Linguistic Analysis of the bioRxiv Preprint Landscape

Authors

- **David N. Nicholson**
  ⓘD [0000-0003-0002-5761](#) · ⓖ [danich1](#)
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · 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, University of Pennsylvania

- **Dongbo Hu**
  · ⓖ [dongbohu](#)
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania

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

- **Lawrence E. Hunter**
  ⓘD [0000-0003-1455-3370](#) · ⓖ [LEHunter](#)
  Center for Computational Pharmacology, University of Colorado Denver School of Medicine, Aurora, CO, USA

- **Casey S. Greene**
  ⓘD [0000-0001-8713-9213](#) · ⓖ [cgreene](#)
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania

Abstract

Preprints allow researchers to make their results quickly and widely accessible to the scientific community. These are scholarly works that have yet to undergo the peer review process and are often hosted within open access repositories such as bioRxiv. The majority of studies involving bioRxiv have focused on preprint metadata; however, a fundamental piece that is missing is an understanding of the language contained within preprints and how it changes through the peer review process. We sought to compare and contrast linguistic features within bioRxiv preprints to their published counterparts within Pubmed Central's Open Access corpus (PMC). We quantified the time delay preprints face while undergoing the peer review process. We generated document embeddings for every article within bioRxiv and PMC. We used these embeddings to idenitfy missing preprint-publication links along with training machine learning models to predict journal endpoints for published articles. We found that topic-specific terms such as "genome", "neuron" and "network" were enriched within bioRxiv compared to PMC, reflecting bioRxiv's uptake in bioinformatics and neuroscience preprints. We found that the leading source of linguistic variation among preprints captured the distinction between quantitative and cellular biology. We discovered that preprints are delayed an average of 16 days as changes are requested from the peer review process. Lastly, we created a web app allows anyone to input a bioRxiv or medRxiv preprint and receive a set of the most linguistically similar journals and articles to serve as potential publication venues for their work.

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 PMCOA 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, PMCOA, 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 KL 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 PMCOA 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 PMCOA corpus. A subset of our PMCOA 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 PMCOA corpus for training and initial evaluation via cross validation. 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-centric 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-centric 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 [45] 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 sklearn's implementation of k-d trees. To run cost effectively on AWS, 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 [46], 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 PMCOA 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 pre-trained 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 [47]. Because the rate of change is rapid, we first analyzed category data and compared our results with previous findings. As in previous reports [47], neuroscience remains the most common category of preprint followed by bioinformatics (Supplemental Figure S1). Microbiology, which was fifth in the most recent report [47], 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 [48]. 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 Comparison of bioRxiv and PubMed Central

**Table 1:** Generated corpora statistics for all corpus used in this project.

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

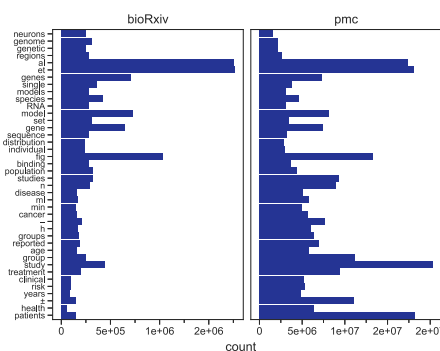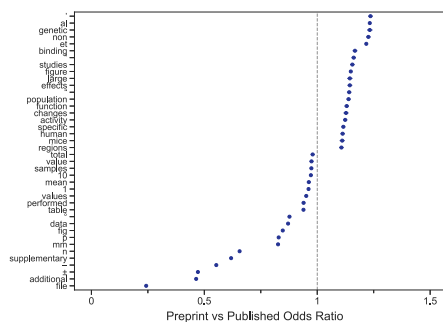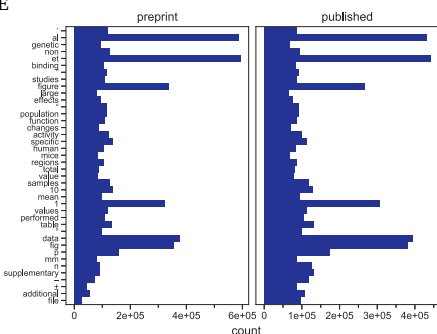| Metric	bioRxiv	PMC	NYTAC
coordinating conjunctions%	3.40%	3.57%	3.18%
pronouns	4,604,432	74,994,125	46,712,553
pronouns%	1.09%	0.87%	3.83%
passives	15,012,441	342,407,363	19,472,053
passive%	3.57%	3.98%	1.60%
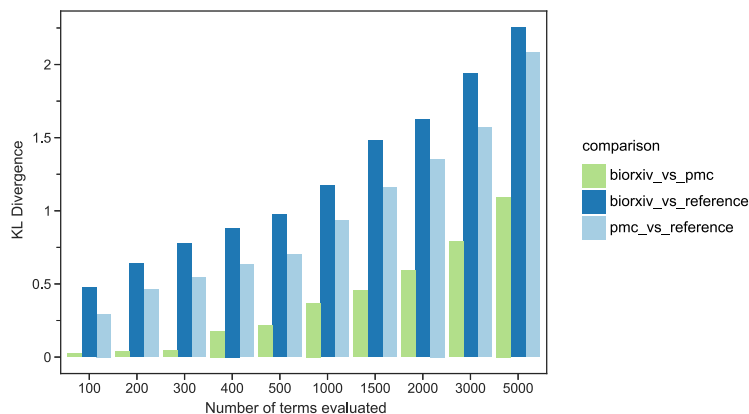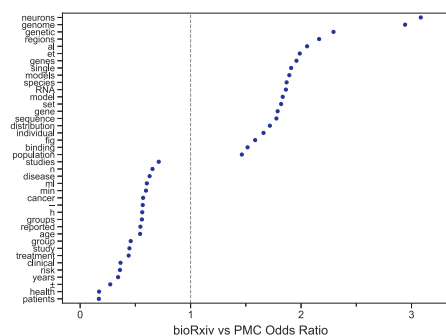
Figure 1: A. This barplot represents the KL divergence between bioRxiv, Pubmed Central and the reference corpus. The y-axis is the KL divergence metric where lower values indicates similar distributions and vice versa for higher values. The x-axis represents the number of highly occuring tokens used to calculate the KL divergence. B. This 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 PubMed Central. The dotted line provides a breaking point between both categories. C. This is a bar chart of token frequency appearing in bioRxiv and PMC respectively. D. This is a point range plot of the odds ratio with respect to preprints. Values greater than one indicate a high association with preprints while values less than one indicate a high association with published articles. The dotted line provides a breaking point between both categories. E. This is a barchart of token frequency appearing in preprints and published versions of preprints respectively.

