2022

Unmasking The Language Of Science Through Textual Analyses On Biomedical Preprints And Published Papers

David Nicholson
University of Pennsylvania

Unmasking The Language Of Science Through Textual Analyses On Biomedical Preprints And Published Papers

Abstract

Scientific communication is essential for science as it enables the field to grow. This task is often accomplished through a written form such as preprints and published papers. We can obtain a high-level understanding of science and how scientific trends adapt over time by analyzing these resources. This thesis focuses on conducting multiple analyses using biomedical preprints and published papers. In Chapter 2, we explore the language contained within preprints and examine how this language changes due to the peer-review process. We find that token differences between published papers and preprints are stylistically based, suggesting that peer-review results in modest textual changes. We also discovered that preprints are eventually published and adopted quickly within the life science community. Chapter 3 investigates how biomedical terms and tokens change their meaning and usage through time. We show that multiple machine learning models can correct for the latent variation contained within the biomedical text. Also, we provide the scientific community with a listing of over 43,000 potential change points. Tokens with notable changepoints such as “sars” and “cas9” appear within our listing, providing some validation for our approach. In Chapter 4, we use the weak supervision paradigm to examine the possibility of speeding up the labeling function generation process for multiple biomedical relationship types. We found that the language used to describe a biomedical relationship is often distinct, leading to a modest performance in terms of transferability. An exception to this trend is Compound-binds-Gene and Gene-interacts-Gene relationship types.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Casey S. Greene

Keywords

Natual Language Processing, Preprints, Pubmed Central, Semantic Change, Text Mining, Web Resources

Subject Categories

Bioinformatics | Computer Sciences | Linguistics

UNMASKING THE LANGUAGE OF SCIENCE THROUGH TEXTUAL ANALYSES ON BIOMEDICAL PREPRINTS AND PUBLISHED PAPERS

David N. Nicholson

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

_____

Casey S. Greene
Professor of Biochemistry and Molecular Genetics

Graduate Group Chairperson

_____

Benjamin F. Voight
Associate Professor of Pharmacology

Dissertation Committee

John H. Holmes, FACE, FACMI, Professor of Medical Informatics in Epidemiology
Benjamin F. Voight, Associate Professor of Pharmacology
Lawrence Hunter, Professor and Director of Computational Pharmacology
Graciela Gonzalez-Hernandez, Vice Chair of Education & Research in the Computational Biomedicine Department

UNMASKING THE LANGUAGE OF SCIENCE THROUGH TEXTUAL ANALYSES ON BIOMEDICAL PREPRINTS AND PUBLISHED PAPERS

To my parents, Jane and Lenwood Nicholson, and my family

ACKNOWLEDGMENT

I would like to thank Casey Greene for being a thoughtful and inspirational mentor. Most of this work was outside his field of expertise; however, his advice and perseverance were greatly appreciated. I would also like to thank my thesis committee: John Holmes, Lawrence Hunter, Graciela Gonzalez-Hernandez, and Benjamin Voight for their insight and advice. Special thanks to the past GCB staff: Li-San Wang, Maureen Kirsch and the present GCB staff: Benjamin Voight and Anna-Cara Apple. Without their assistance navigating graduate school would have been more than a logistics nightmare. I extend my gratitude to Maya Bucan and Junhyong Kim for allowing me to be a part of their T32 training grant.

It has been an honor to be a member of the Greenelab as the environment allowed me to develop as a researcher and discover new ways to develop my knowledge in data science and computational biology. Special thanks to my secondary mentor Daniel Himmelstein. His assistance was invaluable for my initial project in the lab. I can't thank Alexandria Lee enough as she has been my unofficial third mentor and work partner in crime. I'm so grateful for our sessions of venting and idea bouncing as it made the last part of this journey so much manageable.

I would also like to thank other current and past members of the GreeneLab: YoSon Park, Qiwen Hu, Michael Zietz, Jaclyn Taroni, Gregory Way, Amy Campbell, Ben Heil, Jake Crawford, Ariel Hippen, Halie Rando, Milton Pividori, Natalie Davidson, Taylor Reiter. Also, thank you software developers: Dongbo Hu, Vincent Rubinetti, and Faisal Alquaddoomi for their hard work on the frontend and backend portions of the web resources mentioned in this thesis.

Thank you to my collaborator Marvin Thelik for his help on analyzing bioRxiv preprint half-life and graciously providing me job hunting advice. Thanks to Alex Ratner and Steven Bach for their and navigating Snorkel for my text mining project. Also, I would to thank people in the graduate group: Sammy Klasfied, Van Truong and everyone in my cohort for making gradate experience bearable.

Regarding my journey before this Ph.D. program, I want to give special thanks to everyone in the Voight Lab their guidance and expertise helped fuel my drive to get this degree. Furthermore, thanks to Arnaldo and Aliza for running Penn PREP and Penn's SUIP program. I'd like to acknowledge the everyone in the Meyerhoff Scholars program. Without their help and guidance none of this work and effort would have been possible.

Lastly, I'd like to thank my family for their love and support. Especially my parents who sacrificed a great deal to get me here. I would like to thank my college buddies for checking in on my even though I got quiet for being stuck in graduate school for so long. I also would like to thank two special friends I made during this: Antwuan Turner, Reginald Terry. These two were great support systems outside the lab where at times I wanted to quit, and they made sure I didn't.

ABSTRACT

UNMASKING THE LANGUAGE OF SCIENCE THROUGH TEXTUAL ANALYSIS ON BIOMEDICAL PREPRINTS AND PUBLISHED PAPERS.

David N. Nicholson

Casey S. Greene

Scientific communication is essential for science as it enables the field to grow. This task is often accomplished through a written form such as preprints and published papers. We can obtain a high-level understanding of science and how scientific trends adapt over time by analyzing these resources. This thesis focuses on conducting multiple analyses using biomedical preprints and published papers. In Chapter 2, we explore the language contained within preprints and examine how this language changes due to the peer-review process. We find that token differences between published papers and preprints are stylistically based, suggesting that peer-review results in modest textual changes. We also discovered that preprints are eventually published and adopted quickly within the life science community. Chapter 3 investigates how biomedical terms and tokens change their meaning and usage through time. We show that multiple machine learning models can correct for the latent variation contained within the biomedical text. Also, we provide the scientific community with a listing of over 43,000 potential change points. Tokens with notable changepoints such as “sars” and “cas9” appear within our listing, providing some validation for our approach. In Chapter 4, we use the weak supervision paradigm to examine the possibility of speeding up the labeling function generation process for multiple biomedical relationship types. We found that the language used to describe a biomedical relationship is often distinct, leading to a modest performance in terms of transferability. An exception to this trend is Compound-binds-Gene and Gene-interacts-Gene relationship types.

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Textual analysis has been an indispensable field of research within the scientific community. It enables fast comprehension of research from a high-level perspective. Standard tasks in this field involve mining text to find relationships between entities, examining topics or themes within the text, and observing how the text changes given external factors such as the peer-review process. Regarding the life science community, these tasks are typically performed using published literature or social media platforms such as Twitter; however, other forms of text such as preprints can be used. This chapter discusses previous research that used textual analysis to analyze biomedical literature.

Analyzing scientific articles before the publication process

Preprints have become an essential medium in the life science field. They are defined as scholarly articles that have yet to undergo the peer-review process [1,2]. They are commonly hosted within repositories such as arXiv [3], bioRxiv [4], and medRxiv [5]. One of the primary motivations for preprints is communicating science without the long wait times or bias presented by scientific journals [6,7]. For example, these tools have been used to rapidly communicate disease outbreaks [8,9]. In addition to rapid communication, preprints are beginning to emerge as a data resource for textual analysis in the life science community. This section describes past efforts that used preprints for textual analysis.

Most of the analyses involving preprints are heavily concentrated on gauging scientific publicity. Preprints are being posted onto repositories at an exponential rate [10,11]. During their initial posting, preprints receive considerable attention regarding discussion on social media [11]. Also, preprints are being integrated into published literature as they are frequently downloaded and cited [11,12,13]. Overall, these studies highlight that preprints are being adapted into the life science community at a high rate.

Despite the rapid adaptation of preprints, they still face scrutiny from the life science community [14]. The main arguments against preprints are that they take a long time to publish, allow for the

possibility of being scooped, and aren't peer-reviewed [14,15,15,16,17,18,19]. This lack of peer-review can lead to submissions containing inconsistent results or conclusions [8,9] This trend was one of the driving factors for efforts to examine textual differences between preprints and their corresponding published versions [20,21]. Interestingly, these studies found that most differences between preprints and their corresponding published versions were small stylistic changes [20,21]. However, these studies only had limited data to analyze the differences. Despite these discoveries, there has yet to be a study that examined the language contained within preprints from a global perspective. This thesis fills this gap by analyzing preprints hosted on the bioRxiv repository and observing the differences between these preprints and their published counterparts.

Analyzing language adaptations through time

The meaning of words evolves and changes over time. For example, the word "nice" used to mean "foolish or innocent" back in the 15th-17th century; then, it underwent a positive shift toward meaning "pleasant or delightful" [22]. These changes are termed semantic shifts and can occur for various reasons [22]. Analyzing these shifts uncovers the historical context behind words and their meaning. Regarding science, examining these shifts enables swift comprehension of past endeavors and illuminates where research fields are progressing towards. Despite the usefulness of these shifts, there is a modest amount of effort in identifying these shifts in the life science community. This section discusses previous efforts that analyzed semantic shifts both outside of the life science field and within.

The task of examining semantic shifts has been quite successful outside the realm of life science. These studies utilize text resources such as the Google N-Gram corpus [23], New York Times (NYT) dataset [24], or the COHA corpus [25] to perform various tasks. One task used the N-Gram corpus to discover statistical trends behind semantic shifts [26,27], while others both N-Gram and COHA to validate techniques for detecting semantic shifts [28,29,30,31,32]. Other efforts used

NYT to observe how political viewpoints changed over time [33] along with validating techniques to detect semantic changes [34,35].

Regarding the life science community, the majority of studies have been heavily focused on a particular concept or topic. One study analyzed Reddit posts the gauge the audience's viewpoint and usage of the drug fentanyl [36]. Similarly, researchers analyzed Twitter posts to measure the viewpoint of the platform's users on the COVID-19 pandemic [37]. Outside of social media, one study examined how titles and abstracts that mentioned a disease changed through time [38]. Despite these efforts, there has yet to be work done that universally examines semantic shifts for all biomedical terms and concepts. This thesis fills this gap by detecting semantic shifts within the biomedical literature.

Text Mining for Relationship Extraction

This section was adapted from: Nicholson, David N., and Greene, Casey, S. "Constructing knowledge graphs and their biomedical applications" Published in Computational and Structural Biotechnology Journal https://doi.org/10.1016/j.csbj.2020.05.017

Mining text to extract relationships has been prevalent within the textual analysis field. This task provides a medium to identify known discoveries and populate database resources rapidly. There are many ways to perform text mining, and this section discusses the pros and cons of each approach.

