Linguistic Analysis of the bioRxiv Preprint Landscape

This manuscript (permalink) was automatically generated from greenelab/annorxiver_manuscript@8faae39 on October 8, 2020.

Authors

- **David N. Nicholson**
  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)

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

1. What is a preprint
2. Why are preprints important?
3. Mention how preprints are being integrated into scientist's everyday workflow
4. Talk about biorxiv and discuss how it is one of the repositories that maintains preprints along with citation of others such as arxiv/medrxiv etc.
5. Discuss works that analyze biorxiv from an audience perspective (quantifying tweets etc.)
6. Mention the gap which consists of analysing the language of biorxiv preprints (first to do this)
7. ^ Why is this important? What will this allow for future research projects?
8. Provide list of contributions within this manuscript

Methods

Datasets

BioRxiv

BioRxiv [1] is a repository of biological and biomedical research preprints. We downloaded an xml snapshot of this repository on February 3, 2020 from bioRxiv's Amazon S3 resource [2] that contained the full text and image content of 98,023 preprints. Preprints on bioRxiv are versioned, and in our snapshot 26,905 of 98,023 contained more than one version. When preprints had multiple versions, we used only the latest one. Preprints in this snapshot were grouped by researchers submitting to *bioRxiv* into one of twenty-nine different categories. Each preprint was also classified as a new result, confirmatory finding, or contradictory finding. Some preprints in this snapshot have been withdrawn from bioRxiv. When a preprint is withdrawn, its content is replaced with the reason for withdrawal. Because we used the latest version, withdrawn preprints in our analysis contained only statements indicating their removal.

PubMed Central

PubMed Central (PMC) [3] is a repository that contains free-to-read articles. PMC contains two types of contributions: closed access articles from research funded by the United States National Institutes of Health (NIH) appearing after an embargo period and articles published under Gold Open Access [4] publishing schemes. Paper availability within PMC is largely dependent on the journal's participation level [5]. Individual journals have can fully participate in submitting articles to PMC, selectively participate sending only a few few of papers to PMC, only submit papers according to NIH's public access policy [6], or not participate at all. As of September 2019, PMC had 5,725,819 articles available [7]. Out of these 5 million articles, about 3 million were open access and available for text processing systems [8,9]. We downloaded a snapshot of this open access subset on January 31, 2020. This snapshot contains papers such as literature reviews, book reviews, editorials, case reports, research articles and more; however, we used only the research articles.

Comparing Corpora

We compared bioRxiv against Pubmed Central's Open Access corpus (PMCOA) and the New York Times Annotated corpus (NYTAC) [10] to assess the similarities and differences between bioRxiv, PMCOA and NYTAC. Throughout our analysis we encountered non-word symbols (e.g., ±), so we refer words and symbols 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 [11] (version 2.2.3) to preprocess all corpora and filtered out 326 spaCy-provided stopwords.

Following the cleaning process, we calculated the frequency of every token across all corpora. Because many tokens were unique to one set or the other and observed at low frequency, we used the union of the top 100 most frequent tokens from each corpus to compare them. We generated a contingency table for each token and calculated the odds ratio from every generated table. We also calculated the 95% confidence interval for each token's odds ratio [**???**/] and measured corpus similarity by calculating the KL divergence across all three corpora.

Visualizing the Preprint Landscape

Generate Document Representation

We used gensim [**???**] (version 3.8.1) to train a word2vec continuous bag of words (CBOW) [12] 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 otherwise used gensim's default settings. Following training, we generated a document vector for every article within bioRxiv and PubMed Central. This document vector is calculated by taking the average of every token present within a given article, ignoring those that were absent from the word2vec model.

Dimensionality Reduction of Document Embeddings

We used principal component analysis (PCA) [13] to project bioRxiv document vectors into a low dimensional space. We trained this model using scikit-learn's [14] implementation of a randomized solver [15] with a random seed of 100, output of 50 principal components, and default settings for all other parameters. For each principal component we calculated its cosine similarity with 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 (blue) or negative (orange) association.

Discovering Unannotated Preprint-Publication Relationships

Automated procedures are in place to link preprints to peer reviewed versions and many journals require authors to update preprints with a link to the published version. However, automated procedures at *bioRxiv* are often based on exact matching of certain attributes and authors can forget to establish a link after publication. For example, authors can change the title between a preprint and published version (e.g., [16] and [17]), 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 the distance in embedding space could reveal preprint to published version links that were missed by existing automated processes. First, we used CrossRef [18] to identify bioRxiv preprints that were linked to a corresponding published article. We filtered out links that contained papers not in PubMed Central's Open Access corpus. Following the preprocessing step, 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 the 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 the published version. 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 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 candidates to two scientists 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 reviewers. The preprint-publication pairs on which reviewers disagreed were supplied to a third scientist, who carefully reviewed each case and made a final determination. Lastly, 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 verisons.

Journal Recommendation

Determining the best journal venue for a preprint is a non-trivial task as there are too many options for authors to decide. We sought to provide a resource that recommends journals based on a preprint's embedding representation. We illustrate our recommendations as a short list along with a network visualization available at https://greenelab.github.io/annorxiver-journal-recommender/. Since we sought to examine if embeddings were related to publication venue, we used a simple k-nearest neighbors approach with Euclidean distance to recommend journals.

First, we filtered all journals that had fewer than 100 papers in the PMC dataset. 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 our gold standard test set. We used the remainder of the PMC corpus for training and initial evaluation via cross validation. We 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 for recommending journals.

The first approach is based on individual paper proximity, which enabled us to provide an example of the specific article or articles that led to the prediction. Conversely, predictions using this technique could be biased due to the overrepresentation of general topic journals. We call this approach the paper-based classifier. This classifier takes a query article that has been projected onto the embedding space trained on bioRxiv preprints as input and reports the journals of the top ten closest papers. 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.

The second approach is based on close proximity to a journal's centroid. This technique provides recommendations that are more focused on domain-specific publication venues. We call this approach the journal-based classifier. This classifier was trained by computing journal centroids via taking the average embedding of all papers published in each journal. Following the centroid calculation, this classifier takes a query article projected onto the same embedding space as above for

input and reports the top ten nearest journals centroids. Both the paper-based classifier and the journal-based classifier were optimized via 10-fold cross validation. We evaluated performance of both classifiers on our gold standard test set of published preprints.