The linguistic style of the bioRxiv corpus differs from the PMC corpus. We compared and contrasted preprints in bioRxiv, published manuscripts in PMC and newspaper articles from the New York Times

(NYTAC) against eachother. We refer to NYTAC as our reference corpus for the following analysis. We found that bioRxiv is more similar to PMC than to the reference in terms of token frequencies and corpora statistics (Figure ??A and Table 1). When comparing bioRxiv and PMC to the reference, topic associated and measurement related tokens appear highly enriched (Supplemental Figures S2 and S3). Furthermore, we found that tokens such as "neuron", "genome", "RNA" and "network" had a high odds ratio, while tokens such as "patient", "health", ±, and "ml" to have a low odds ratio when comparing bioRxiv to PMC (Figure ??B and ??C). This separation of tokens suggests that articles focused on clinical trials and patient research are more prevalent in PMC than to bioRxiv. This separation also suggests that bioRxiv has a predominance of preprints focused on neuroscience and bioinformatic topics. In regard to writing, citation styles diversify from the familiar "et al." form as preprints transition through the publication process. Additionally, published articles have an increase of typesetting (±) and measurement symbols ("ml", "age") compared to preprints.

A preprint's linguistic style can change once a preprint has undergone the revision process prior to being published. We quantified this linguistic difference by calculating the odds ratio of tokens appearing in the union of bioRxiv preprints and their published counterparts within PMC. Tokens with an odds ratio greater than one are mainly centered on paper/figure references and research specific terms (Figure ??D and ??E). Tokens with an odds ratio of less than one are focused on data availability, and research measurements such as number of cases and controls or significance testing (Figure ??D and ??E). This enrichment suggests that a key piece in the publication process is verifying that essential parts of research (e.g. data availability, specific measurements) are obvious to future readers within the scientific community.

Topic Analysis of bioRxiv's Principal Components
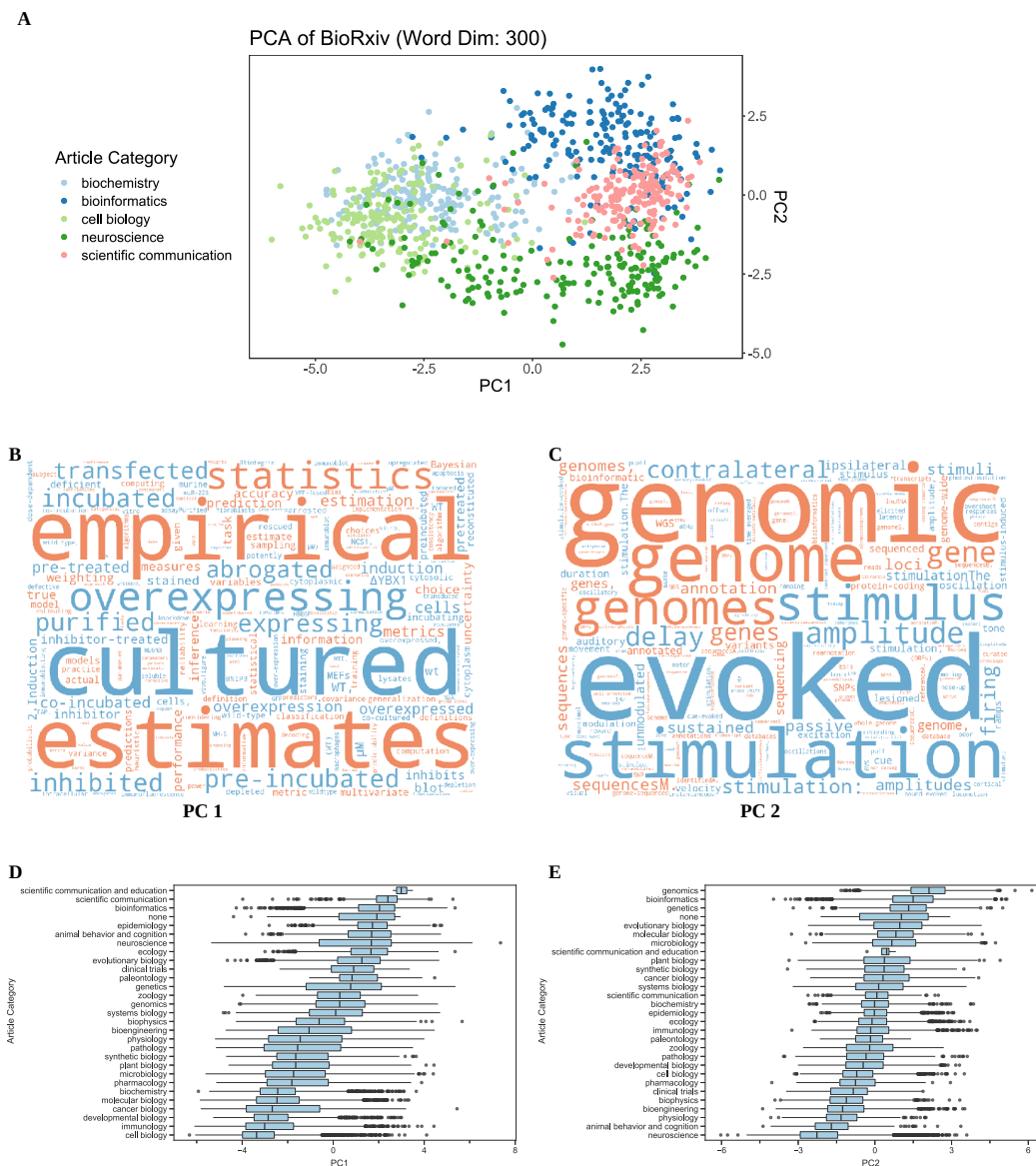
Figure 2: A. The scatter plot at the top right (A) is a visualization of documents being plotted along the PC directions. Article categories were hand-picked based on the concepts captured by each PC. B. The word cloud depicts the cosine similarity score between tokens and the second PC. Tokens in orange are most similar to a PC's positive direction while tokens in blue are most similar to a PC's negative direction. The size of each token indicates the magnitude of the similarity score. C. The word cloud depicts the cosine similarity score between tokens and the first PC. Tokens in orange are most similar to a PC's positive direction while tokens in blue are most similar to a PC's negative direction. The size of each token indicates the magnitude of the similarity score. D. This is box plot shows preprints in each article category projected along the PC1 direction. Negative values indicate molecular biology concepts, while positive values indicate quantitative biology concepts. E. This is box plot shows preprints in each article category projected along the PC2 direction. Negative values indicate neuroscience concepts, while positive values indicate bioinformatic concepts.