Rule-Based Relationship Extraction

Rule-based extraction consists of identifying essential keywords and grammatical patterns to detect relationships of interest. Keywords are established via expert knowledge or through the use of pre-existing ontologies, while grammatical patterns are constructed via experts curating parse trees. Parse trees are tree data structures that depict a sentence's grammatical structure and come in two forms: a constituency parse tree (Figure 1) and a dependency parse tree (Figure 2). Both trees use part of speech tags, labels that dictate the grammatical role of a word such as

noun, verb, adjective, etc., for construction, but represent the information in two different forms. Constituency parse trees break a sentence into subphrases (Figure 1) while dependency path trees analyze the grammatical structure of a sentence (Figure 2). Many text mining approaches [39,40,41] use such trees to generate features for machine learning algorithms and these approaches are discussed in later sections. In this section, we focus on approaches that use rule-based extraction as a primary strategy to detect sentences that allude to a relationship. Grammatical patterns can simplify sentences for easy extraction [42,43]. Jonnalagadda et al. used a set of grammar rules inspired by constituency trees to reshape complex sentences with simpler versions [42] and these simplified versions were manually curated to determine the presence of a relationship. By simplifying sentences, this approach achieved high recall but had low precision [42]. Other approaches used simplification techniques to make extraction easier [44,45,46,47]. Tudor et al. simplified sentences to detect protein phosphorylation events [46]. Their sentence simplifier broke complex sentences that contain multiple protein events into smaller sentences that contain only one distinct event. By breaking these sentences down the authors were able to increase their recall; however, sentences that contained ambiguous directionality or multiple phosphorylation events were too complex for the simplifier. As a consequence, the simplifier missed some relevant sentences [46]. These errors highlight a crucial need for future algorithms to be generalizable enough to handle various forms of complex sentences.

Pattern matching is a fundamental approach used to detect relationship asserting sentences. These patterns can consist of phrases from constituency trees, a set of keywords or some combination of both [48,49,50,51,52,53]. Xu et al. designed a pattern matcher system to detect sentences in PubMed abstracts that indicate drug-disease treatments [52]. This system matched drug-disease pairs from ClinicalTrials.gov to drug-disease pairs mentioned in abstracts. This matching process aided the authors in identifying sentences that can be used to create simple patterns, such as “Drug in the treatment of Disease” [52], to match other sentences in a wide

variety of abstracts. The authors hand curated two datasets for evaluation and achieved a high precision score of 0.904 and a low recall score of 0.131 [52]. This low recall score was based on constructed patterns being too specific to detect infrequent drug pairs. Besides constituency trees, some approaches used dependency trees to construct patterns [39,54]. Depending upon the nature of the algorithm and text, dependency trees could be more appropriate than constituency trees and vice versa. The performance difference between the two trees remains as an open question for future exploration.

Rule-based methods provide a basis for many relationship extraction systems. Approaches in this category range from simplifying sentences for easy extraction to identifying sentences based on matched key phrases or grammatical patterns. Both require a significant amount of manual effort and expert knowledge to perform well. A future direction is to develop ways to automate the construction of these hand-crafted patterns, which would accelerate the process of creating these rule-based systems.

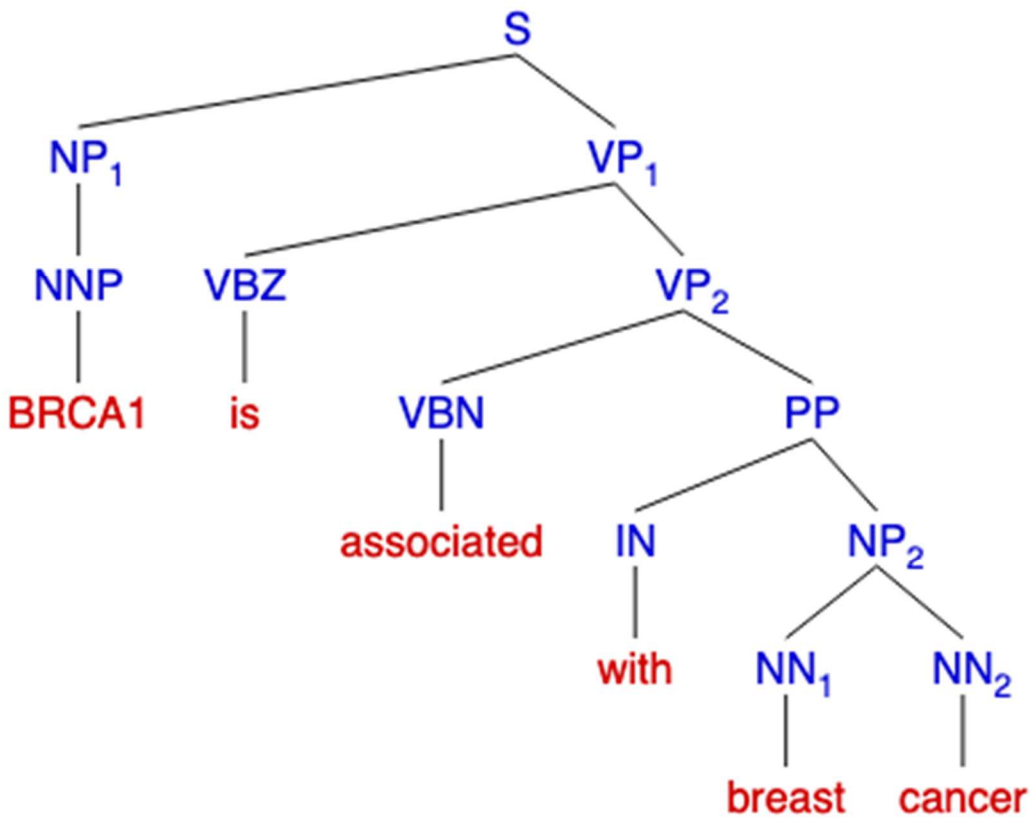


Figure 1 Constituency Parse Tree for "BRCA1 is associated with breast cancer"

A visualization of a constituency parse tree using the following sentence: "BRCA1 is associated with breast cancer" [55]. This type of tree has the root start at the beginning of the sentence. Each word is grouped into subphrases depending on its correlating part of speech tag. For example, the word "associated" is a past participle verb (VBN) that belongs to the verb phrase (VP) subgroup.

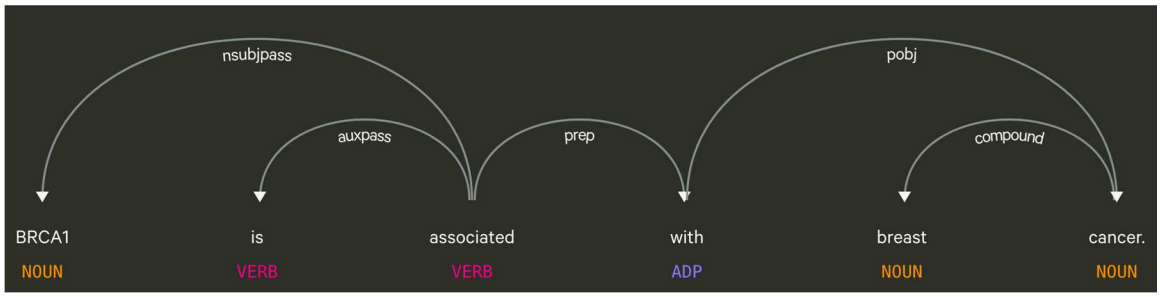


Figure 2 Dependency Parse tree for the sentence "BRCA1 is associated with breast cancer"

A visualization of a dependency parse tree using the following sentence: "BRCA1 is associated with breast cancer" [56]. For these types of trees, the root begins with the main verb of the

sentence. Each arrow represents the dependency shared between two words. For example, the dependency between BRCA1 and associated is nsubjpass, which stands for passive nominal subject. This means that "BRCA1" is the subject of the sentence and it is being referred to by the word "associated".

Extracting Relationships Without Labels

Unsupervised extractors draw inferences from textual data without the use of annotated labels. These methods involve some form of clustering or statistical calculations. In this section we focus on methods that use unsupervised learning to extract relationships from text.

An unsupervised extractor can exploit the fact that two entities may appear together in text. This event is referred to as co-occurrence and studies that use this phenomenon can be found in Table 1. Two databases DISEASES [57] and STRING [58] were populated using a co-occurrence scoring method on PubMed abstracts, which measured the frequency of co-mention pairs within individual sentences as well as the abstracts themselves. This technique assumes that each individual co-occurring pair is independent from one another. Under this assumption mention pairs that occur more than expected were presumed to implicate the presence of an association or interaction. This approach identified 543,405 disease gene associations [57] and 792,730 high confidence protein-protein interactions [58] but is limited to only PubMed abstracts.

Full text articles are able to dramatically enhance relationship detection [59,60]. Westergaard et al. used a co-occurrence approach, similar to DISEASES [57] and STRING [58], to mine full articles for protein-protein interactions and other protein related information [59]. The authors discovered that full text provided better prediction power than using abstracts alone, which suggests that future text mining approaches should consider using full text to increase detection power.

Unsupervised extractors often treat different biomedical relationships as multiple isolated problems. An alternative to this perspective is to capture all different types at once. Clustering is an approach that performs simultaneous extraction. Percha et al. used a biclustering algorithm on generated dependency parse trees to group sentences within PubMed abstracts [61]. Each

cluster was manually curated to determine which relationship each group represented. This approach captured 4,451,661 dependency paths for 36 different groups [61]. Despite the success, this approach suffered from technical issues such as dependency tree parsing errors. These errors resulted in some sentences not being captured by the clustering algorithm [61]. Future clustering approaches should consider simplifying sentences to prevent this type of issue. Overall unsupervised methods provide a means to rapidly extract relationship asserting sentences without the need of annotated text. Approaches in this category range from calculating co-occurrence scores to clustering sentences and provide a generalizable framework that can be used on large repositories of text. Full text has already been shown to meaningfully improve the performance of methods that aim to infer relationships using cooccurrences [59], and we should expect similar benefits for machine learning approaches. Furthermore, we expect that simplifying sentences would improve unsupervised methods and should be considered as an initial preprocessing step.

Table 1 Approaches that mainly use a form of co-occurrence.

Study	Relationship of Interest
CoCoScore [62]	Protein-Protein Interactions, Disease-Gene and Tissue-Gene Associations
Rastegar-Mojarad et al. [63]	Drug Disease Treatments
CoPub Discovery [64]	Drug, Gene and Disease interactions
Westergaard et al. [59]	Protein-Protein Interactions
DISEASES [57]	Disease-Gene associations
STRING [65]	Protein-Protein Interactions
Singhal et al. [66]	Genotype-Phenotype Relationships

Supervised Relationship Extraction

Supervised extractors use labeled sentences to construct generalized patterns that bisect positive examples (sentences that allude to a relationship) from negative ones (sentences that do not allude to a relationship). Most of these approaches have flourished due to pre-labelled publicly available datasets (Table 2). These datasets were constructed by curators for shared open tasks [67,68] or as a means to provide the scientific community with a gold standard [68,69,70]. Approaches that use these available datasets range from using linear classifiers such as support vector machines (SVMs) to non-linear classifiers such as deep learning techniques. The rest of this section discusses approaches that use supervised extractors to detect relationship asserting sentences.

Some supervised extractors involve the mapping of textual input into a high dimensional space. SVMs are a type of classifier that can accomplish this task with a mapping function called a kernel [41,71]. These kernels take information such as a sentence's dependency tree [39,40], part of speech tags [41] or even word counts [71] and map them onto a dense feature space. Within this space, these methods construct a hyperplane that separates sentences in the positive class (illustrates a relationship) from the negative class (does not illustrate a relationship). Kernels can be manually constructed or selected to cater to the relationship of interest [40,41,71,71]. Determining the correct kernel is a nontrivial task that requires expert knowledge to be successful. In addition to single kernel methods, a recent study used an ensemble of SVMs to extract disease-gene associations [72]. This ensemble outperformed notable disease-gene association extractors [54,73] in terms of precision, recall and F1 score. Overall, SVMs have been shown to be beneficial in terms of relationship mining; however, major focus has shifted to utilizing deep learning techniques which can perform non-linear mappings of high dimensional data.