We used SAUCIE [19] to train a model that uses the latent space of a neural network to learn an embedding suitable for visualization. This model enabled us to visualized the PMC corpus and to efficiently embed new papers and preprints with the space. We trained this model using a learning rate of 0.001, lambda_b of 0, lambda_c of 0.001, and lambda_d of 0.001 for 2000 iterations. 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 PubMed Central

bioRxiv Repository

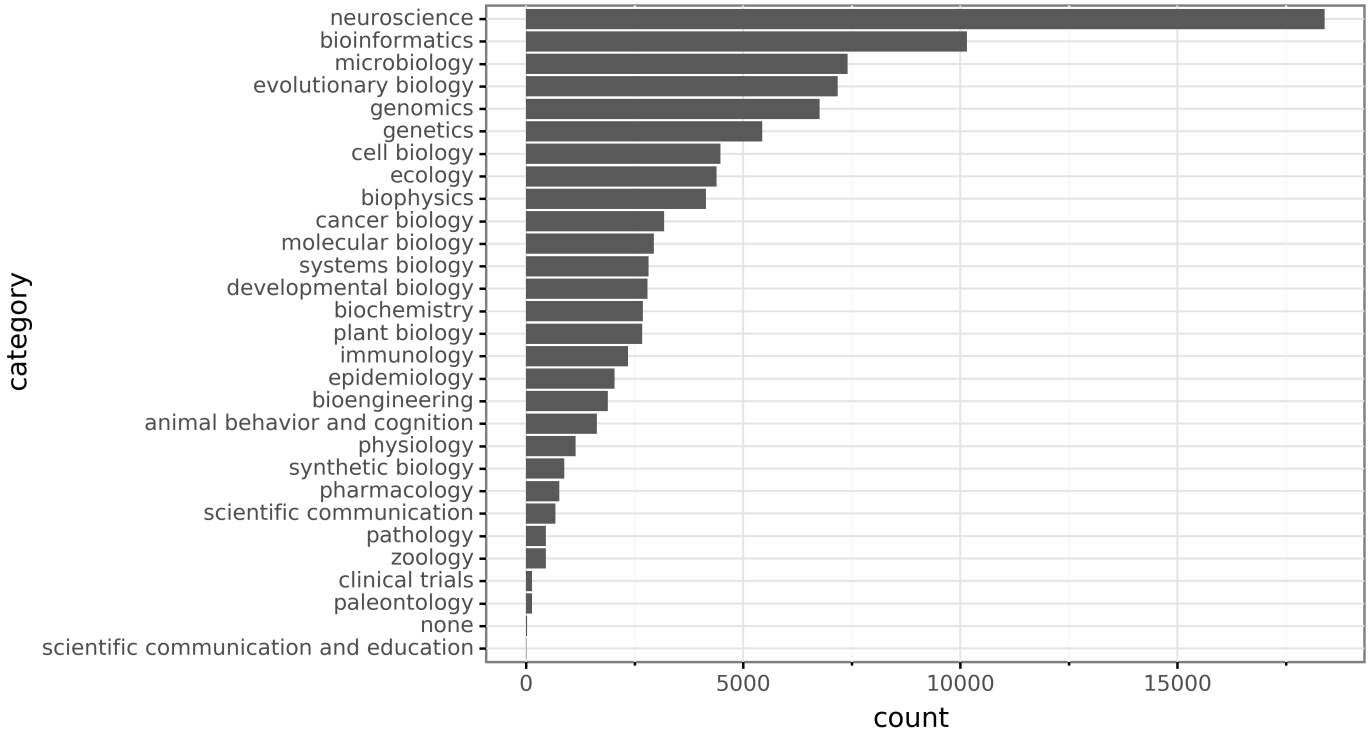

Figure 1: Neuroscience and bioinformatics are the two most common topics for preprints on bioRxiv. This bar chart depicts the number of preprints that fall into each author-selected topic area.

Each preprint on bioRxiv has an author-selected topic area, and the predominant area in past reports has been neuroscience [20]. Our analysis of the full text release of bioRxiv confirms this previous finding (Figure 1). The author-selected topic area abundances that we found in the full text largely matched previous studies [20,21]. One exception was microbiology, which has a larger share of preprints than in a previous report from 2018 [20] (Figure 1). Authors also select from three article types when they upload their preprints. We found that nearly all preprints were categorized as new results, which is consistent with previous findings [21].

Global Comparison of bioRxiv and PubMed Central

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

Metric	bioRxiv	PMC	NYTAC

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

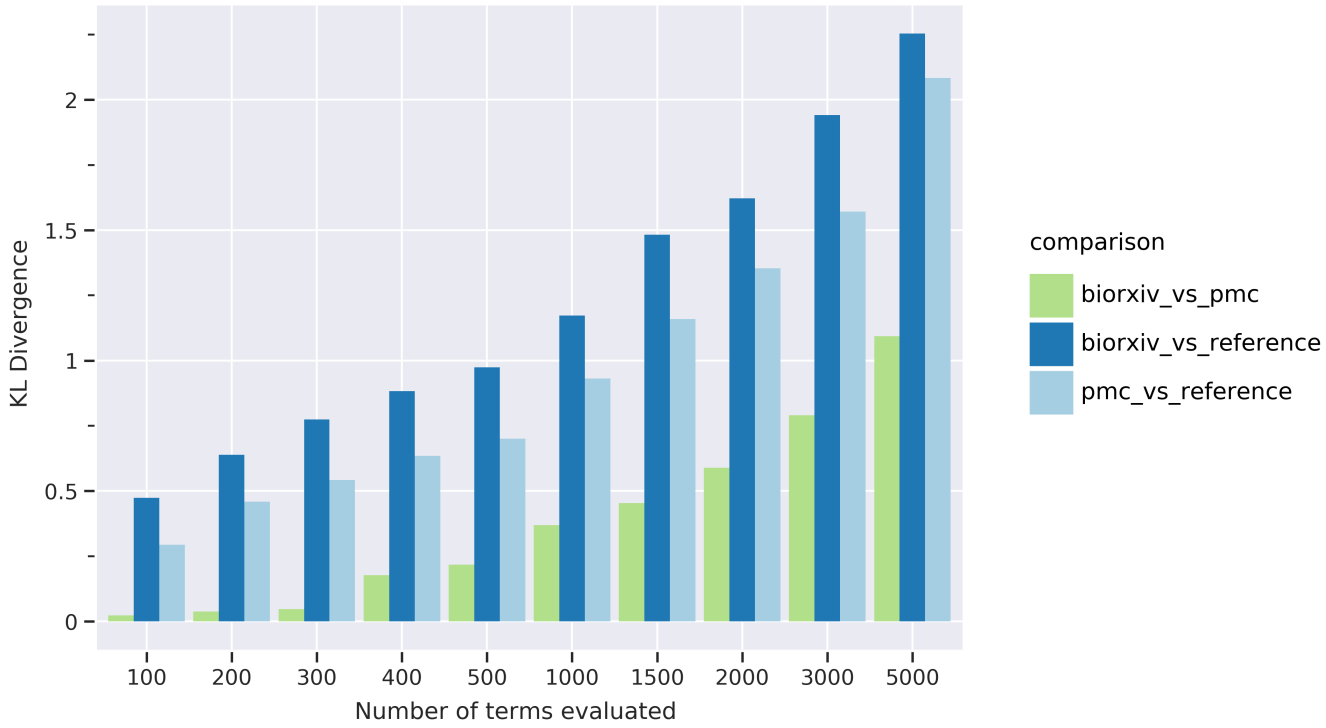

Figure 2: BioRxiv is more similar to PubMed Central than to the reference corpus. This barplot represents the KL divergence between bioRxiv, Pubmed Central and the reference corpus. The y-axis is the KL divergence metric where lower values indicates similar distributions and vice versa for higher values. The x-axis represents the number of highly occuring tokens used to calculate the KL divergence.