We explored the primary differences between the full text of bioRxiv preprints by performing principal components analysis on generated document embeddings. We visualized the correspondence between tokens and the loadings for each principal component (Figure 2). We also visualized documents projected on selected principal components (Figure 2A). The first principal component separates bioRxiv preprints that encompass molecular biology results with preprints that contain quantitative biology results (Figure 2B). This highlights the bisection of biomedical research where majority of results can be categorized under the molecular biology category or the quantitative biology category. Furthermore, this bisecting trend is evident across individual preprint categories as most categories lie on either side of the first principal component (Figure 2D). As with the first principal component we also provide example preprints from the systems biology category to reinforce this concept (Supplemental Table 2).

The second principal component represents the concept of neuroscience vs bioinformatics (Figure 2C). This principal component suggests that the bulk of preprints within bioRxiv are largely focused around neuroscience and bioinformatic concepts. This split is evident in Figure 2E as enriched categories along this principal component are quite related to neuroscience (negative end) or bioinformatics (positive end). We provide example preprints from the systems biology category to reinforce this concept (Supplemental Table 3). More principal component word clouds can be found on our journal recommender website and within our online repository (see Software and Data Availability).

Identifying preprints that were not linked with their corresponding publications
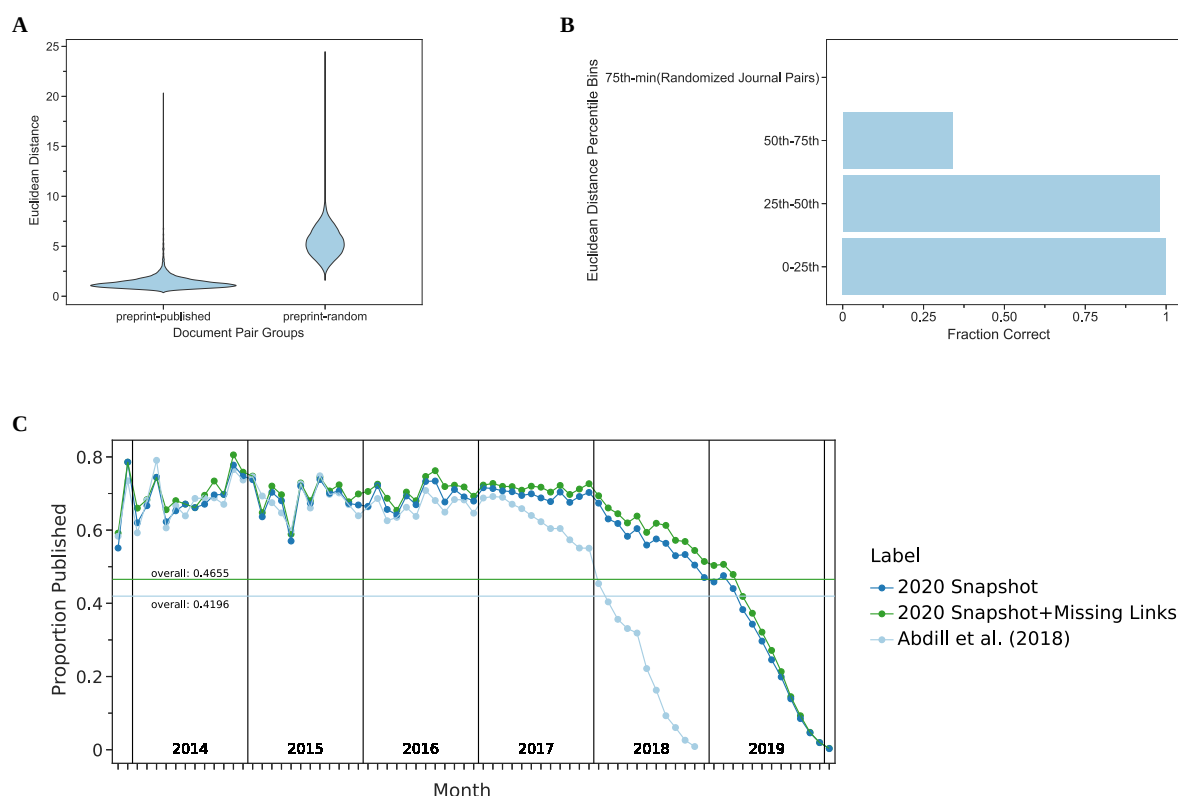
Figure 3: A. This violin plot shows the distribution of distances between the preprint-corresponding published version category and the preprint-randomly sampled articles category. B. This bar chart depicts the fraction of true positives over the total number of pairs in each bin. Each bin contains a total of 200 annotated pairs and is based on the percentiles of the preprint-published distribution. C. This line plot shows 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. [47]. The dark blue line represents the updated publication rate without missing links added while the dark green line is the updated publication rate with missing links added. The horizontal lines represent the overall proportion of preprints that are published.

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. Sproadically, there are cases where *bioRxiv* may miss detecting a link or authors may forget to notiy *bioRxiv* of their recent publcation. These missing links can make it more difficult to identify the latest version of scientific manuscripts and estimate the fraction of articles that are eventually published [47]. We used distance in the document space to identify preprints without an annotated publication but contained very similar content to published articles. We found that distances between preprints and their corresponding

published versions were lower than preprints paired with a random article published in the same journal (Figure 3A). This observation suggests that pairs with low embedding distances could be considered a true match, so we separated articles into quantiles based on the distribution of distances between true preprint-publication pairs. We curated 50 potential preprint-publication pairs from each of four quantiles and achieved a high inter-rater reliability of 91.7% (Cohen's Kappa [49]) for this task. Out of these two hundred pairs we found that approximately 98% of pairs with an embedding distance in the 0-25th and 25th-50th percentile bins were 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. [47]. Our filtering criteria were intentionally stringent, so the increased estimate of publication rate amounts to a few percent (Figure 3C). Many of these missed annotations were for preprints posted in the 2017-2018 interval. Compared to preprints published in 2019 and later, the preprints posted in 2017-2018 are old enough to have a high chance of being published; however, it is interesting that the rate for older preprints was not observed to be higher.