Deep learning is an increasingly popular class of techniques that can construct their own features within a high dimensional space [74,75]. These methods use different forms of neural networks, such as recurrent or convolutional neural networks, to perform classification.

Recurrent neural networks (RNN) are designed for sequential analysis and use a repeatedly updating hidden state to make predictions. An example of a recurrent neural network is a long short-term memory (LSTM) network [76]. Cocos et al. [77] used a LSTM to extract drug side effects from de-identified twitter posts, while Yadav et al. [78] used an LSTM to extract protein-protein interactions. Others have also embraced LSTMs to perform relationship extraction [77,79,80,81,82]. Despite the success of these networks, training can be difficult as these networks are highly susceptible to vanishing and exploding gradients [83,84]. One proposed solution to this problem is to clip the gradients while the neural network trains [85]. Besides the gradient problem, these approaches only peak in performance when the datasets reach at least tens of thousands of data points [86].

Convolutional neural networks (CNNs), which are widely applied for image analysis, use multiple kernel filters to capture small subsets of an overall image [75]. In the context of text mining an image is replaced with words within a sentence mapped to dense vectors (i.e., word embeddings) [87,88]. Peng et al. used a CNN to extract sentences that mentioned protein-protein interactions [89] and Zhou et al. used a CNN to extract chemical-disease relations [90]. Others have used CNNs and variants of CNNs to extract relationships from text [91,92,93]. Just like RNNs, these networks perform well when millions of labeled examples are present [86]; however, obtaining these large datasets is a non-trivial task. Future approaches that use CNNs or RNNs should consider solutions to obtaining these large quantities of data through means such as weak supervision [94], semi-supervised learning [95] or using pre-trained networks via transfer learning [96,97].

Semi-supervised learning [95] and weak supervision [94] are techniques that can rapidly construct large datasets for machine learning classifiers. Semi-supervised learning trains

classifiers by combining labeled data with unlabeled data. For example, one study used a variational auto encoder with a LSTM network to extract protein-protein interactions from PubMed abstracts and full text [98]. This is an elegant solution for the small dataset problem but requires labeled data to start. This dependency makes finding under-studied relationships difficult as one would need to find or construct examples of the missing relationships at the start.

Weak or distant supervision takes a different approach by using noisy or even erroneous labels to train classifiers [94,99,100,101]. Under this paradigm, sentences are labeled based on their mention pair being present (positive) or absent (negative) in a database and, once labeled, a machine learning classifier can be trained to extract relationships from text [94]. For example, Thomas et al. [102] used distant supervision to train a SVM to extract sentences mentioning protein-protein interactions (PPI). Their SVM model achieved comparable performance against a baseline model; however, the noise generated via distant supervision was difficult to eradicate [102]. A number of efforts have focused on combining distant supervision with other types of labeling strategies to mitigate the negative impacts of noisy knowledge bases [103,104,105]. Combining distant supervision with other types of labeling strategies remains an active area of investigation with numerous associated challenges and opportunities. This thesis investigates one strategy that involved reusing multiple labeling sources to speed up the efforts of labeling sentences under the weak supervision paradigm.

Table 2 A set of publicly available datasets for supervised text mining.

Dataset	Type of Sentences
AIMed [106]	Protein-Protein Interactions
BioInfer [107]	Protein-Protein Interactions
LLL [108]	Protein-Protein Interactions
IEPA [109]	Protein-Protein Interactions
HPRD5 [69]	Protein-Protein Interactions

EU-ADR [110]	Disease-Gene Associations
BeFree [73]	Disease-Gene Associations
CoMAGC [70]	Disease-Gene Associations
CRAFT [111]	Disease-Gene Associations
Biocreative V CDR [68]	Compound induces Disease
Biocreative IV ChemProt [67]	Compound-Gene Bindings

CHAPTER 2

Examining linguistic shifts between preprints and publications

This chapter was originally published as: Nicholson DN, Rubinetti V, Hu D, Thielk M, Hunter LE, Greene CS (2022) Examining linguistic shifts between preprints and publications. PLoS Biol 20(2): e3001470. https://doi.org/10.1371/journal.pbio.3001470

This is a co-authored paper where the majority of scientific work was performed by Nicholson DN who was advised by Greene CS and Hunter LE. Thielk M. contributed to the half-life analysis for preprint categories. Rubinetti V. and Hu D. assisted with the creation of the preprint similarity search website.

Introduction

The dissemination of research findings is key to science. Initially, much of this communication happened orally [6]. During the 17th century, the predominant form of communication shifted to personal letters shared from one scientist to another [6]. Scientific journals didn't become a predominant mode of communication until the 19th and 20th centuries when the first journal was created [6,112,113]. Although scientific journals became the primary method of communication, they added high maintenance costs and long publication times to scientific discourse [112,113]. Some scientists' solutions to these issues have been to communicate through preprints, which are scholarly works that have yet to undergo peer review process [1,2].

Preprints are commonly hosted on online repositories, where users have open and easy access to these works. Notable repositories include arXiv [3], bioRxiv [4] and medRxiv [114]; however, there are over 60 different repositories available [115]. The burgeoning uptake of preprints in life sciences has been examined through research focused on metadata from the bioRxiv repository. For example, life science preprints are being posted at an increasing rate [10]. Furthermore,

these preprints are being rapidly shared on social media, routinely downloaded, and cited [116]. Some preprint categories are shared on social media by both scientists and non-scientists [117]. About two-thirds to three-quarters of preprints are eventually published [13,118] and life science articles that have a corresponding preprint version are cited and discussed more often than articles without them [12,119,120]. Preprints take an average of 160 days to be published in the peer-reviewed literature [18], and those with multiple versions take longer to publish[18].

The rapid uptake of preprints in the life sciences also poses challenges. Preprint repositories receive a growing number of submissions [14]. Linking preprints with their published counterparts is vital to maintaining scholarly discourse consistency, but this task is challenging to perform manually [15,119,121]. Errors and omissions in linkage result in missing links and consequently erroneous metadata. Furthermore, repositories based on standard publishing tools are not designed to show how the textual content of preprints is altered due to the peer review process [14]. Certain scientists have expressed concern that competitors could scoop them by making results available before publication [14,19]. Preprint repositories by definition do not perform in-depth peer review, which can result in posted preprints containing inconsistent results or conclusions [15,16,17,120]; however, an analysis of preprints posted at the beginning of 2020 revealed that over 50% underwent minor changes in the abstract text as they were published, but over 70% did not change or only had simple rearrangements to panels and tables [122]. Despite a growing emphasis on using preprints to examine the publishing process within life sciences, how these findings relate to the text of all documents in bioRxiv has yet to be examined.

Textual analysis uses linguistic, statistical, and machine learning techniques to analyze and extract information from text [123,124]. For instance, scientists analyzed linguistic similarities and differences of biomedical corpora [125,126]. Scientists have provided the community with a number of tools that aide future text mining systems [127,128,129] as well as advice on how to train and test future text processing systems [130,131,132]. Here, we use textual analysis to

examine the bioRxiv repository, placing a particular emphasis on understanding the extent to which full-text research can address hypotheses derived from the study of metadata alone.

To understand how preprints relate to the traditional publishing ecosystem, we examine the linguistic similarities and differences between preprints and peer-reviewed text and observe how linguistic features change during the peer review and publishing process. We hypothesize that preprints and biomedical text will appear to have similar characteristics, especially when controlling for the differential uptake of preprints across fields. Furthermore, we hypothesize that document embeddings [87,133] provide a versatile way to disentangle linguistic features along with serving as a suitable medium for improving preprint repository functionality. We test this hypothesis by producing a linguistic landscape of bioRxiv preprints, detecting preprints that change substantially during publication, and identify journals that publish manuscripts that are linguistically similar to a target preprint. We encapsulate our findings through a web app that projects a user-selected preprint onto this landscape and suggests journals and articles that are linguistically similar. Our work reveals how linguistically similar and dissimilar preprints are to peer-reviewed text, quantifies linguistic changes that occur during the peer review process, and highlights the feasibility of document embeddings concerning preprint repository functionality and peer review's effect on publication time.

Materials and Methods

Corpora Examined

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

BioRxiv Corpus

BioRxiv [4] is a repository for life sciences preprints. We downloaded an XML snapshot of this repository on February 3rd, 2020, from bioRxiv's Amazon S3 bucket [134]. This snapshot contained the full text and image content of 98,023 preprints. Preprints on bioRxiv are versioned, and in our snapshot, 26,905 out of 98,023 contained more than one version. When preprints had multiple versions, we used the latest one unless otherwise noted. Authors submitting preprints to bioRxiv can select one of twenty-nine different categories and tag the type of article: a new result, confirmatory finding, or contradictory finding. A few preprints in this snapshot were later withdrawn from bioRxiv; when withdrawn, their content is replaced with the reason for withdrawal. We encountered a total of 72 withdrawn preprints within our snapshot. After removal, we were left with 97,951 preprints for our downstream analyses.

PubMed Central Open Access Corpus

PubMed Central (PMC) is a digital archive for the United States National Institute of Health's Library of Medicine (NIH/NLM) that contains full text biomedical and life science articles [135]. Paper availability within PMC is mainly dependent on the journal's participation level [136]. Articles appear in PMC as either accepted author manuscripts (Green Open Access) or via open access publishing at the journal (Gold Open Access [137]). Individual journals have the option to fully participate in submitting articles to PMC, selectively participate sending only a few papers to PMC, only submit papers according to NIH's public access policy [138], or not participate at all; however, individual articles published with the CC BY license may be incorporated. As of September 2019, PMC had 5,725,819 articles available [139]. Out of these 5 million articles, about 3 million were open access (PMCOA) and available for text processing systems [128,140]. PMC also contains a resource that holds author manuscripts that have already passed the peer review process [141]. Since these manuscripts have already been peer-reviewed, we excluded them from our analysis as the scope of our work is focused on examining the beginning and end of a preprint's life cycle. We downloaded a snapshot of the PMCOA corpus on January 31st,

2020. This snapshot contained many types of articles: literature reviews, book reviews, editorials, case reports, research articles, and more. We used only research articles, which align 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 [24] is a collection of newspaper articles from the New York Times dating from January 1st, 1987, to June 19th, 2007. This collection contains over 1.8 million articles where 1.5 million of those articles have undergone manual entity tagging by library scientists [24]. We downloaded this collection on August 3rd, 2020, from the Linguistic Data Consortium (see Software and Data Availability section) and used the entire collection as a negative control for our corpora comparison analysis.

Mapping bioRxiv preprints to their published counterparts

We used CrossRef [142] to identify bioRxiv preprints linked to a corresponding published article. We accessed CrossRef on July 7th, 2020, and successfully linked 23,271 preprints to their published counterparts. Out of those 23,271 preprint-published pairs, only 17,952 pairs had a published version present within the PMCOA corpus. For our analyses that involved published links, we only focused on this subset of preprints-published pairs.

Comparing Corpora

We compared the bioRxiv, PMCOA, and NYTAC corpora to assess the similarities and differences between them. We used the NYTAC corpus as a negative control to assess the similarity between two life sciences repositories compared with non-life sciences text. All corpora contain multiple words that do not have any meaning (e.g. conjunctions, prepositions, etc.) or occur with a high frequency. These words are termed stopwords and are often removed to improve text processing pipelines. Along with stopwords, all corpora contain both words and non-word entities (e.g., numbers or symbols like $\pm$), which we refer to together as tokens to avoid confusion. We calculated the following characteristic metrics for each corpus: the number of documents, the number of sentences, the total number of tokens, the number of stopwords, the

average length of a document, the average length of a sentence, the number of negations, the number of coordinating conjunctions, the number of pronouns and the number of past tense verbs. SpaCy is a lightweight and easy-to-use python package designed to preprocess and filter text [56]. We used spaCy's "en_core_web_sm" model [56] (version 2.2.3) to preprocess all corpora and filter out 326 stopwords using spaCy's default settings.