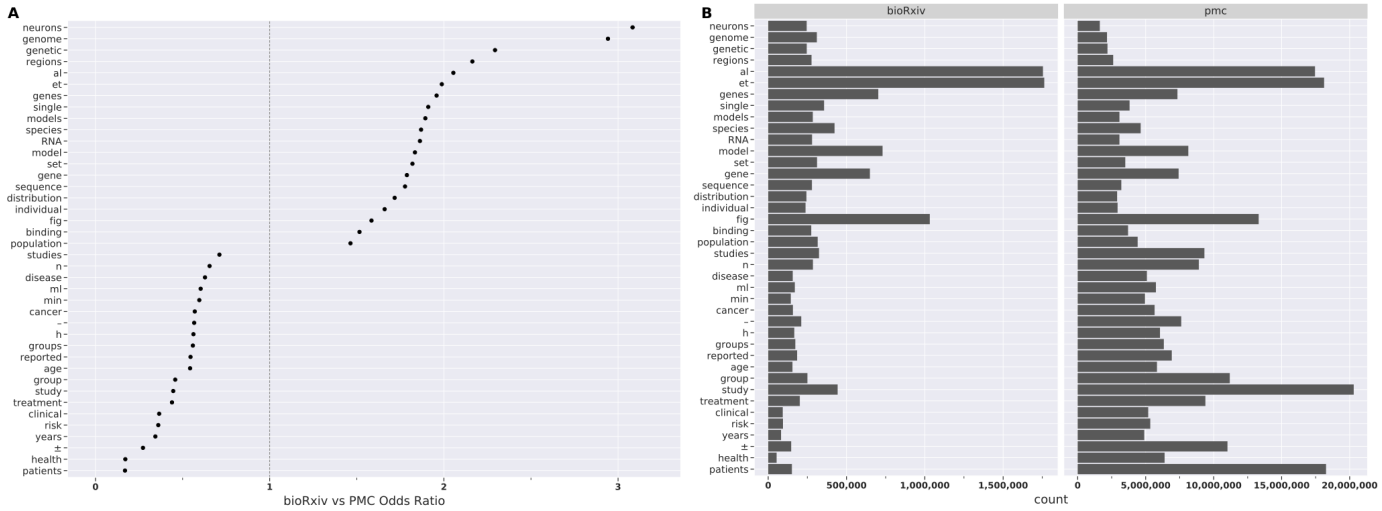

Figure 3: BioRxiv is more focused on biological discoveries rather than disease treatments and clinical trials. The plot on the left (A) is a point range plot of the odds ratio with respect to bioRxiv. Values greater than one indicate a high association with bioRxiv whereas values less than one indicate high association with PubMed Central. 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 PMC 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 2 and Table 1). When comparing bioRxiv and PMC to the reference, topic associated and measurement related tokens appear highly enriched (Supplemental Figures 13 and 14). 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 3). 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 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.

Published Preprint Differences

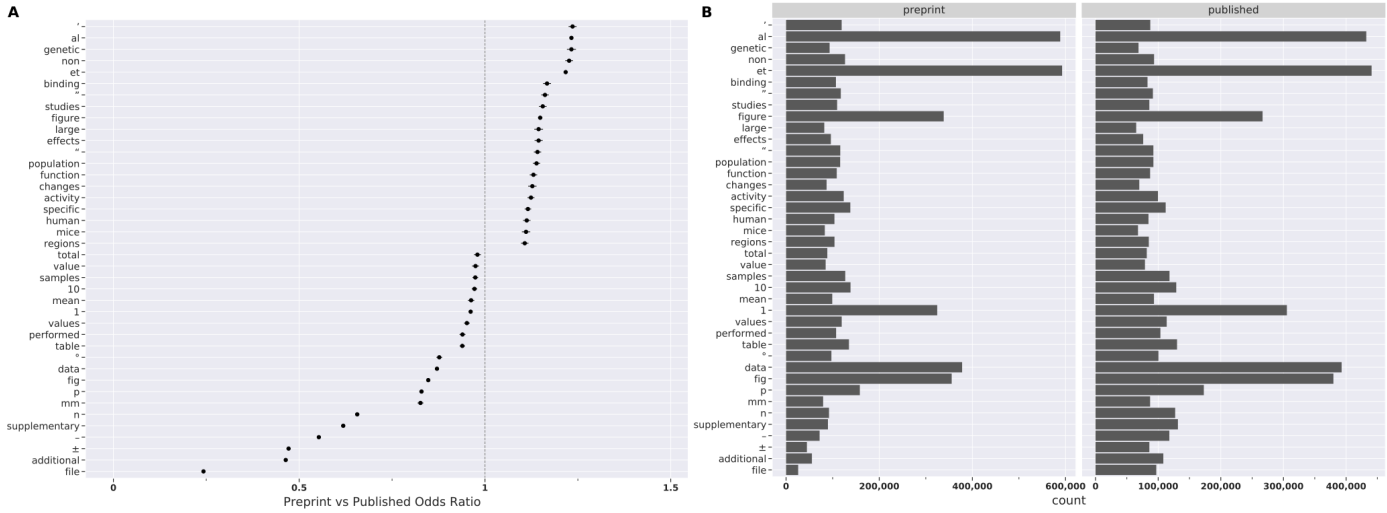

Figure 4: Top scoring tokens for preprints are focused on figure citations whereas their published versions are more focused on data availability. The plot on the left (A) is a point range plot of the odds ratio with respect to preprints. Values greater than one indicate a high association with preprints while values less than one indicate a high association

with published articles. The dotted line provides a breaking point between both categories. The plot on the right (B) is a barchart of token frequency appearing in preprints and published versions of preprints respectively.