Factors that affect the time between preprinting and publication
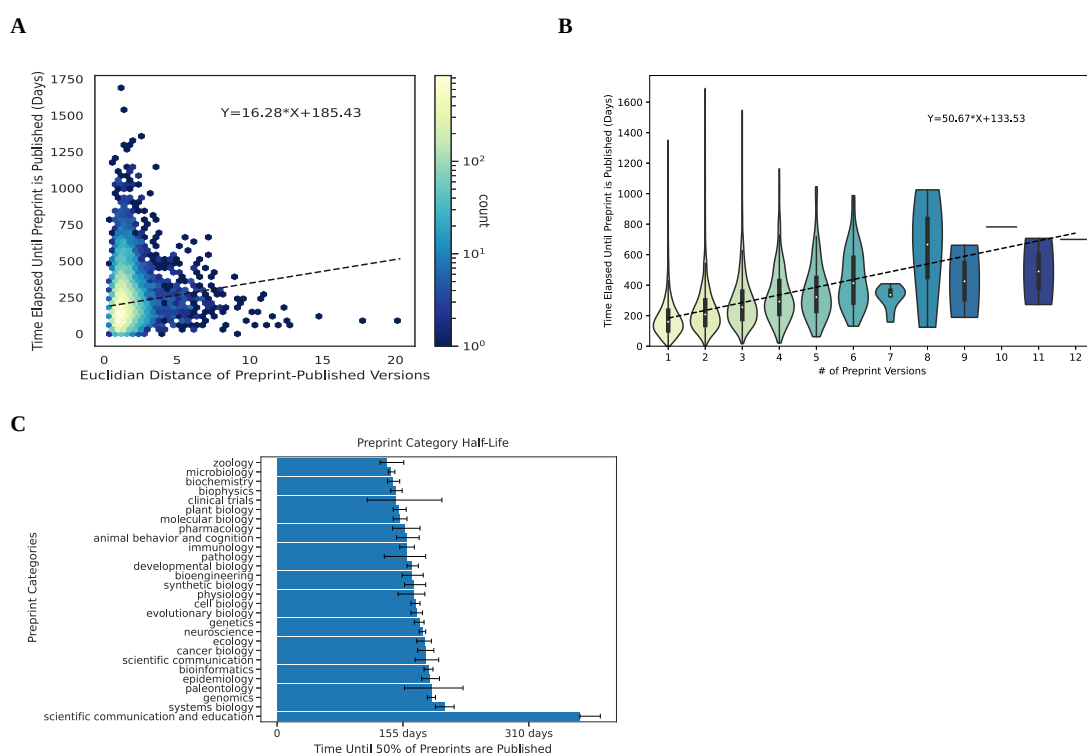
Figure 4: A.This hexbin plot depicts the amount of time it takes a preprint to be published against the distances of a preprint's first version and its corresponding published version. The x-axis represents the Euclidean distance between document representations, while the y-axis represents the number of days elapsed between 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 where more dense regions have a brighter color compared to their counterparts. B.This violin plot depicts the amount of time it takes a preprint to be published against the number of versions posted for a specific preprint. The x-axis represents the number of different versions a preprint has on bioRxiv, while the y-axis represents the number of days elapsed between a preprint posted on bioRxiv and the time a preprint is published. Inside each violin is a boxplot for the respective preprint version distribution. C.This bargraph depicts the amount of time it takes to get half of the total number of preprints published. The x-axis reprints days until 50% of preprints are published and the y-axis reprints the different preprint categories. The error bars represent 95% confidence intervals for each preprint's category half-life.

Preprints undergo multiple review checkpoints before they are published within a journal [50]. Oftentimes these checkpoints may result in rejection or revisions requested by a reviewer [50]. These

negative outcomes result in authors may having to drastically edit their preprint, which greatly impedes a preprint reaching a published endpoint. We sought to quantify the extent to which preprints are stalled when faced with a setback from the peer-review process. On average preprints are delayed approximately 16 days for every distance unit change (Figure 4A). We found that the average distance between two preprints' in the bioinformatics category was 5.068, which suggests that a single distance unit represents a fifth of a preprint's total text being changed. Sometimes preprints have to undergo drastic revisions that result in a new version being created. We found that on average it takes 51 days for authors to construct a new version of a preprint (Figure 4B). Both the document distance trend and the version number trend confirm that the larger the revision the longer it takes for a preprint to be published.

Preprints in certain categories take less time to publish than others. we sought to quantify the time each category takes to publish half their total number of preprints. Every preprint category takes at least 124 days to publish half of their respective preprints (Figure 4C). Categories that took the least amount of time were microbiology and zoology, while scientific communication and education took the most time (Figure 4C). Overall, this suggests that preprints in the microbiology and zoology categories may face less peer-review setbacks compared to other categories.

Recommending Journals Based on Preprint Representation

We sought to identify journals that might publish a preprint based on the text of a paper. We trained two different classifiers to predict the journal endpoints for already published papers. One classifier uses the nearest journal centroids, which attempts to capture the topic area of a journal. The other classifier aims to be more granular and recommends journals based on close proximity of individual papers. Both classifiers achieved a substantial increase over the random baseline; however, our predictors are not perfect (Supplemental Figure S4). This is expected as our dataset contains 2516 different journals where some journals publish papers that cover very specific topic while others publish papers that have a broad set of covered topics. Our journal centroid classifier performed better than the nearest paper classifier on the held out test set (Supplemental Figure S4). Overall, our software provides a starting point for authors to use the text of their preprints to identify potentially suitable publication venues.

Home Screen (A)

Most Similar Papers Screen (B)

Journal Centroid Recommendations (C)

Pubmed Central Landscape (D)