Following that cleaning process, we calculated the frequency of every token across all corpora. Because many tokens were unique to one set or the other and observed at low frequency, we focused on the union of the top 0.05% (~100) most frequently occurring tokens within each corpus. We generated a contingency table for each token in this union and calculated the odds ratio along with the 95% confidence interval [143]. We measured corpora similarity by calculating the Kullback–Leibler (KL) divergence across all corpora along with token enrichment analysis. KL divergence is a metric that measures the extent to which two distributions differ from each other. A low value of KL divergence implicates that two distributions are similar and vice versa for high values. The optimal number of tokens used to calculate the KL divergence is unknown, so we calculated this metric using a range of the 100 most frequently occurring tokens between two corpora to the 5000 most frequently occurring tokens.

Constructing a Document Representation for Life Sciences Text

We sought to build a language model to quantify linguistic similarities of biomedical preprints and articles. Word2vec is a suite of neural networks designed to model linguistic features of tokens based on their appearance in the text. These models are trained to either predict a token based on its sentence context, called a continuous bag of words (CBOW) model, or predict the context based on a given token, called a skipgram model [87]. Through these prediction tasks, both networks learn latent linguistic features which are helpful for downstream tasks, such as identifying similar tokens. We used gensim [144] (version 3.8.1) to train a CBOW [87] model over all the main text within each preprint in the bioRxiv corpus. Determining the best number of dimensions for token embeddings can be a non-trivial task; however, it has been shown that

optimal performance is between 100-1000 dimensions [145]. We chose to train the CBOW model using 300 hidden nodes, a batch size of 10000 tokens, and for 20 epochs. We set a fixed random seed and used gensim's default settings for all other hyperparameters. Once trained, every token present within the CBOW model is associated with a dense vector representing latent features captured by the network. We used these token vectors to generate a document representation for every article within the bioRxiv and PMCOA corpora. We used spaCy to lemmatize each token for each document and then took the average of every lemmatized token present within the CBOW model and the individual document [133]. Any token present within the document but not in the CBOW model is ignored during this calculation process.

Visualizing and Characterizing Preprint Representations

We sought to visualize the landscape of preprints and determine the extent to which their representation as document vectors corresponded to author-supplied document labels. We used principal component analysis (PCA) [146] to project bioRxiv document vectors into a low-dimensional space. We trained this model using scikit-learn's [147] implementation of a randomized solver [148] with a random seed of 100, an output of 50 principal components (PCs), and default settings for all other hyperparameters. After training the model, every preprint within the bioRxiv corpus receives a score for each generated PC. We sought to uncover concepts captured within generated PCs and used the cosine similarity metric to examine these concepts. This metric takes two vectors as input and outputs a score between -1 (most dissimilar) and 1 (most similar). We used this metric to score the similarity between all generated PCs and every token within our CBOW model for our use case. We report the top 100 positive and negative scoring tokens as word clouds. The size of each word corresponds to the magnitude of similarity, and color represents a positive (orange) or negative (blue) association.

Discovering Unannotated Preprint-Publication Relationships

The bioRxiv maintainers have automated procedures to link preprints to peer-reviewed versions, and many journals require authors to update preprints with a link to the published version. However, this automation is primarily based on the exact matching of specific preprint attributes. If authors change the title between a preprint and published version (e.g., [149] and [150]), then this change will prevent bioRxiv from automatically establishing a link. Furthermore, if the authors do not report the publication to bioRxiv, the preprint and its corresponding published version are treated as distinct entities despite representing the same underlying research. We hypothesize that close proximity in the document embedding space could match preprints with their corresponding published version. If this finding holds, we could use this embedding space to fill in links missed by existing automated processes. We used the subset of paper-preprint pairs annotated in CrossRef as described above to calculate the distribution of available preprint to published distances. We calculated this distribution by taking the Euclidean distance between the preprint's embedding coordinates and the coordinates of its corresponding published version. We also calculated a background distribution, which consisted of the distance between each preprint with an annotated publication and a randomly selected article from the same journal. We compared both distributions to determine if there was a difference between both groups as a significant difference would indicate that this embedding method can parse preprint-published pairs apart. After comparing the two distributions, we calculated distances between preprints without a published version link with PMCOA articles that weren't matched with a corresponding preprint. We filtered any potential links with distances greater than the minimum value of the background distribution as we considered these pairs to be true negatives. Lastly, we binned the remaining pairs based on percentiles from the annotated pairs distribution at the [0,25th percentile), [25th percentile, 50th percentile), [50th percentile, 75th percentile), and [75th percentile, minimum background distance). We randomly sampled 50 articles from each bin and shuffled these four sets to produce a list of 200 potential preprint-published pairs with a

randomized order. We supplied these pairs to two co-authors to manually determine if each link between a preprint and a putative matched version was correct or incorrect. After the curation process, we encountered eight disagreements between the reviewers. We supplied these pairs to a third scientist, who carefully reviewed each case and made a final decision. Using this curated set, we evaluated the extent to which distance in the embedding space revealed valid but unannotated links between preprints and their published versions.

Measuring Time Duration for Preprint Publication Process

Preprints can take varying amounts of time to be published. We sought to measure the time required for preprints to be published in the peer-reviewed literature and compared this time measurement across author-selected preprint categories as well as individual preprints. First, we queried bioRxiv's application programming interface (API) to obtain the date a preprint was posted onto bioRxiv as well as the date a preprint was accepted for publication. We did not include preprint matches found by our paper matching approach (see 'Discovering Unannotated Preprint-Publication Relationships'). We measured time elapsed as the difference between the date a preprint was first posted on bioRxiv and its publication date. Along with calculating the time elapsed, we also recorded the number of different preprint versions posted onto bioRxiv.

We used this captured data to apply the Kaplan-Meier estimator [151] via the KaplanMeierFitter function from the lifelines [152] (version 0.25.6) python package to calculate the half-life of preprints across all preprint categories within bioRxiv. We considered survival events as preprints that have yet to be published. We encountered 123 cases where the preprint posting date was subsequent to the publication date, resulting in a negative time difference, as previously reported [153]. We removed these preprints for this analysis as they were incompatible with the rules of the bioRxiv repository.

We measured the textual difference between preprints and their corresponding published version after our half-life calculation by calculating the Euclidean distance for their respective embedding representation. This metric can be difficult to understand within the context of textual differences,

so we sought to contextualize the meaning of a distance unit. We first randomly sampled with replacement a pair of preprints from the Bioinformatics topic area as this was well represented within bioRxiv and contains a diverse set of research articles. Next, we calculated the distance between two preprints 1000 times and reported the mean. We repeated the above procedure using every preprint within bioRxiv as a whole. These two means serve as normalized benchmarks to compare against as distance units are only meaningful when compared to other distances within the same space. Following our contextualization approach, we performed linear regression to model the relationship between preprint version count with a preprint's time to publication. We also performed linear regression to measure the relationship between document embedding distance and a preprint's time to publication. For this analysis, we retained preprints with negative time within our linear regression model, and we observed that these preprints had minimal impact on results. We visualize our version count regression model as a violin plot and our document embeddings regression model as a square bin plot.

Building Classifiers to Detect Linguistically Similar Journal Venues and Published Articles

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

Training models to identify which journal publishes similar articles is challenging as not all journals are the same. 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. Furthermore, some journals focus on publishing articles within a concentrated topic area, while others cover many dispersive topics. Therefore, we designed two approaches to account for these characteristics. Our first approach focuses on articles that account for a journal's variation of publication topics.

This approach allows for topically similar papers to be retrieved independently of their respective journal. Our second approach is centered on journals to account for varying publication rates. This approach allows more selective or less popular journals to have equal representation to their high publishing counterparts.

Our article-based approach identifies most similar manuscripts to the preprint query, and we evaluated the journals that published these identified manuscripts. We embedded each query article into the space defined by the word2vec model (see above section 'Constructing a Document Representation for Life Sciences Text'). Once embedded, we selected manuscripts close to the query via Euclidean distance in the embedding space. Once identified, we return articles along with journals that published these identified articles.

We constructed a journal-based approach to accompany the article-based classifier while accounting for the overrepresentation of these high publishing frequency journals. We identified the most similar journals for this approach by constructing a journal representation in the same embedding space. We computed this representation by taking the average embedding of all published papers within a given journal. We then projected a query article into the same space and returned journals closest to the query using the same distance calculation described above. Both models were constructed using the scikit-learn k-Nearest Neighbors implementation [147] with the number of neighbors set to 10 as this is an appropriate number for our use case. We consider a prediction to be a true positive if the correct journal appears within our reported list of neighbors and evaluate our performance using 10-fold cross-validation on the training set along with test set evaluation.

Web Application for Discovering Similar Preprints and Journals

We developed a web application that places any bioRxiv or medRxiv preprint into the overall document landscape and identifies topically similar papers and journals (similar to [154]). Our application attempts to download the full text xml version of any preprint hosted on the bioRxiv or medRxiv server and uses the lxml package (version num) to extract text. If the xml version isn't

available our application defaults to downloading the pdf version and uses PyMuPDF [155] to extract text from the pdf. The extracted text is fed into our CBOW model to construct a document embedding representation. We pass this representation onto our journal and article classifiers to identify journals based on the ten closest neighbors of individual papers and journal centroids. We implemented this search using the scikit-learn implementation of k-d trees. To run it more cost-effectively in a cloud computing environment with limited available memory, we sharded the k-d trees into four trees.

The app provides a visualization of the article's position within our training data to illustrate the local publication landscape, We used SAUCIE [156], an autoencoder designed to cluster single-cell RNA-seq data, to build a two-dimensional embedding space that could be applied to newly generated preprints without retraining, a limitation of other approaches that we explored for visualizing entities expected to lie on a nonlinear manifold. We trained this model on document embeddings of PMC articles that did not contain a matching preprint version. We used the following parameters to train the model: a hidden size of 2, a learning rate of 0.001, lambda_b of 0, lambda_c of 0.001, and lambda_d of 0.001 for 5000 iterations. When a user requests a new document, we can then project that document onto our generated two-dimensional space; thereby, allowing the user to see where their preprint falls along the landscape. We illustrate our recommendations as a shortlist and provide access to our network visualization at our website (https://greenelab.github.io/preprint-similarity-search/).

Analysis of the Preprints in Motion Collection

Our manuscript describes the large-scale analysis of bioRxiv. Concurrent with our work, another set of authors performed a detailed curation and analysis of a subset of bioRxiv [122] that was focused on preprints posted during the initial stages of the COVID-19 pandemic. The curated analysis was designed to examine preprints at a time of increased readership [157] and includes certain preprints posted from January 1st, 2020 to April 30th, 2020 [122]. We sought to contextualize this subset, which we term "Preprints in Motion" after the title of the preprint [122],

within our global picture of the bioRxiv preprint landscape. We extracted all preprints from the set reported in Preprints in Motion [122] and retained any entries in the bioRxiv repository. We manually downloaded the XML version of these preprints and mapped them to their published counterparts as described above. We used Pubmed Central's DOI converter [158] to map the published article DOIs with their respective PubMed Central IDs. We retained articles that were included in the PMCOA corpus and performed a token analysis as described to compare these preprints with their published versions. As above, we generated document embeddings for every obtained preprint and published article. We projected these preprint embeddings onto our publication landscape to visually observe the dispersion of this subset. We performed a time analysis that paralleled our approach for the full set of preprint-publication pairs to examine relationships between linguistic changes and the time to publication. The "Preprints in Motion" subset includes recent papers, and the longest time to publish in that set was 195 days; however, our bioRxiv snapshot contains both older preprint-published pairs and many with publication times longer than this timepoint. The optimum comparison would be to consider only preprints posted on the same days as preprints with the "Preprints in Motion" collection. However, based on our results examining publication rate over time, these preprints may not have made it entirely through the publication process. We performed a secondary analysis to control for the time since posting, where we filtered the bioRxiv snapshot to only contain publication pairs with publication time of less than or equal to 195 days.

