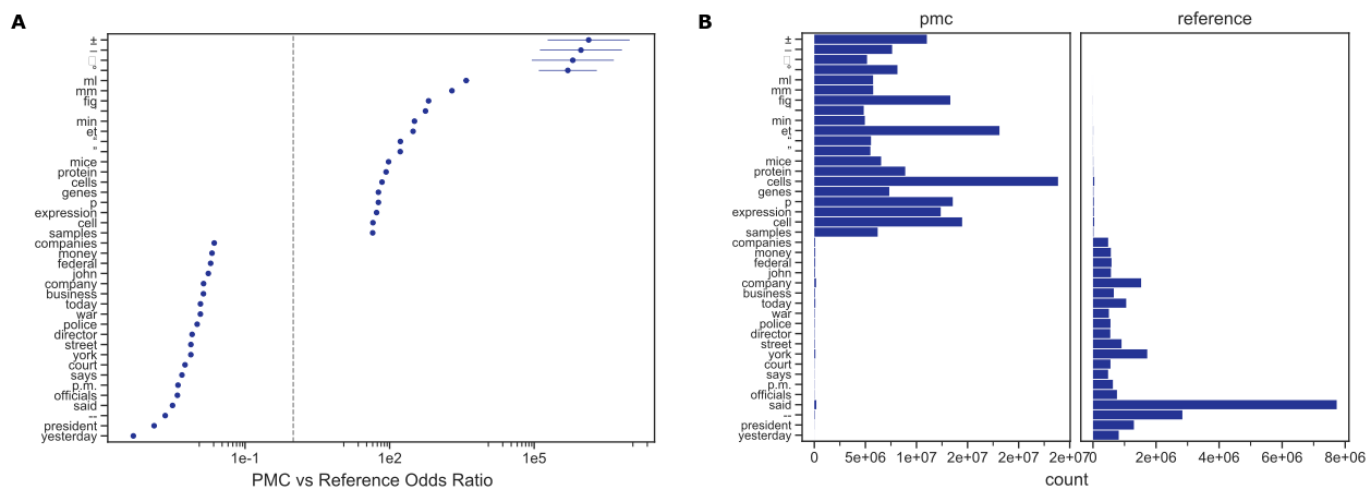
Figure S3: Typesetting symbols and biologically relevant tokens are highly enriched when comparing PubMed Central (PMC) to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to PMC. Values greater than one indicate a high association with PMC whereas values less than one indicate high association with the New York Times. The dotted line provides a breaking point between both categories. The plot on the right (B) is a bar chart of token frequency appearing in PMC and New York Times respectively.
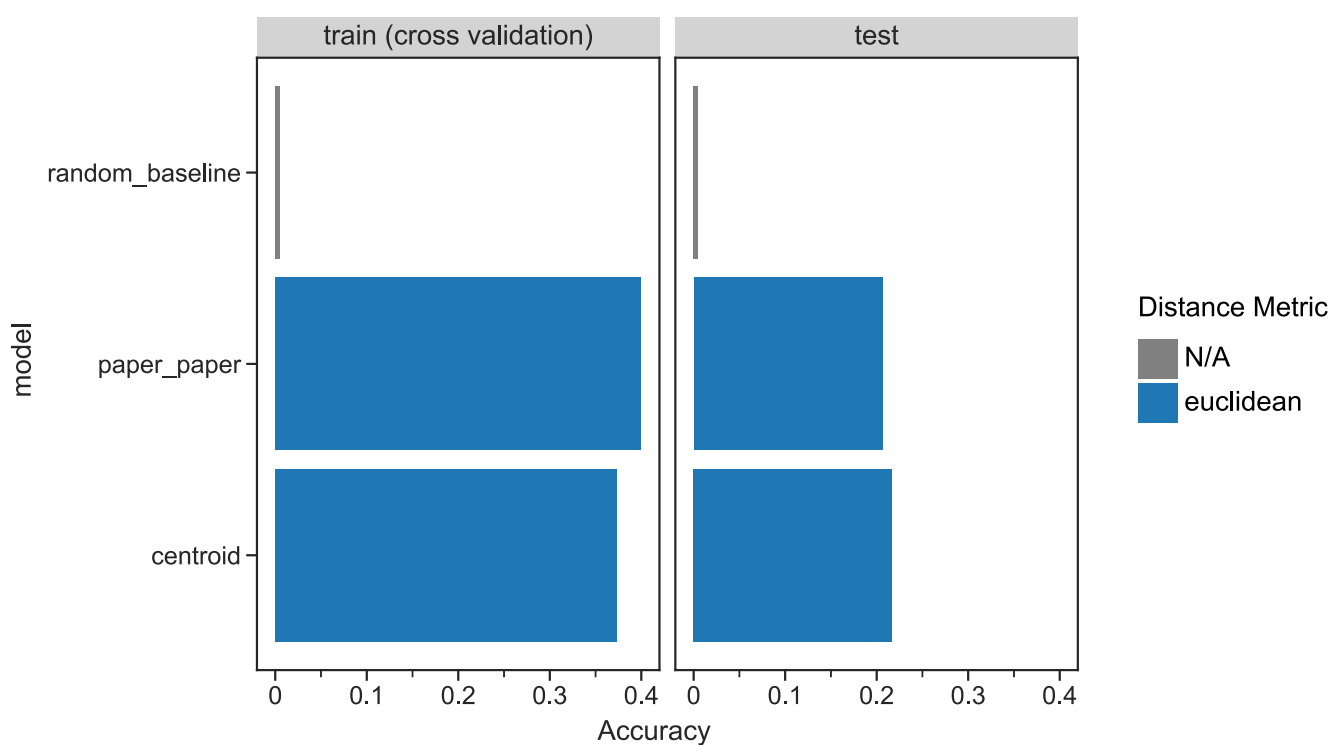
Figure S4: Both classifiers outperform the randomized baseline when predicting a paper's journal endpoint. This bargraph shows each model's accuracy in respect to predicting the training and test set.