Square Bin Properties (E)
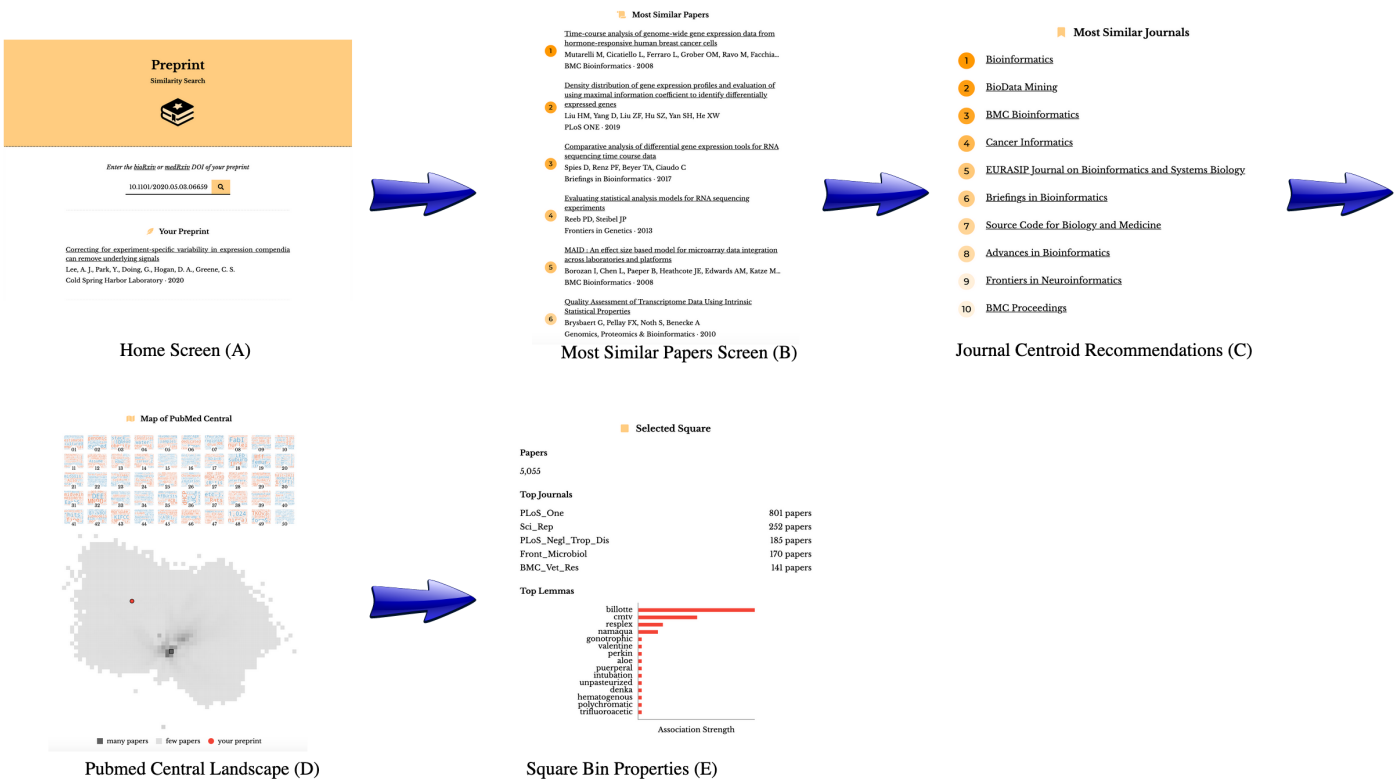
Figure 5: Here is the workflow of the journal recommender web-app. Starting with the homescreen users can paste in a *bioRxiv* or *medRxiv* DOI, which sends a request to biorxiv or medrxiv (A). Next our app preprocesses the preprint and returns a listing of the top ten most similar papers (B) and the top ten closest journals to the query (C). Following the listing, our app manually plots the preprint query onto the Pubmed Central Landscape (D). Lastly, users can click on a square within the landscape, which will show bin statistics as well as associated word-odd ratios (E).

We constructed an online app that provides users with journal suggestions based on their preprint content. Users supply DOIs from *bioRxiv* or *medRxiv*. The application then downloads the article, 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. It also embeds the document into the overall PMC landscape for visualization and allows the user to examine principal components and term enrichment for each bin within the landscape (Figure 5).

Discussion

We analyzed the language contained used in preprints and examined how it changes through the publication process. We found that bioRxiv and PubMed Central (PMC) have similar word frequency distributions, which suggests that the overall manner of writing is consistent with the biomedical literature. At the token level, those most strongly associated with bioRxiv are related to neuroscience and bioinformatics, which are also fields that have seen high uptake of preprinting [47]. We noticed that a multitude of preprints highly associated with the first principal component have restrictive or no copyright license (Supplemental Table 2). This finding highlights the ongoing problem of restricted access within the scientific community [51,52]. We also found that the second principal component for our language embedding differentiated neuroscience and bioinformatics papers.

We examined preprints that were textually similar to published articles and found numerous missing links between preprints and their publsihed counterparts. This observation led us to find that the life sciences preprint publication rate is higher than previously estimated (Figure ??). Preprint-publication similarity also predicts journal endpoints with modest performance for already published articles. This observation enabled us to provide a web application that allows users to identify the papers and journals that are most similar to a *bioRxiv* or *medRxiv* preprint.

Conclusion and Future Directions

Our linguistic analysis did not reveal substantial changes in the language during the peer-reviewed publishing process. The tokens most strongly associated with the peer reviewed form, as opposed to the preprint form, were associated with data availability and statistical reporting. We found that embeddings of preprints and publications could be compared and that distance in this space was meaningful in terms of topic area and the journal of eventual publication. Being able to identify similar preprints and publications using text content makes it feasible to begin tackling more detailed questions, and our analytical software is all open source to enable others to build upon them. The analysis of preprints' full text can support new tools that accelerate publishing, integrity checks, and other critically important contributions.

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.

Acknowledgements

Competing Interest

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

46. **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

47. **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

48. **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

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

50. **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

51. **Biologists debate how to license preprints**
Lindsay McKenzie
Nature (2017-06-16) https://doi.org/b9fb
DOI: 10.1038/nature.2017.22161

52. **The licensing of *bioRxiv* preprints**
Daniel Himmelstein
Satoshi Village (2016-12-05) https://blog.dhimmel.com/biorxiv-licenses/

53. **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

54. **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

55. **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

56. **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

57. **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

58. **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

59. **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

60. **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

61. **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

62. **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

63. **Pangenome Analysis of Enterobacteria Reveals Richness of Secondary Metabolite Gene Clusters and their Associated Gene Sets**
Omkar S. Mohite, Colton J. Lloyd, Jonathan M. Monk, Tilmann Weber, Bernhard O. Palsson
bioRxiv (2019-09-25) https://doi.org/gg94bk
DOI: 10.1101/781328

64. **QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci**
Fan Lin, Jue Fan, Seung Y. Rhee
bioRxiv (2019-04-29) https://doi.org/gg94bd
DOI: 10.1101/484204

65. **Identification of candidate genes underlying nodulation-specific phenotypes in *Medicago truncotula* through integration of genome-wide association studies and co-expression networks**
Jean-Michel Michno, Liana T. Burghardt, Junqi Liu, Joseph R. Jeffers, Peter Tiffin, Robert M. Stupar, Chad L. Myers
Cold Spring Harbor Laboratory (2018-08-16) https://doi.org/gg94bb
DOI: 10.1101/392779

66. **Raw sequence to target gene prediction: An integrated inference pipeline for ChIP-seq and RNA-seq datasets**
Nisar Wani, Khalid Raza
Cold Spring Harbor Laboratory (2017-11-16) https://doi.org/gg9394
DOI: 10.1101/220152

67. **The y-ome defines the thirty-four percent of *Escherichia coli* genes that lack experimental evidence of function**
Sankha Ghatak, Zachary A. King, Anand Sastry, Bernhard O. Palsson
bioRxiv (2018-12-03) https://doi.org/gg9396
DOI: 10.1101/328591

68. **The effects of time-varying temperature on delays in genetic networks**
Marcella M Gomez, Richard M Murray, Matthew R Bennett
Cold Spring Harbor Laboratory (2015-09-24) https://doi.org/gg939x
DOI: 10.1101/019687

69. **An analog to digital converter creates nuclear localization pulses in yeast calcium signaling**
Ian S Hsu, Bob Strome, Sergey Plotnikov, Alan M Moses
Cold Spring Harbor Laboratory (2018-06-28) https://doi.org/gg9397
DOI: 10.1101/357939