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 [118]. Because the rate of change is rapid, we first analyzed category data and

compared our results with previous findings. As in previous reports [118], neuroscience remains the most common category of preprints, followed by bioinformatics (Supplemental Figure 18). Microbiology, which was fifth in the most recent report [118], has now surpassed evolutionary biology and genomics to move into third. When authors upload their preprints, they select from three result category types: new results, confirmatory results, or contradictory results. We found that nearly all preprints (97.5%) were categorized as new results, consistent with reports on a smaller set [159]. The results taken together suggest that while bioRxiv has experienced dramatic growth, how it is being used appears to have remained consistent in recent years.

Global analysis reveals similarities and differences between bioRxiv and PMC

Table 3 : Summary statistics for the bioRxiv, PMC, and NYTAC corpora.

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

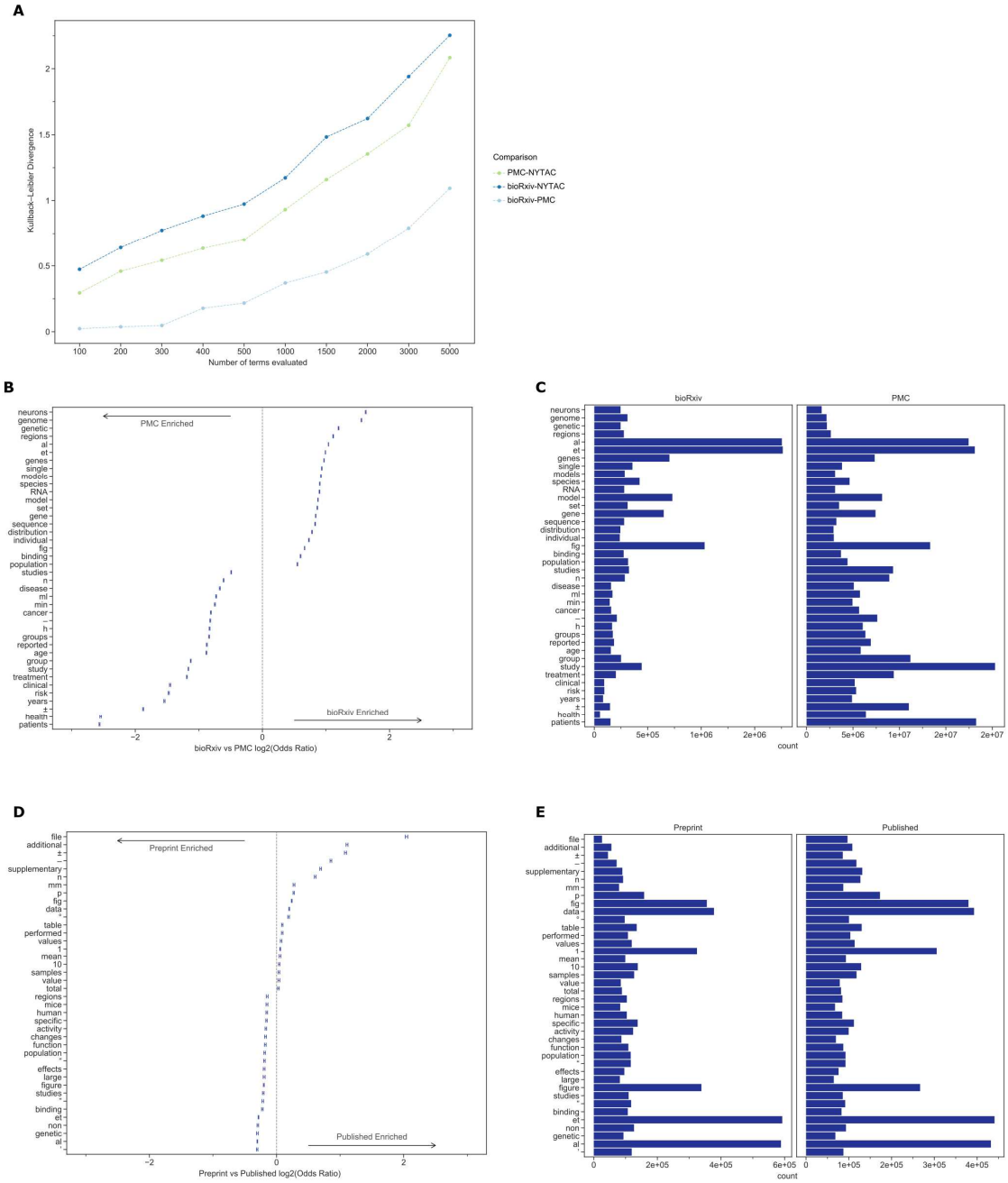


Figure 3 Corpora Comparison between bioRxiv and PMCOA

A. *The Kullback–Leibler divergence measures the extent to which the distributions, not specific tokens, differ from each other. The token distribution of bioRxiv and PMC corpora is more similar than these biomedical corpora are to the NYTAC one.* ***B.*** *The significant differences in token frequencies for the corpora appear to be driven by the fields with the highest uptake of bioRxiv, as terms from neuroscience and genomics are relatively more abundant in bioRxiv. We plotted the 95% confidence interval for each reported token.* ***C.*** *Of the tokens that differ between bioRxiv*

and PMC, the most abundant in bioRxiv are "et" and "al" while the most abundant in PMC is "study." ***D.*** *The significant differences in token frequencies for preprints and their corresponding published version often appear to be associated with typesetting and supplementary or additional materials. We plotted the 95% confidence interval for each reported token.* ***E.*** *The tokens with the largest absolute differences in abundance appear to be stylistic. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-one.*

Documents within bioRxiv were slightly longer than those within PMCOA, but both were much longer than those from the control (NYTAC) (Table 3). The average sentence length, the fraction of pronouns, and the use of the passive voice were all more similar between bioRxiv and PMC than they were to NYTAC(Table 3). The Kullback–Leibler (KL) divergence of term frequency distributions between bioRxiv and PMCOA were low, especially among the top few hundred tokens (Figure 3A). As more tokens were incorporated, the KL divergence started to increase but remained much lower than the biomedical corpora compared against NYTAC. We provide a listing of the top 100 most frequently occurring tokens from all three corpora in our supplement (Supplemental Table 12). These findings support our notion that bioRxiv is linguistically similar to the PMCOA repository.

The terms "neurons", "genome", and "genetic", which are common in genomics and neuroscience, were more common in bioRxiv than PMCOA while others associated with clinical research, such as "clinical" "patients" and "treatment" were more common in PMCOA (Figure 3B, 3C and Supplementary Figure 19). When controlling for the differences in the body of documents to identify textual changes associated with the publication process, we found that tokens such as "et" "al" were enriched for bioRxiv while "±", "–" were enriched for PMCOA (Figure 3D, 3E). When removing special and single-character tokens, data availability and presentation related terms "fle", "supplementary", "fig" appeared enriched for published articles, and research related terms "mice", "activity", "neurons" appeared enriched for bioRxiv (Supplementary Figure 20). Furthermore, we found that specific changes appeared to be related to journal styles: "figure" was more common in bioRxiv while "fig" was relatively more common in PMCOA. Other changes appeared to be associated with an increasing reference to content external to the manuscript

itself: the tokens “supplementary”, “additional” and “file” were all more common in PMCOA than bioRxiv, suggesting that journals are not simply replacing one token with another but that there are more mentions of such content after peer review.

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

Following our analysis of tokens, we examined the principal components of document embeddings derived from bioRxiv. We found that the top principal components separated methodological approaches and research fields. Preprints from certain topic areas that spanned approaches from informatics-related to cell biology could be distinguished using these principal components (see Supplementary Results).

Document embedding similarities reveal unannotated preprint-publication pairs

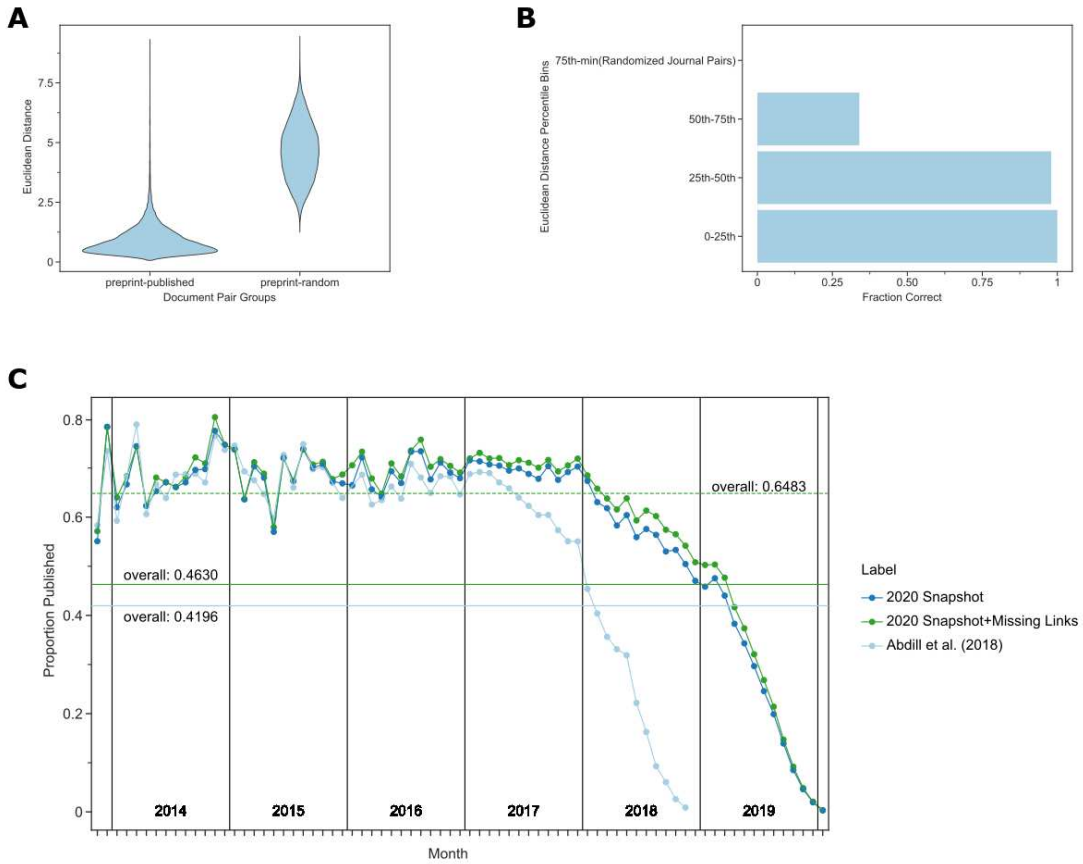


Figure 4 Filling in preprint and corresponding publication links

A. *Preprints are closer in document embedding space to their corresponding peer-reviewed publication than they are to random papers published in the same journal.* ***B.*** *Potential preprint-publication pairs that are unannotated but within the 50th percentile of all preprint-publication pairs in the document embedding space are likely to represent true preprint-publication pairs. We depict the fraction of true positives over the total number of pairs in each bin. Accuracy is derived from the curation of a randomized list of 200 potential pairs (50 per quantile) performed in duplicate with a third rater used in the case of disagreement.* ***C.*** *Most preprints are eventually published. We show the publication rate of preprints since bioRxiv first started. The x-axis represents months since bioRxiv started, and the y-axis represents the proportion of preprints published given the month they were posted. The light blue line represents the publication rate previously estimated by Abdill et al. [118]. The dark blue line represents the updated publication rate using only CrossRef-derived annotations, while the dark green line includes annotations derived from our embedding space approach. The horizontal lines represent the overall proportion of preprints published as of the time of the annotated snapshot. The dashed horizontal line represents the overall proportion published preprints for preprints posted before 2019. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-two.*