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 4). 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 4). 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 5: The top two principal components (PCs) appear to capture the concepts of molecular biology vs quantitative biology (PC1) and neuroscience vs bioinformatics (PC2). The word clouds (A, C) depict the cosine similarity score between tokens and the first two PCs. 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. The scatter plot at the top right (B) is a visualization of documents being plotted along the PC directions. Article categories were hand-picked based on the concepts captured by each PC.

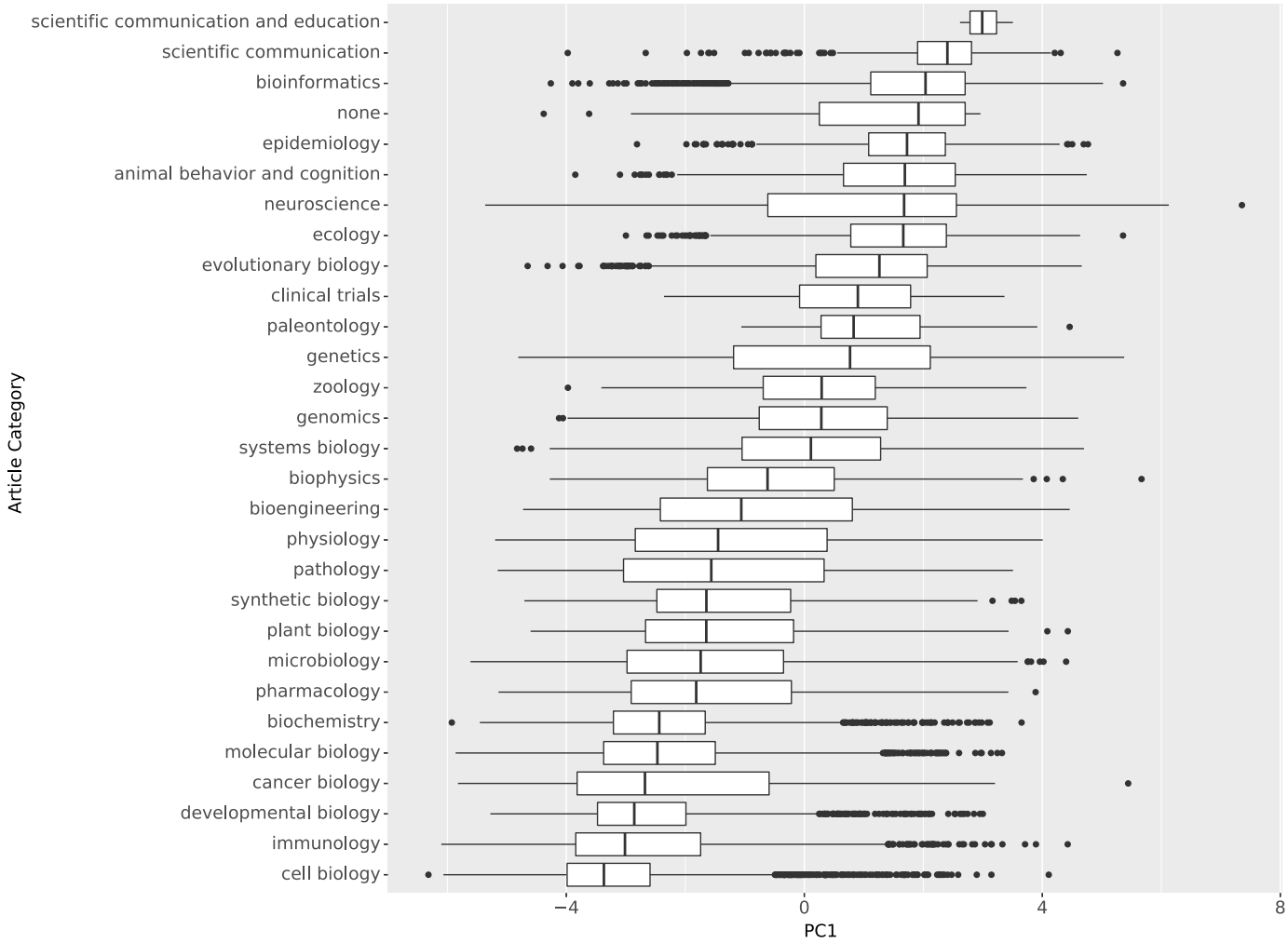

Figure 6: Preprint categories have a diverse spread of quantitative and molecular biology results. This is box plot shows preprints in each article category projected along the PC1 direction. Negative values indicate molecular biology concepts, while positive values indicate quantitative biology concepts.