70. **Nicotinic modulation of hierarchal inhibitory control over prefrontal cortex resting state dynamics: modeling of genetic modification and schizophreniarelated pathology**
Marie Rooy, Fani Koukouli, Uwe Maskos, Boris Gutkin
Cold Spring Harbor Laboratory (2018-04-13) https://doi.org/gg9395
DOI: 10.1101/301051

71. **Electrical propagation of vasodilatory signals in capillary networks**
Pilhwa Lee
Cold Spring Harbor Laboratory (2019-11-13) https://doi.org/gg94bn
DOI: 10.1101/840280

72. **Dendritic spine geometry and spine apparatus organization govern the spatiotemporal dynamics of calcium**
Miriam Bell, Tom Bartol, Terrence Sejnowski, Padmini Rangamani
Cold Spring Harbor Laboratory (2019-05-29) https://doi.org/gg9399
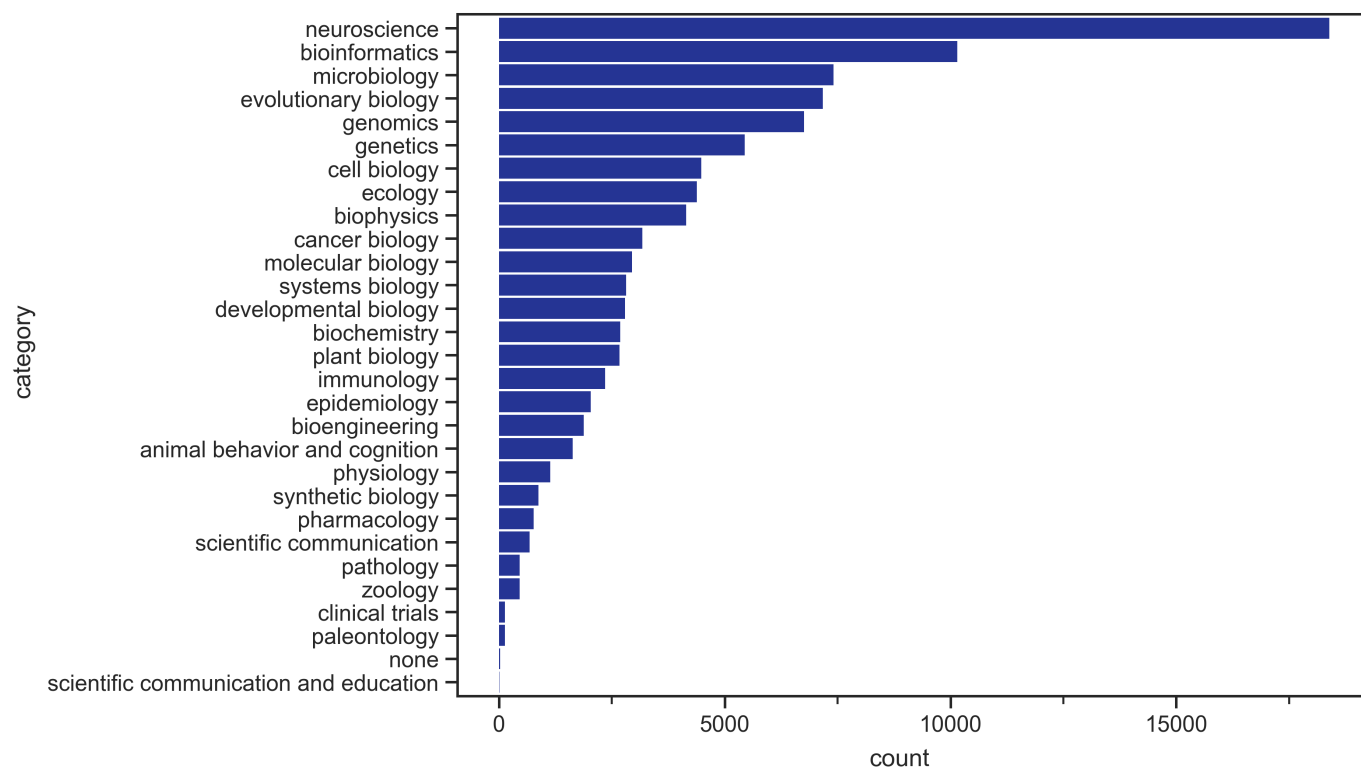DOI: 10.1101/386367