Distances between preprints and their corresponding published versions were nearly always lower than preprints paired with a random article published in the same journal (Figure 4A). This suggested that embedding distances may predict the published form of preprints. We directly tested this by selecting low-distance but unannotated preprint-publication pairs and curating the extent to which they represented matching documents. Approximately 98% of our 200 pairs with an embedding distance in the 0-25th and 25th-50th percentile bins were successfully matched with their published counterpart (Figure 4B). These two bins contained 1,542 preprint-article pairs, suggesting that many preprints may have been published but not previously connected with their published versions. There is a particular enrichment for preprints published but unlinked within the 2017-2018 interval (Figure 4C). We expected a higher proportion of such preprints before 2019 (many of which may not have been published yet); however, observing relatively few missed annotations before 2017 was against our expectations. There are several possible explanations for this increasing fraction of missed annotations. As the number of preprints posted on bioRxiv grows, it may be harder for bioRxiv to establish a link between preprints and their published counterparts simply due to the scale of the challenge. It is possible that the set of authors participating in the preprint ecosystem is changing and that new participants may be less likely to report missed publications to bioRxiv. Finally, as familiarity with preprinting grows, it is possible that authors are posting preprints earlier in the process and that metadata fields that bioRxiv uses to establish a link may be less stable.

Preprints with more versions or more text changes relative to their published counterpart took longer to publish

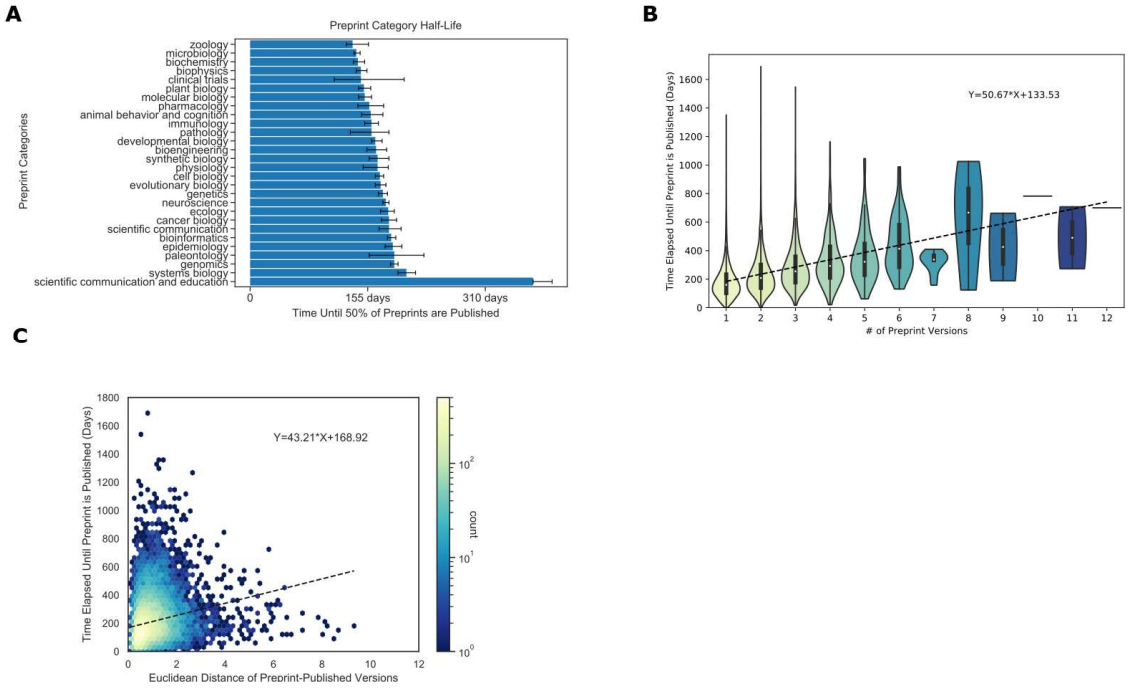


Figure 5 Time taken for preprints to become published

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

The process of peer review includes several steps, which take variable amounts of time [160], and we sought to measure if there is a difference in publication time between author-selected categories of preprints (Figure 5A). Of the most abundant preprint categories microbiology was

the fastest to publish (140 days, (137, 145 days) [95% CI]) and genomics was the slowest (190 days, (185, 195 days) [95% CI]) (Figure 5A). We did observe category-specific differences; however, these differences were generally modest, suggesting that the peer review process did not differ dramatically between preprint categories. One exception was the Scientific Communication and Education category, which took substantially longer to be peer-reviewed and published (373 days, (373, 398 days) [95% CI]). This hints that there may be differences in the publication or peer review process or culture that apply to preprints in this category.

Examining peer review's effect on individual preprints, we found a positive correlation between preprints with multiple versions and the time elapsed until publication (Figure 5B). Every additional preprint version was associated with an increase of 51 days before a preprint was published. This time duration seems broadly compatible with the amount of time it would take to receive reviews and revise a manuscript, suggesting that many authors may be updating their preprints in response to peer reviews or other external feedback. The embedding space allows us to compare preprint and published documents to determine if the level of change that documents undergo relates to the time it takes them to be published. Distances in this space are arbitrary and must be compared to reference distances. We found that the average distance of two randomly selected papers from the bioinformatics category was 4.470, while the average distance of two randomly selected papers from bioRxiv was 5.343. Preprints with large embedding space distances from their corresponding peer-reviewed publication took longer to publish (Figure 5C): each additional unit of distance corresponded to roughly forty-three additional days.

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

Preprints with similar document embeddings share publication venues

We developed an online application that returns a listing of published papers and journals closest to a query preprint in document embedding space. This application uses two k-nearest neighbor classifiers that achieved better performance than our baseline model (Supplemental Figure 21) to

identify these entities. Users supply our app with digital object identifiers (DOIs) from bioRxiv or medRxiv, and the corresponding preprint is downloaded from the repository. Next, the preprint's PDF is converted to text, and this text is used to construct a document embedding representation. This representation is supplied to our classifiers to generate a listing of the ten papers and journals with the most similar representations in the embedding space (Figures 6A, 6B and 6C). Furthermore, the user-requested preprint's location in this embedding space is then displayed on our interactive map, and users can select regions to identify the terms most associated with those regions (Figures 6D and 6E). Users can also explore the terms associated with the top 50 PCs derived from the document embeddings, and those PCs vary across the document landscape. You can access this application using the following url: https://greenelab.github.io/preprint-similarity-search/

Figure 6 Preprint Similarity Search Walkthrough

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

and (**C**) the ten closest journals. **D.** The app also displays the location of the query preprint in PMC. **E.** Users can select a square within the landscape to examine statistics associated with the square, including the top journals by article count in that square and the odds ratio of tokens.

Contextualizing the Preprints in Motion Collection

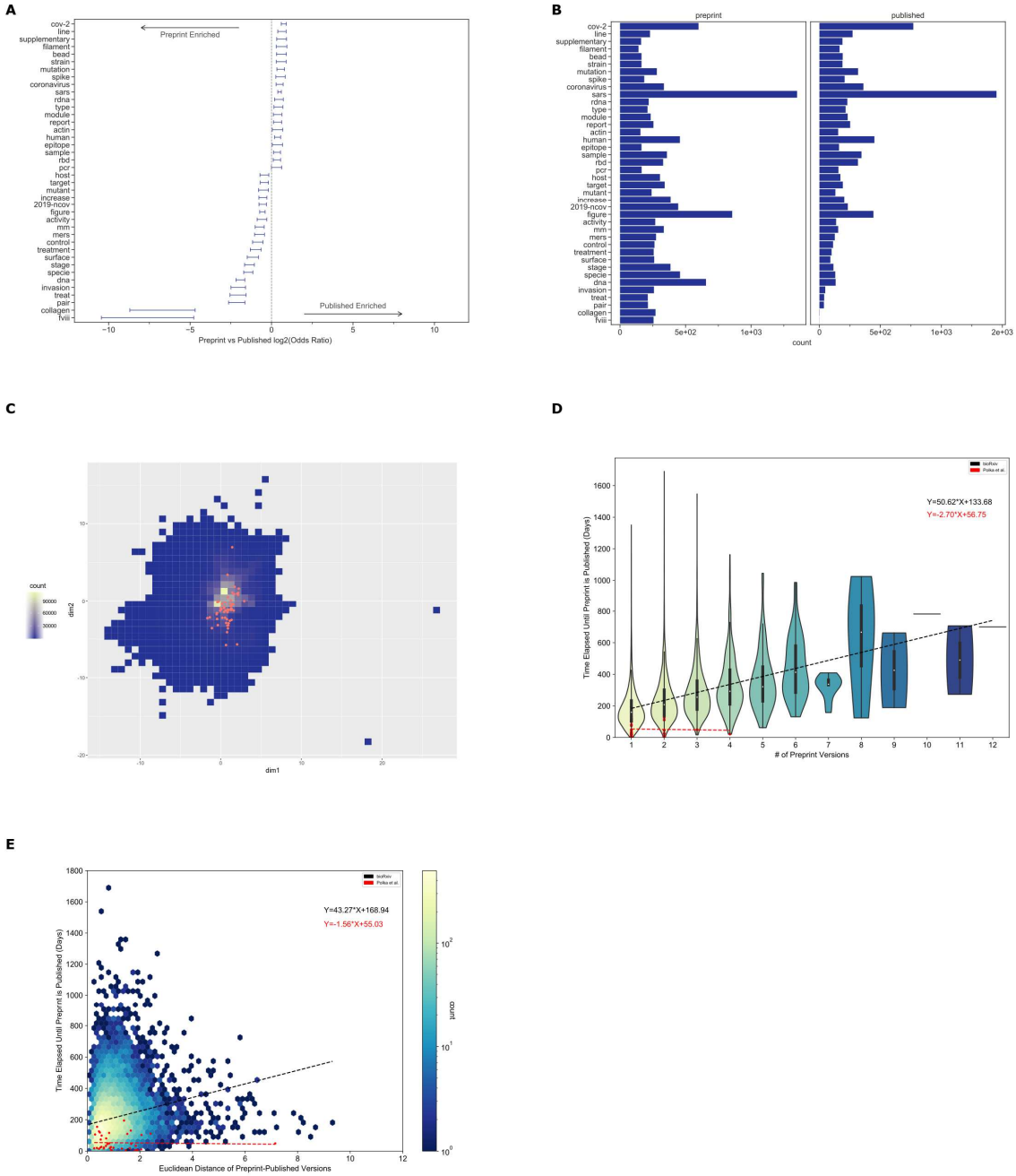


Figure 7 Contextualing Preprints in Motion

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

The Preprints in Motion collection included a set of preprints posted during the first four months of 2020. We examined the extent to which preprints in this set were representative of the patterns that we identified from our analysis on all of bioRxiv. As with all of bioRxiv, typesetting tokens changed between preprints and their paired publications. Our token-level analysis identified certain patterns consistent with our findings across bioRxiv (Figure 7A and 7B). However, in this set, we also observe changes likely associated with the fast-moving nature of COVID-19 research: the token “2019-ncov” became less frequently represented while “sars” and “cov-2” became more represented, likely due to a shift in nomenclature from “2019-nCoV” to “SARS-CoV-2”. The Preprints in Motion were not strongly colocalized in the linguistic landscape, suggesting that the collection covers a diverse set of research approaches (Figure 7C). Preprints in this collection were published faster than the broader set of bioRxiv preprints (Figure 7D and 7E). We see the same trend when filtering the broader bioRxiv set to only contain preprints

published within the same timeframe as this collection (Supplemental Figures 22A and 22B). The relationship between time to publication and the number of versions (Figure 7D and Supplemental Figure 22A) and the relationship between time to publication and the amount of linguistic change (Figure 7E and Supplemental Figure 22B) were both lost in the Preprints in Motion set. Our findings suggest that Preprints in Motion changed during publication in ways aligned with changes in the full preprint set but that peer review was accelerated in ways that broke the time dependencies observed with the full bioRxiv set.