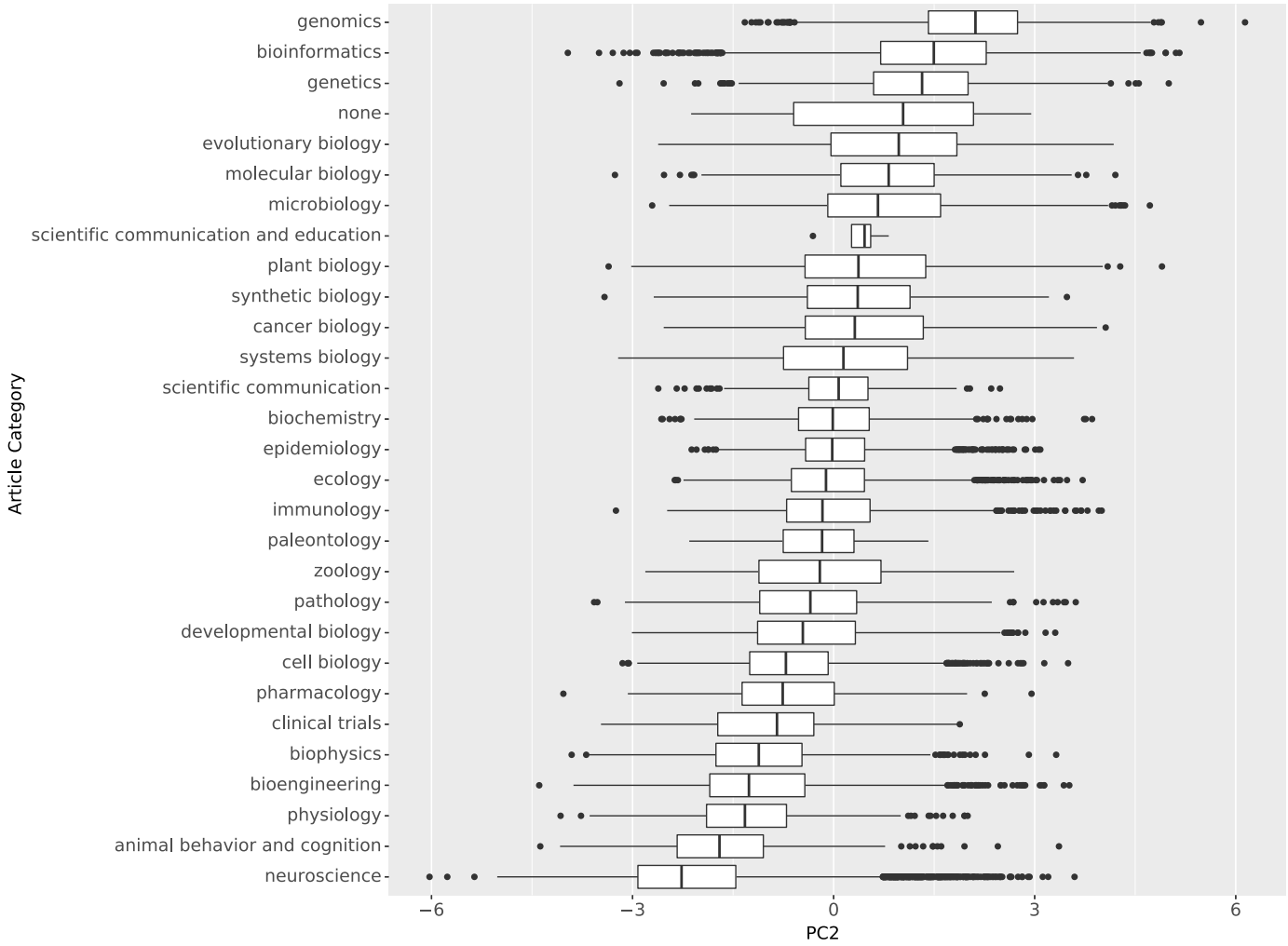

Figure 7: The second PC groups neuroscience related preprint categories and bioinformatics related preprint categories together. This is box plot shows preprints in each article category projected along the PC2 direction. Negative values indicate neuroscience concepts, while positive values indicate bioinformatic concepts.

Table 2: Top five 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	FigureLink
Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis [22]	4.700554908074704	None	https://www.b

Title [citation]	PC_1	License	FigureLink
			biorxiv.org/content/biorxiv/early/2017/10/02/197400/F1

Title [citation]	PC_1	License	FigureLink
			.large.jpg
Machine learning of stochastic gene network phenotypes [23]	4.410660604449826	CC-BY-NC-ND	https://www.biorxiv.org/content/biorxiv/ea

Title [citation]	PC_1	License	FigureLink
			rly/2019/10/31/825943/F5.large.jpg
Notions of similarity for computational biology models [24]	4.355295926618207	CC-BY-NC-ND	https://www.biorxi

Title [citation]	PC_1	License	FigureLink
			v.org/content/biorxiv/early/2016/03/21/044818/F1.larg

Title [citation]	PC_1	License	FigureLink
			e.jpg
GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation [25]	4.351517618262304	CC-BY-NC-ND	https://www.biorxiv.org/content/biorxiv/early/2

Title [citation]	PC_1	License	FigureLink
			019/09/18/769299/F1.large.jpg
SBpipe: a collection of pipelines for automating repetitive simulation and analysis tasks [26]	4.321847854182741	CC-BY-NC-ND	https://www.biorxiv.org

Title [citation]	PC_1	License	FigureLink
			/content/biorxiv/early/2017/02/09/107250/F1.large.jpg

Title [citation]	PC_1	License	FigureLink
Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection [27]	-4.263964235099807	CC-BY-ND	https://www.biorxiv.org/content/biorxiv/early/2018/1

Title [citation]	PC_1	License	FigureLink
			1/07/45504 8/F1.large.jpg
Systems analysis by mass cytometry identifies susceptibility of latent HIV-infected T cells to targeting of p38 and mTOR pathways [28]	-4.279016673409032	CC-BY-NC-ND	https://www.biorxiv.org/cont

Title [citation]	PC_1	License	FigureLink
			ent/biorxiv/early/2018/07/19/371922/F1.large.jpg
NADPH consumption by L-cystine reduction creates a metabolic vulnerability upon glucose deprivation [29]	-4.592209988884236	None	http

Title [citation]	PC_1	License	FigureLink
			s://www.biorxiv.org/content/biorxiv/early/2019/08/13/

Title [citation]	PC_1	License	FigureLink
			733162/F1.large.jpg
Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease [30]	-4.736613689905791	None	https://www.biorxiv.org/content/b

Title [citation]	PC_1	License	FigureLink
			iorxiv/early/2019/08/28/745943/F1.large.jpg
AKT but not MYC promotes reactive oxygen species-mediated cell death in oxidative culture [31]	-4.826793742506695	None	https://w

Title [citation]	PC_1	License	FigureLink
			www.biorxiv.org/content/biorxiv/early/2019/09/01/754577

Title [citation]	PC_1	License	FigureLink
			2/F1.large.jpg

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 {5)A,C). We also visualized documents projected on selected principal components (Figure {5)B). The first principal component separates bioRxiv preprints that encompass molecular biology results with preprints that contain quantitative biology results (Figure {5)C). 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 6). We also provide example preprints from the systems biology category to reinforce this concept (Table 2).

The second principal component represents the concept of neuroscience vs bioinformatics (Figure {5)A). 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 7 as enriched categories along this principal component are quite related to neuroscience (negative end) or bioinformatics (positive end). As with the first principal component 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 (greenelab.github.io/annorxiver-journal-recommender) and within our online repository (see Data Availability).

Identifying preprints that were not linked with their corresponding publications

Figure 8: The distances between preprints and their published version was on average lower than the distance between preprints and a randomly selected published article in the same journal. This violin plot shows the distribution of distances between both categories.