Supplemental Figures

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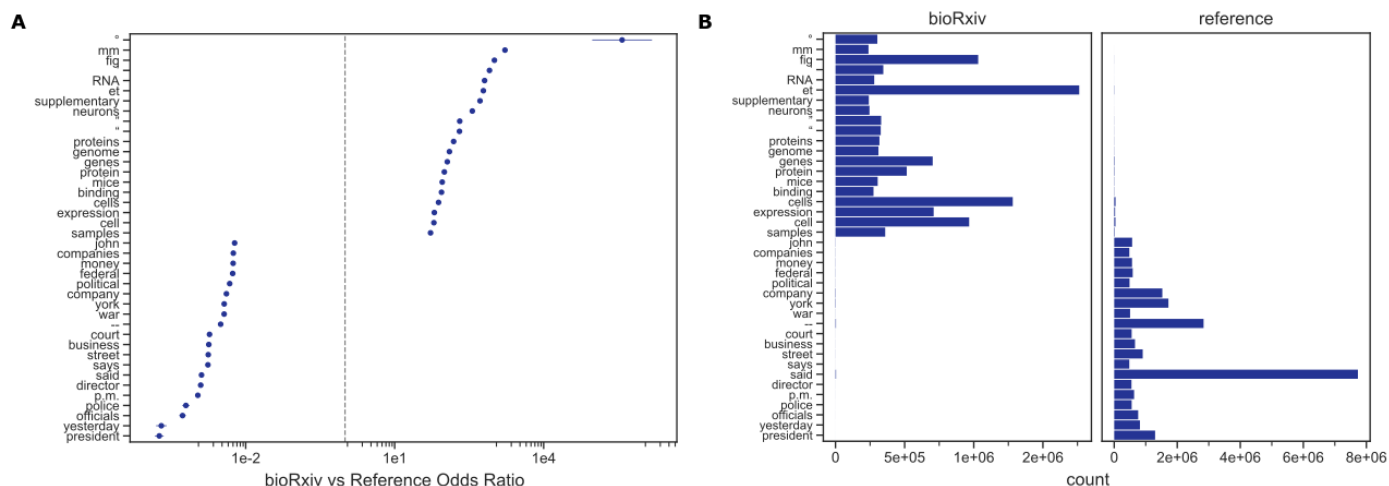
Figure S2: Topic associated tokens are highly enriched when comparing bioRxiv to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to bioRxiv. Values greater than one indicate a high association with bioRxiv whereas values less than one indicate high association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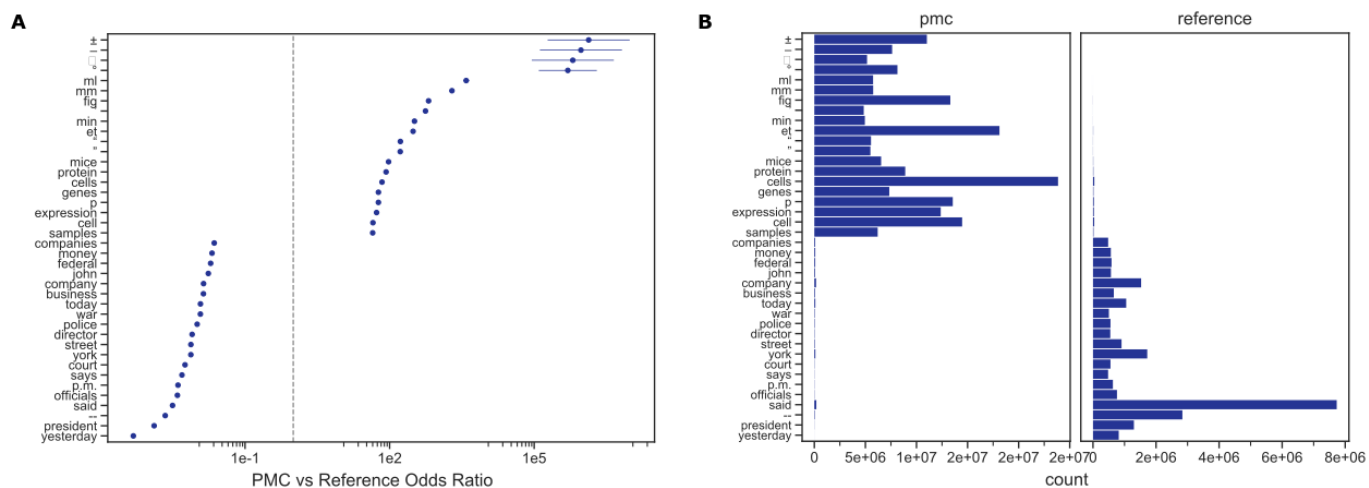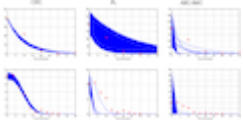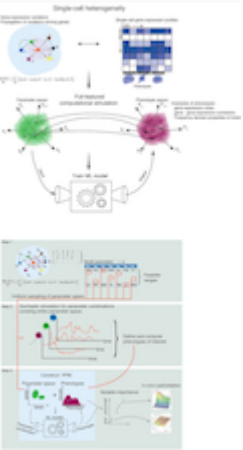
Figure S3: Typesetting symbols and biologically relevant tokens are highly enriched when comparing PubMed Central (PMC) to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to PMC. Values greater than one indicate a high association with PMC whereas values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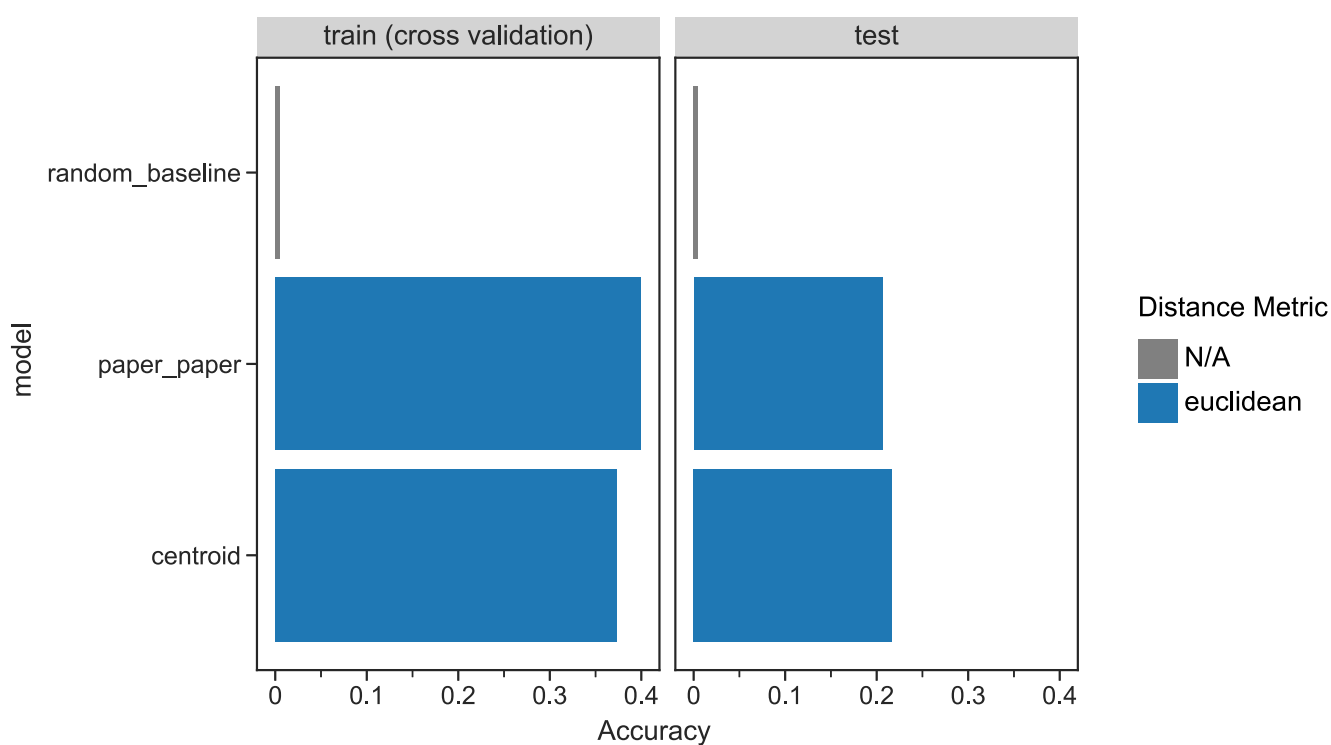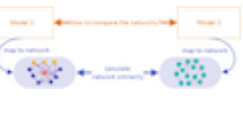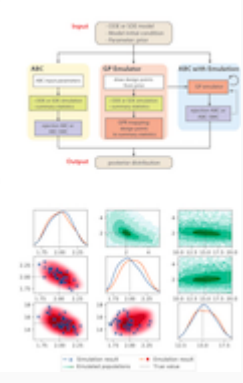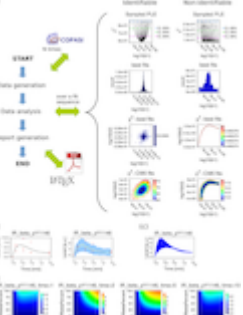
Figure S4: Both classifiers outperform the randomized baseline when predicting a paper's journal endpoint. This bargraph shows each model's accuracy in respect to predicting the training and test set.