Discussion and Conclusions

BioRxiv is a constantly growing repository that contains life science preprints. Over 77% of bioRxiv preprints with a corresponding publication in our snapshot were successfully detected within Pubmed Central's Open Access Corpus (PMCOA). This suggests that most work from groups participating in the preprint ecosystem is now available in final form for literature mining and other applications. Most research on bioRxiv preprints has examined their metadata; we examine the text content as well. Throughout this work, we sought to analyze the language within these preprints and understand how it changes in response to peer review.

Our global corpora analysis found that writing within bioRxiv is consistent with the biomedical literature in the PMCOA repository, suggesting that bioRxiv is linguistically similar to PMCOA. Token-level analyses between bioRxiv and PMCOA suggested that research fields drive significant differences; e.g., more patient-related research is prevalent in PMCOA than bioRxiv. This observation is expected as preprints focused on medicine are supported by the complementary medRxiv repository [114]. Token-level analyses for preprints and their corresponding published version suggest that peer review may focus on data availability and incorporating extra sections for published papers; however, future studies are needed to ascertain individual token level changes as preprints venture through the publication process. One future avenue of research could examine the differences between only preprints and

accepted author manuscripts within Pubmed Central to identify changes prior to journal publication.

Document embeddings are a versatile way to examine language contained within preprints, understanding peer review's effect on preprints, and provide extra functionality for preprint repositories. Our approach to generate document embeddings was focused on interpretability instead of predictive performance; however, using more advanced strategies to generate document vectors such as Doc2Vec [133] or BERT [161] should increase predictive performance. Examining linguistic variance within document embeddings of life science preprints revealed that the largest source of variability was informatics. This observation bisects the majority of life science research categories that have integrated preprints within their publication workflow. This embedding space could also be used to quantify sentiment trends or other linguistic features. Furthermore, methodologies for uncovering latent scientific knowledge [162] may be applicable in this embedding space.

Preprints are typically linked with their published articles via bioRxiv manually establishing links or authors self-reporting that their preprint has been published; however, gaps can occur as preprints change their appearance through multiple versions or authors do not notify bioRxiv. Our work suggests that document embeddings can help fill in missing links within bioRxiv. Furthermore, our analysis reveals that the publication rate for preprints is higher than previously estimated, even though our analysis can only account for published open access papers. Our results raise the lower bound of the total preprint publication fraction; however, the true fraction is necessarily higher. Future work, especially that which aims to assess the fraction of preprints that are eventually published, should account for the possibility of missed annotations.

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

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

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

Supplemental Section

Document embeddings derived from bioRxiv reveal fields and subfields

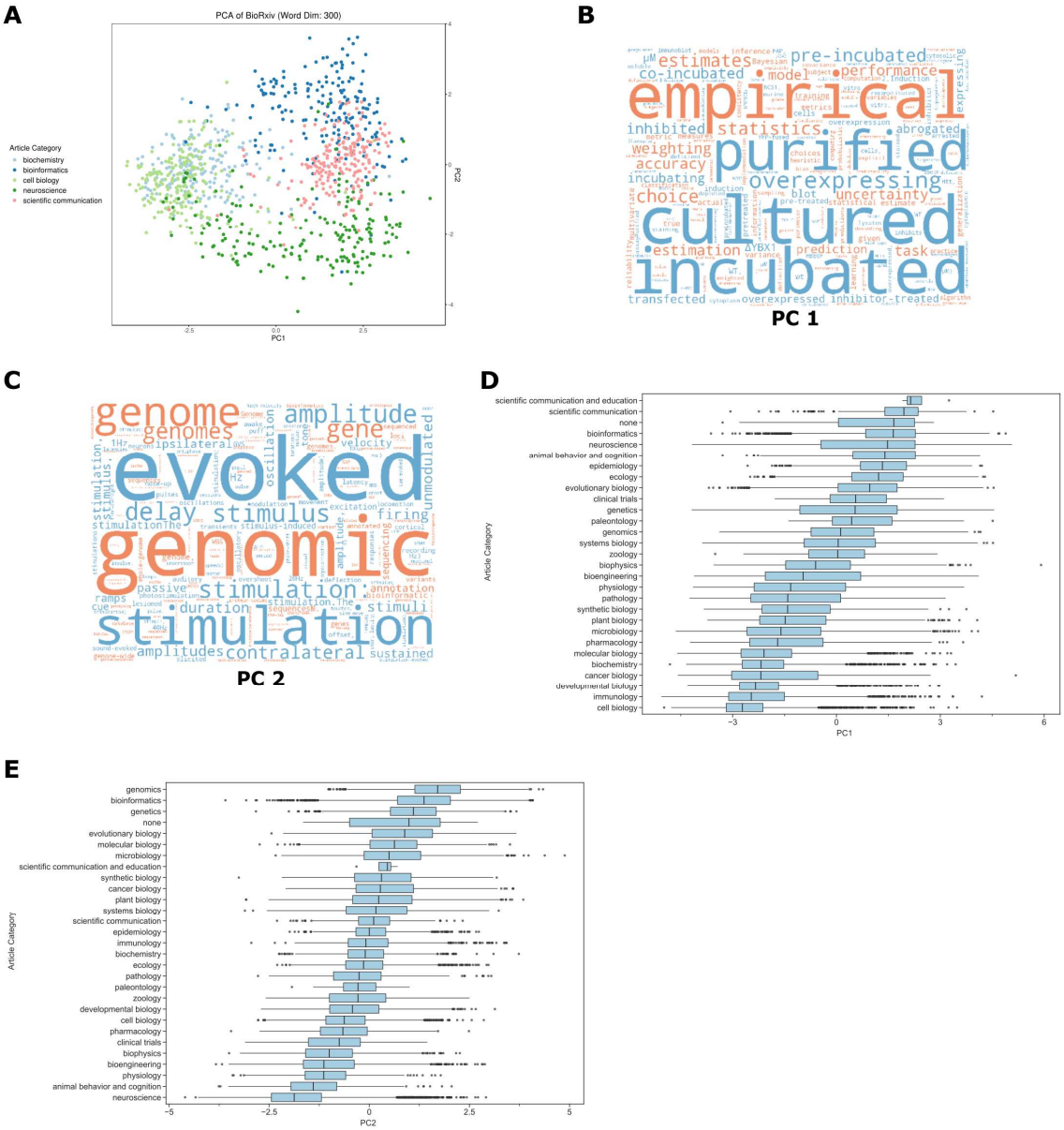


Figure 8 PCA analysis on preprint document embeddings

A. *Principal components (PC) analysis of bioRxiv word2vec embeddings groups documents based on author-selected categories. We visualized documents from key categories on a scatterplot for the first two PCs. The first PC separated cell biology from informatics-related fields, and the second PC separated bioinformatics from neuroscience fields.* ***B.*** *A word cloud visualization of PC1. Each word cloud depicts the cosine similarity score between tokens and the first PC. Tokens in orange were most similar to the PC's positive direction, while tokens in blue were most similar to the PC's negative direction. The size of each token indicates the magnitude of the similarity.* ***C.*** *A word cloud visualization of PC2, which separated bioinformatics from neuroscience. Similar to the first PC, tokens in orange were most similar to the PC's positive*

direction, while tokens in blue were most similar to the PC's negative direction. The size of each token indicates the magnitude of the similarity. ***D.*** *Examining PC1 values for each article by category created a continuum from informatics-related fields on the top through cell biology on the bottom. Specific article categories (neuroscience, genetics) were spread throughout PC1 values.* ***E.*** *Examining PC2 values for each article by category revealed fields like genomics, bioinformatics, and genetics on the top and neuroscience and behavior on the bottom. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s1.*

Document embeddings provide a means to categorize the language of documents in a way that takes into account the similarities between terms [163,164,165]. We found that the first two PCs separated articles from different author-selected categories (Supplementary Figure 8A). Certain neuroscience papers appeared to be more associated with the cellular biology direction of PC1, while others seemed to be more associated with the informatics-related direction (Supplementary Figure 8A). This suggests that the concepts captured by PCs were not exclusively related to their field.

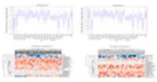
Visualizing token-PC similarity revealed tokens associated with certain research approaches (Supplementary Figures 8B and 8C). Token association of PC1 shows the separation of cell biology and informatics-related fields through tokens: "empirical", "estimates" and "statistics" depicted in orange and "cultured" and "overexpressing" shown in blue (Supplementary Figure 8B and Supplementary Table 5). Association for PC2 shows the separation of bioinformatics and neuroscience via tokens: "genomic", "genome" and "genomes" depicted in orange and "evoked", "stimulus" and "stimulation" shown in blue (Supplementary Figure 8C and Supplementary Table 6).

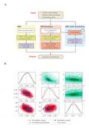
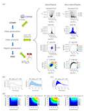
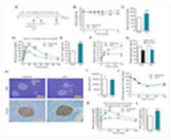
Examining the value for PC1 across all author-selected categories revealed an ordering of fields from cell biology to informatics-related disciplines (Supplementary Figure 8D). These results suggest that a primary driver of the variability within the language used in bioRxiv could be the divide between informatics and cell biology approaches. A similar analysis for PC2 suggested that neuroscience and bioinformatics present a similar language continuum (Supplementary Figure 8E). This result supports the notion that bioRxiv contains an influx of neuroscience and

bioinformatics-related research results. For both of the top two PCs, the submitter-selected category of systems biology preprints was near the middle of the distribution and had a relatively large interquartile range when compared with other categories (Supplementary Figures 8D and 8E), suggesting that systems biology is a broader subfield containing both informatics and cellular biology approaches.

Examining the top five highest-scoring and bottom five lowest-scoring systems biology preprints along PC1 reinforces its dichotomous theme (Supplementary Table 4). Preprints with the highest values [166,167,168,169,170] included software packages, machine learning analyses, and other computational biology manuscripts, while preprints with the lowest values [171,172,173,174,175] were focused on cellular signaling and protein activity. We provide the rest of our 50 generated PCs in our online repository (see Software and Data Availability).

Table 4 PC1 divided the author-selected category of systems biology preprints along an axis from computational to molecular approaches.