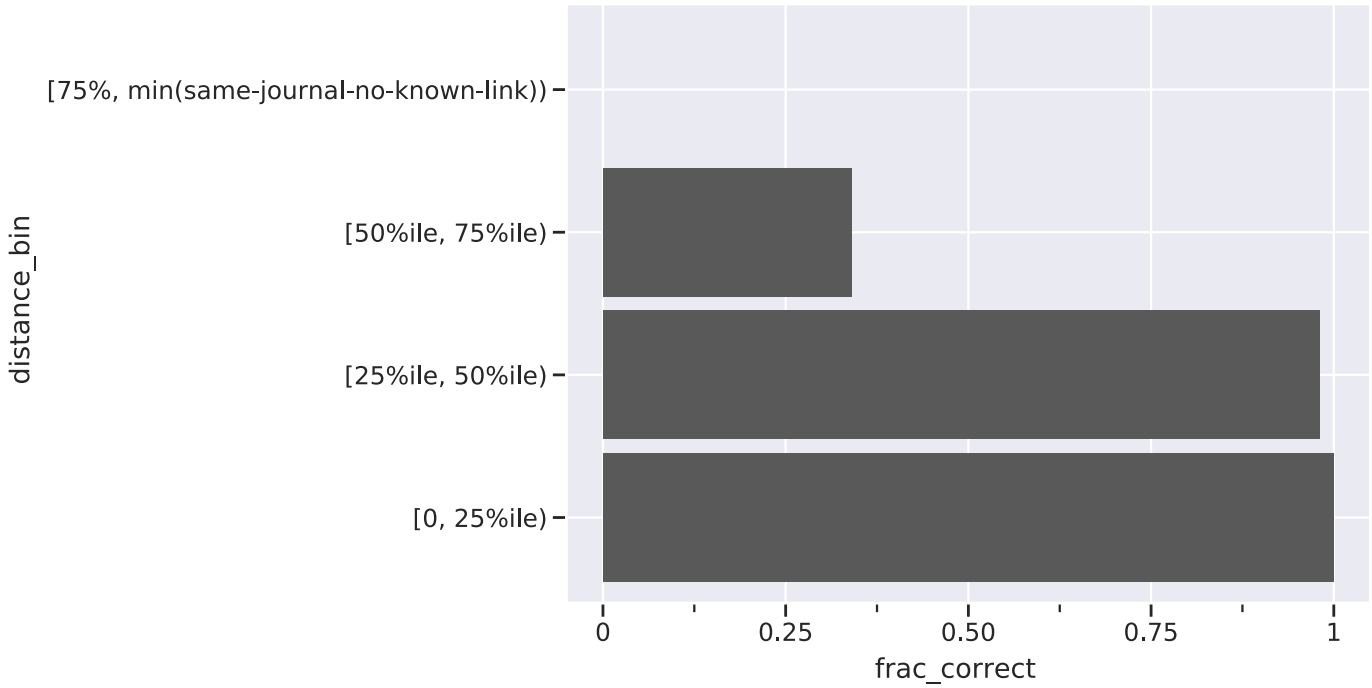

Figure 9: The proportion of publication-preprint pairs decreased as the distance for publication-preprint pairs increased. This bar chart depicts the fraction of true positives over the total number of pairs in each bin. Each bin contains a total of 200 annotated pairs and is based on the percentiles of the preprint-published distribution.

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. Missing preprint-publication links can make it more difficult to identify the latest version of scientific manuscripts and estimate the fraction

of articles that are eventually published [20]. 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 8). 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 in duplicate, and found a high inter-rater reliability for this task achieving a Cohen's Kappa [32] of 91.7%. 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 9). These two bins contained 1,720 preprint-article pairs, suggesting that many preprints have been published but not previously connected with their published versions.

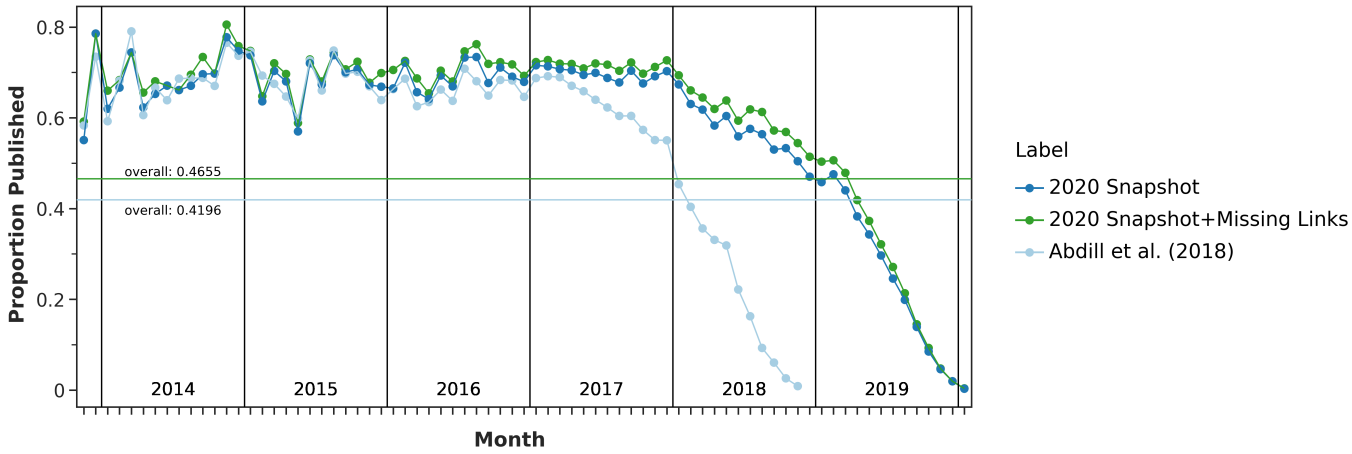

Figure 10: The overall fraction of published preprints is higher than originally estimated in [20]. This line plot shows the publication rate of preprints since bioRxiv first started. The x-axis represents months since bioRxiv started and the y-axis represents the proportion of preprints published. The light blue line represents the publication rate estimated by Abdill et al. [20]. 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.

We overlaid these new annotations onto existing annotations to reassess the overall preprint publication rate reported by Abdill et al. [20]. Our filtering criteria were intentionally stringent, so the increased estimate of publication rate amounts to a few percent (Figure 10). Many of these missed annotations were for preprints posted in the 2017-2018 interval. As opposed to those published in 2019 and later, these preprints are old enough that they are likely to have been published but it was interesting that the rate was not observed to be higher for older preprints.

Recommending Journals Based on Preprint Representation

Figure 11: 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.

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 uses the journals that published the nearest papers. Both classifiers outperformed a randomized baseline. A classifier that aimed to predict centroids performed better on the held out test set compared than the nearest paper classifier (Figure 11). There are 2516 journals in our dataset, so the baseline performance of a classifier is quite low. We were able to achieve a substantial increase with respect to random performance at predicting the journals that a paper was published in. However, the predictor is not perfect (Figure 11), which we should expect because there are multiple journals that cover certain topic areas and others have a very broad set of covered topic areas. Still, our software provides a starting point for authors to use the text of their preprints to identify potentially suitable publication venues.