Supplemental Tables

Table 2: Top and bottom five systems biology preprints projected onto the PC1 direction. These preprints contain quantitative and molecular biology concepts respectively.

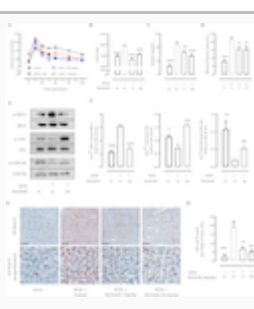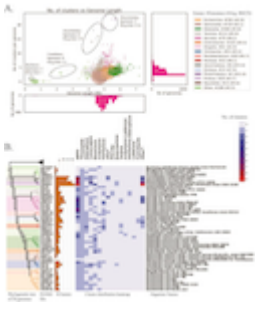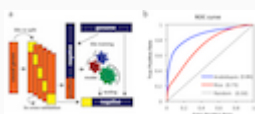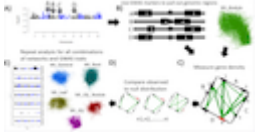| Title [citation]	PC_1	License	Figure Thumbnail	Figure Link
Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis [53]	4.700554908074704	None		https://www.biorxiv.org/content/biorxiv/early/2017/10/02/197400/F1.large.jpg
Machine learning of stochastic gene network phenotypes [54]	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 [55]	4.355295926618207	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2016/03/21/044818/F1.large.jpg
GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation [56]	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 [57]	4.321847854182741	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2017/02/09/107250/F1.large.jpg
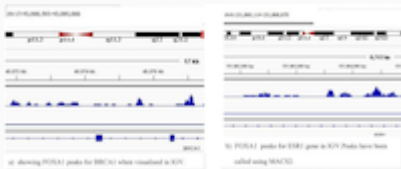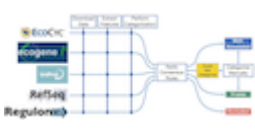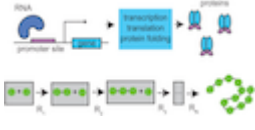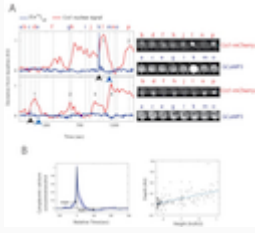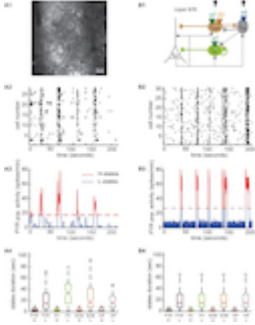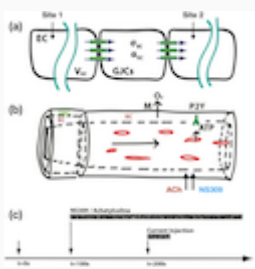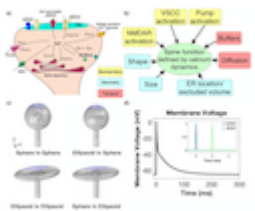
Title [citation]	PC_1	License	Figure Thumbnail	Figure Link
Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection [58]	-4.263964235099807	CC-BY-ND		https://www.biorxiv.org/content/biorxiv/early/2018/11/07/455048/F1.large.jpg
Systems analysis by mass cytometry identifies susceptibility of latent HIV-infected T cells to targeting of p38 and mTOR pathways [59]	-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 [60]	-4.592209988884236	None		https://www.biorxiv.org/content/biorxiv/early/2019/08/13/733162/F1.large.jpg
Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease [61]	-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 [62]	-4.826793742506695	None		https://www.biorxiv.org/content/biorxiv/early/2019/09/01/754572/F1.large.jpg

Table 3: Top and bottom five systems biology preprints projected onto the PC2 direction. These preprints contain bioinformatis and neuroscience concepts respectively.

Title [citation]	PC_2	License	Figure Thumbnail	Figure Link

Title [citation]	PC_2	License	Figure Thumbnail	Figure Link
Pangenome Analysis of Enterobacteria Reveals Richness of Secondary Metabolite Gene Clusters and their Associated Gene Sets [63]	3.5865702659438883	CC-BY-ND		https://www.biorxiv.org/content/biorxiv/early/2019/09/25/781328/F1.large.jpg
QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci [64]	3.470388383023157	None		https://www.biorxiv.org/content/biorxiv/early/2019/04/29/484204/F1.large.jpg
Identification of candidate genes underlying nodulation-specific phenotypes in Medicago truncotula through integration of genome-wide association studies and co-expression networks [65]	3.3814906334073953	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2018/08/16/392779/F1.large.jpg
Raw sequence to target gene prediction: An integrated inference pipeline for ChIP-seq and RNA-seq datasets [66]	3.3632576028389742	None		https://www.biorxiv.org/content/biorxiv/early/2017/11/16/220152/F3.large.jpg
The y-ome defines the thirty-four percent of Escherichia coli genes that lack experimental evidence of function [67]	3.28742786641417	CC-BY		https://www.biorxiv.org/content/biorxiv/early/2018/12/03/328591/F1.large.jpg

Title [citation]	PC_2	License	Figure Thumbnail	Figure Link
The effects of time-varying temperature on delays in genetic networks [68]	-2.7047102478958056	None		https://www.biorxiv.org/content/biorxiv/early/2015/09/24/019687/F1.large.jpg
An analog to digital converter creates nuclear localization pulses in yeast calcium signaling [69]	-2.775745000260895	None		https://www.biorxiv.org/content/biorxiv/early/2018/06/28/357939/F1.large.jpg
Nicotinic modulation of hierarchal inhibitory control over prefrontal cortex resting state dynamics: modeling of genetic modification and schizophreniarelated pathology [70]	-3.047342382798414	None		https://www.biorxiv.org/content/biorxiv/early/2018/04/13/301051/F1.large.jpg
Electrical propagation of vasodilatory signals in capillary networks [71]	-3.107715578793087	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2019/11/13/840280/F1.large.jpg
Dendritic spine geometry and spine apparatus organization govern the spatiotemporal dynamics of calcium [72]	-3.21533499072831	CC-BY-NC-ND		https://www.biorxiv.org/content/biorxiv/early/2019/05/29/386367/F1.large.jpg