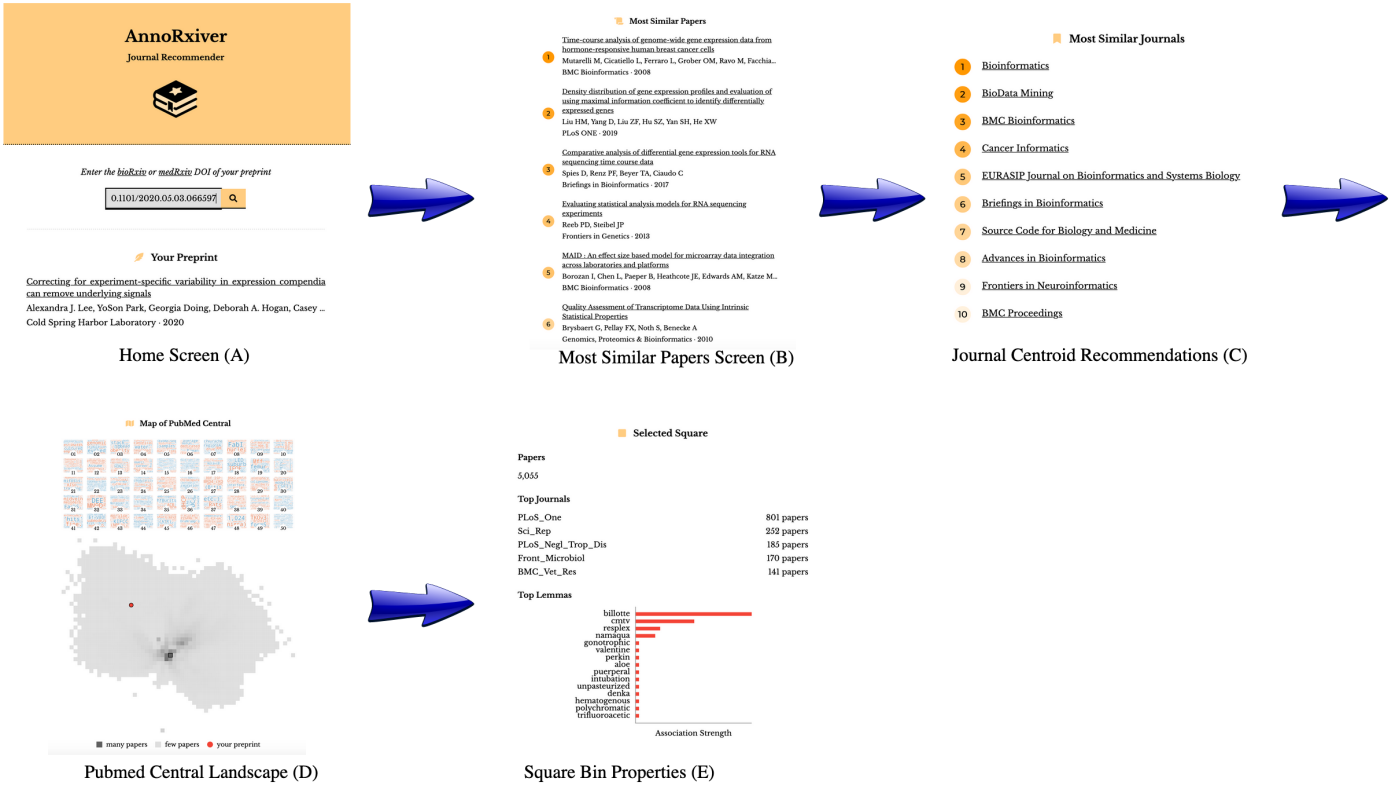

Figure 12: Here is the workflow of the journal recommender web-app. Starting with the homescreen users can paste in a *bioRxiv* or *medRxiv* doi, which sends a request to biorxiv or medrxiv (A). Next our app preprocesses the preprint and returns a listing of the top ten most similar papers (B) and the 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 12).

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 [20]. We noticed that a multitude of preprints highly associated with the first principal component have restrictive or no copyright license (Table {#tbl:five_pc1_table}). This finding highlights the ongoing problem of restricted access within the scientific community [33,34]. 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 preprints that had been published and not previously linked, which led us to find that the life sciences preprint publication rate is higher than previously estimated. Preprint-publication similarity also predicts the journals that will publish a manuscript. 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

The authors would like to thank Ariel Hippen Anderson for evaluating potential missing preprint to published version links. We also would like to thank Richard Sever and the *bioRxiv* team for their assistance with access to and support with questions about preprint full text downloaded from *bioRxiv*. This work was supported by Grant GBMF4552 from the Gordon Betty Moore Foundation and by NIH T32HG00046, Computational Genomics training grant, from the National Human Genome Reserach Institute (NHGRI).

References

1. **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

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

3. **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

4. **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

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

6. **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

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

8. **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

9. **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

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

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

12. **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

13. **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

14. **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)

15. **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

16. **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

17. **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

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

19. **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
bioRxiv (2020-03-05) https://doi.org/gg9353
DOI: 10.1101/2020.03.04.975177

20. **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

21. **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

22. **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

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

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

25. **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
bioRxiv (2019-09-18) https://doi.org/gg94bj
DOI: 10.1101/769299

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

27. **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

28. **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
bioRxiv (2018-07-19) https://doi.org/gg9398
DOI: 10.1101/371922

29. **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

30. **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

31. **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

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

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

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

35. **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

36. **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

37. **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
bioRxiv (2018-08-16) https://doi.org/gg94bb
DOI: 10.1101/392779

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

39. **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

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

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

42. **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
bioRxiv (2018-04-13) https://doi.org/gg9395
DOI: 10.1101/301051

43. **Electrical propagation of vasodilatory signals in capillary networks**
Pilhwa Lee
bioRxiv (2019-11-13) https://doi.org/gg94bn
DOI: 10.1101/840280

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

Supplemental Figures

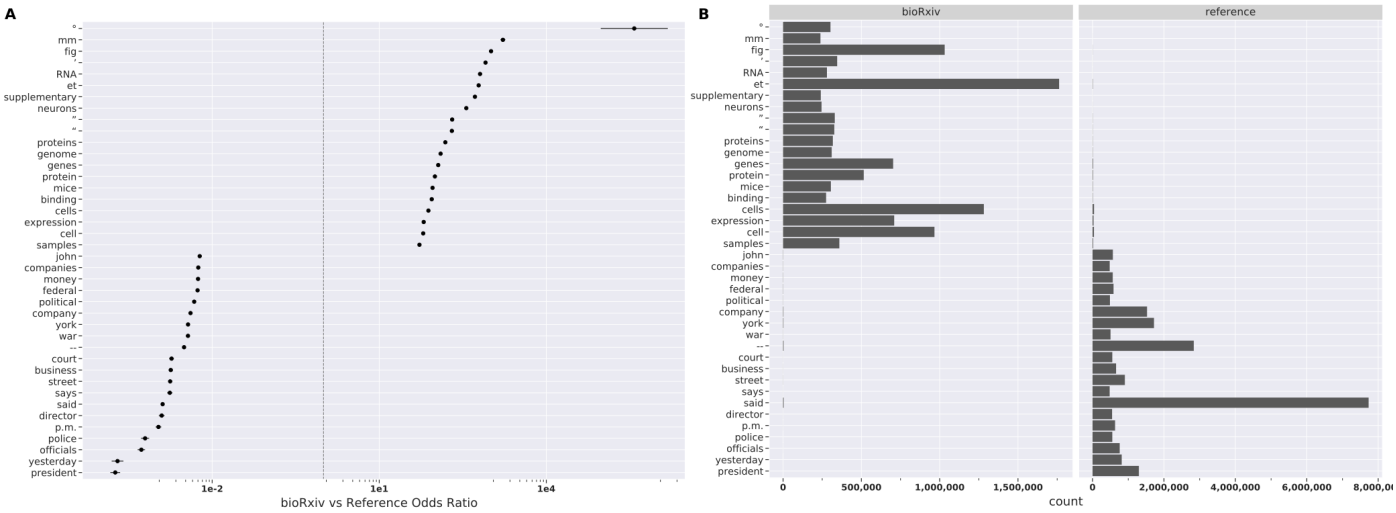

Figure 13: Topic associated tokens are highly enriched when comparing bioRxiv to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to bioRxiv. Values greater than one indicate a high association with bioRxiv whereas values less than one indicate high association 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.

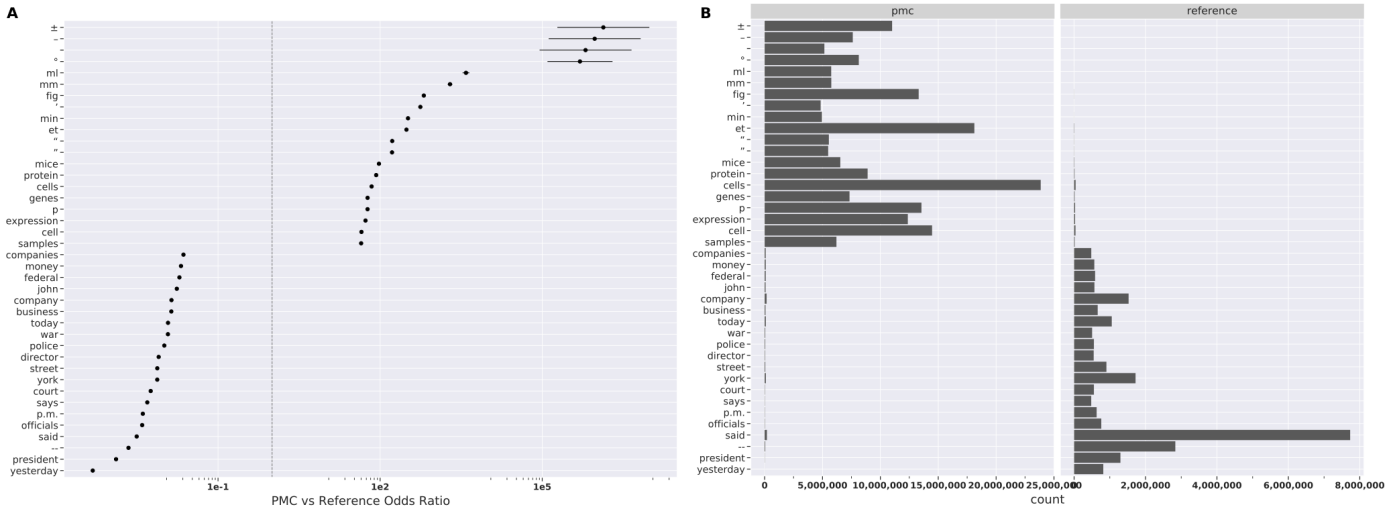

Figure 14: Typesetting symbols and biologically relevant tokens are highly enriched when comparing PubMed Central (PMC) to the New York Times. The plot on the left (A) is a point range plot of the odds ratio with respect to PMC. Values greater than one indicate a high association with PMC whereas values 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.

Supplemental Tables

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

Title [citation]	PC_2	License	FigureLink
Pangenome Analysis of Enterobacteria Reveals Richness of Secondary Metabolite Gene Clusters and their Associated Gene Sets [35]	3.5865702659438883	CC-BY-ND	https://www.biorxiv.org/

Title [citation]	PC_2	License	FigureLink
			content/biorxiv/early/2019/09/25/781328/F1.large.jpg

Title [citation]	PC_2	License	FigureLink
QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci [36]	3.470388383023157	None	https://www.biorxiv.org/content/biorxiv/early/2019/04

Title [citation]	PC_2	License	FigureLink
			/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 [37]	3.381490633407 3953	CC-BY-NC-ND	https://www.biorxiv.org/conte

Title [citation]	PC_2	License	FigureLink
			nt/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 [38]	3.3632576028389742	None	https

Title [citation]	PC_2	License	FigureLink
			:/www.biorxiv.org/content/biorxiv/early/2017/11/16/2

Title [citation]	PC_2	License	FigureLink
			20152/F3.large.jpg
The y-ome defines the thirty-four percent of Escherichia coli genes that lack experimental evidence of function [39]	3.2874278664 1417	CC-BY	https://www.biorxiv.org/content/bi

Title [citation]	PC_2	License	FigureLink
			orxiv/early/2018/12/03/328591/F1.large.jpg
The effects of time-varying temperature on delays in genetic networks [40]	-2.7047102478958056	None	https://w

Title [citation]	PC_2	License	FigureLink
			w.biorxiv.org/content/biorxiv/early/2015/09/24/019687

Title [citation]	PC_2	License	FigureLink
			/F1.large.jpg
An analog to digital converter creates nuclear localization pulses in yeast calcium signaling [41]	-2.7757450002 60895	None	https://www.biorxiv.org/content/biorxiv

Title [citation]	PC_2	License	FigureLink
			/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 [42]	-3.0473423827 98414	None	https://www.bio

Title [citation]	PC_2	License	FigureLink
			rxiv.org/content/biorxiv/early/2018/04/13/301051/F1.l

Title [citation]	PC_2	License	FigureLink
			arge.jpg
Electrical propagation of vasodilatory signals in capillary networks [43]	-3.1077155787 93087	CC-BY-NC-ND	https://www.biorxiv.org/content/biorxiv/earl

Title [citation]	PC_2	License	FigureLink
			y/2019/11/13/840280/F1.large.jpg
Dendritic spine geometry and spine apparatus organization govern the spatiotemporal dynamics of calcium [44]	-3.2153349907 2831	CC-BY-NC-ND	https://www.biorxiv.

Title [citation]	PC_2	License	FigureLink
			org/content/biorxiv/early/2019/05/29/3863677/F1.large.

Title [citation]	PC_2	License	FigureLink
			jpg