Title [citation]	PC1	License	Figure Thumbnail
Conditional Robust Calibration (CRC): a new computational Bayesian methodology for model parameters estimation and identifiability analysis [166]	4.522818390064091	None	
FPtool a software tool to obtain in silico genotype-phenotype signatures and fingerprints based on	4.348956760251298	CC-BY	

massive model simulations [167]			
GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation [168]	4.259104249060651	CC-BY-NC-ND	
Notions of similarity for computational biology models [169]	4.079855550647664	CC-BY-NC-ND	
SBpipe: a collection of pipelines for automating repetitive simulation and analysis tasks [170]	4.022240241143516	CC-BY-NC-ND	
Bromodomain inhibition reveals FGF15/19 as a target of epigenetic regulation and metabolic control [171]	- 3.4783803547922414	None	
Inhibition of Bruton's tyrosine kinase reduces NF-kB and NLRP3 inflammasome activity preventing insulin resistance and microvascular disease [172]	- 3.6926161167521476	None	

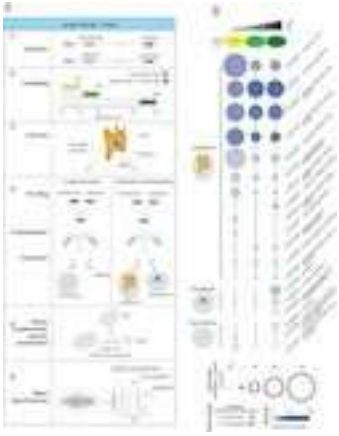
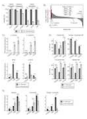
Spatiotemporal proteomics uncovers cathepsin-dependent host cell death during bacterial infection [173]	-3.728443135960558	CC-BY-ND	
NADPH consumption by L-cystine reduction creates a metabolic vulnerability upon glucose deprivation [174]	-3.7363965062637288	None	
AKT but not MYC promotes reactive oxygen species-mediated cell death in oxidative culture [175]	-3.8769231933681176	None	

Table 5 Top and bottom five cosine similarity scores between tokens and the PC1 axis.

Cosine Similarity (PC1, word)	word
0.6399154807185836	empirical
0.5995356000266072	estimates
0.5918321530159384	choice
0.5905550757923625	statistics
0.5832932491448216	performance
0.5803836474390357	accuracy
0.5757250459195589	weighting

0.5753027342288192	estimation
0.5730092178610916	uncertainty
0.5720493442813257	task
-0.4484093198386865	abrogated
-0.4490583645152233	transfected
-0.4500847285921068	incubating
-0.4531550791501111	inhibited
-0.4585422153514687	co-incubated
-0.4774721756292901	pre-incubated
-0.4793057689825842	overexpressing
-0.4839313193713342	purified
-0.4869885872803974	incubated
-0.5040798110023075	cultured

Table 6 Top and bottom five cosine similarity scores between tokens and the PC2 axis.

Cosine Similarity (PC2, word)	word
0.65930201597598	genomic
0.6333515216782134	genome
0.5974018685580009	gene
0.5796531207938461	genomes
0.5353687686155728	annotation
0.5310140161149529	sequencing
0.5197350376908197	sequencesM.
0.5181781615670665	genome,
0.5168781637087506	bioinformatic

0.513853407439108	WGS
-0.4589201401582101	duration
-0.4690482252758019	stimuli
-0.4712875761979691	amplitudes
-0.4772723570301678	contralateral
-0.4813219679071856	stimulation:
-0.4946709932017581	delay
-0.5111990014804086	stimulus
-0.5251288188682695	amplitude
-0.543586881182879	stimulation
-0.5467022203294039	evoked

CHAPTER 3

Detecting semantic shifts in biomedical literature through an intra-year and inter-year approach

This chapter is set to appear as a preprint with the following citation: Nicholson DN, Alquaddoomi F, Rubinetti V, Greene CS Detecting semantic shifts in biomedical literature through an intra-year and inter-year approach.

This is a co-authored paper where the main scientific contributions were by Nicholson DN who was advised by Greene CS. Alquaddoomi F and Rubinetti V assisted with the creation of the word-lapse website backend and front end respectively.

Introduction

Language is constantly evolving, and the meaning that we ascribe to words changes over time. For example, the word “nice” was used to mean foolish or innocent back in the 15th-17th century; then, it underwent a positive shift to its current meaning of “pleasant or delightful”[22]. These shifts occur for many reasons. For example, writers may use new metaphors or substitute words for others with similar meanings in a process known as metonymy [22]. Studying these shifts can provide a nuanced understanding of how language adapts to describe our world.

Scientific fields of inquiry also change, sometimes rapidly, as researchers devise and test new hypotheses and applications. For example, the repurposing of the CRISPR-Cas9 system to a pervasive tool for genome editing has altered how we discuss molecular entities. Microbes use this as an immune system to defend against viruses. Scientists repurposed this system for genome editing [176], leading to changes in the use of the term. Science is a field with substantial written communication [6], both via published papers [135] and preprints [4,177]. Examining scientific manuscripts with computational linguistics can reveal longitudinal trends in scientific research.

Studying changes in the use of word meanings is called semantic shift detection. Approaches for semantic shift detection examine time series datasets that capture word usage patterns, both with respect to frequency and structure. Typically, these time series are generated for individual words by training a unique model on text binned by a selected time period [32,178,179]. Methods are then applied to identify "change points" where a word's meaning has changed [180].

Semantic shifts have been examined in many sources. Analysis has included newspapers [35,182,183], books [178], reddit [36], and Twitter [184]. Researchers have examined topics in information retrieval [185], and in biomedicine COVID-19 has been examined multiple times [38,186,187]. The amount of open access biomedical literature has dramatically increased in the last two decades, laying the groundwork for the large-scale analysis of semantic shifts in biomedicine.

We examine these semantic shifts in this rapidly growing body of open access text. We include both published papers and preprints in our analysis. We found that novel strategies integrating multiple models for each year sidestepped the challenge of instability in the machine learning models and allowed us to estimate intra- and inter-year variability. We identify semantic change points for each token. We examine key cases and provide the full set of research products, including change points and machine learning models, as openly licensed tools for the community. We also created a webserver that allows users to analyze tokens of interest on the fly, examining both the most similar terms within a year and temporal trends.

Methods

Biomedical Corpora Examined

Pubtator Central

Pubtator Central is an open-access resource containing annotated abstracts and full-text annotated with entity recognition systems for biomedical concepts [128]. The methods used are TaggerOne [188] to tag diseases, chemicals, and cell line entities, GNormPlus [189] to tag genes,

SR4GN [190] to tag species, and tmVar [191] to tag genetic mutations. We initially downloaded this resource on December 07th, 2021, and processed over 30 million documents. This resource contains documents that date back to the pre-1800s to the year 2021; however, due to the low sample size in early years, we only used documents published from 2000 to 2021. The resource was subsequently updated with documents from 2021. We also downloaded a later version on March 09th, 2022, and merged both versions using each document's doc_id field to produce the corpus used in this analysis. We divided documents by publication year and then preprocessed each using spacy's en_core_web_sm model [56]. We replaced each tagged word or phrase with its corresponding entity type and entity id for every sentence that contained an annotation. Then, we used spacy to break sentences into individual tokens and normalized each token to its root form via lemmatization. After preprocessing, we used every sentence to train multiple natural language models designed to represent words based on their context.

Biomedical Preprints

BioRxiv [4] and MedRxiv [177] are repositories that contain preprints for the life science community. MedRxiv mainly focuses on preprints that mention patient research, while bioRxiv focuses on general biology. We downloaded a snapshot of both resources on March 4th, 2022, using their respective Amazon S3 bucket [192,193]. This snapshot contained 172,868 BioRxiv preprints and 37,517 MedRxiv preprints. These resources allow authors to post multiple versions of a single preprint. To prevent duplication bias, we filtered every preprint to its most recent version and sorted each preprint into its respective posted year. Unlike Pubtator Central, these filtered preprints do not contain any annotations. Therefore, we used TaggerOne [188] to tag every chemical and disease entity and GNormplus [189] to tag every gene and species entity for our preprint set. Once tagged, we used spacy to preprocess every preprint as described in our Pubtator Central section.

Constructing Word Embeddings for Semantic Change Detection

Word2vec [87] is a natural language processing model designed to model words based on their respective neighbors in the form of dense vectors. This suite of models comes in two forms, a skipgram model and a continuous bags of words (CBOW) model. The skipgram model generates these vectors by having a shallow neural network predict a word's neighbors given the word, while the CBOW model predicts the word given its neighbors. We used the CBOW model to construct word vectors for each year. Despite the power of these word2vec models, these models are known to differ both due to randomization within year and year-to-year variability across years [194,195,196,197]. To control for run-to-run variability, we examined both intra-year and inter-year relationships. Each year, we trained ten different CBOW models using the following parameters: vector size of 300, 10 epochs, minimum frequency cutoff of 5, and a window size of 16 for abstracts. Every model has its own unique vector space following training, making it difficult to compare two models without a correction step. We used orthogonal Procrustes [198] to align models. We aligned all trained CBOW models for the Pubtator Central dataset to the first model trained in 2021. Likewise, we aligned all CBOW models for the BioRxiv/MedRxiv dataset to the first model trained in 2021. We used UMAP [199] to visually examine the aligned models. We trained this model using the following parameters: cosine distance metric, random_state of 100, 25 for n_neighbors, a minimum distance of 0.99, and 50 n_epochs.

Detecting semantic changes across time

Once word2vec models are aligned, the next step is to detect semantic change.

Semantic change events are often detected through time series analysis [200]. We constructed a time series sequence for every token by calculating its distance within a given year (intra-year) and across each year (inter-year). We used the model pairs constructed from the same year to calculate an intra-year distance. Then, we calculated the cosine distance between each token and its corresponding counterpart for every generated pair. Cosine distance is a metric bounded between zero and two, where a score of zero means two vectors are the same, and a score of

two means both vectors are different. For the inter-year distance, we used the Cartesian product of every model between two years and calculated the distance between tokens in the same way as the intra-year distance. Following both calculations, we combined both metrics by taking the ratio of the average inter-year distance over the average intra-year distance. Through this approach, tokens with high intra-year instability will be penalized and vice-verse for more stable tokens. Along with token distance calculations, it has been shown that including token frequency improves results compared to using distance alone [201]. We calculated token frequency as the ratio of token frequency in the more recent year over the frequency of the previous year. Then, we combined both the frequency and distance ratios to make the final metric.

Following time series construction, we performed change point detection, which is a process that uses statistical techniques to detect abnormalities within a given time series. We used the CUSUM algorithm [181] to detect these abnormalities. This algorithm uses a rolling sum of the differences between two timepoints and checks whether the sum is greater than a threshold. A changepoint is considered to have occurred if the sum is greater than a threshold. We used the 99th percentile on every generated timepoint as the threshold. Then, we ran the CUSUM algorithm using a drift of 0 and default settings for all other parameters.

Results

Models can be aligned and compared within and between years

We examined how the usage of tokens in biomedical text changes over time. Our evaluation was derived from machine learning models designed to predict the actual token given a portion of its surrounding tokens. Each token was represented as a vector in a coordinate space constructed by these models. However, training these models is stochastic, which results in arbitrary coordinate spaces. Model alignment is an essential step in allowing word2vec models to be compared [26,202]. Before alignment, each model has its own unique coordinate space (Figures 9A), and each word is represented within that space (Figure 9B). Alignment projects every model

onto a shared coordinate space (Figure 9C), enabling direct token comparison. We randomly selected 100 tokens to confirm that alignment worked as expected. In aligned models, tokens in the global spcae were more similar to themselves within year than between years, while identical tokens in unaligned models were completely distinct (Figure 9D). Local distances were unaffected by alignment (Figure 9D), as token-neighbor distances were unaffected by the alignment procedure.

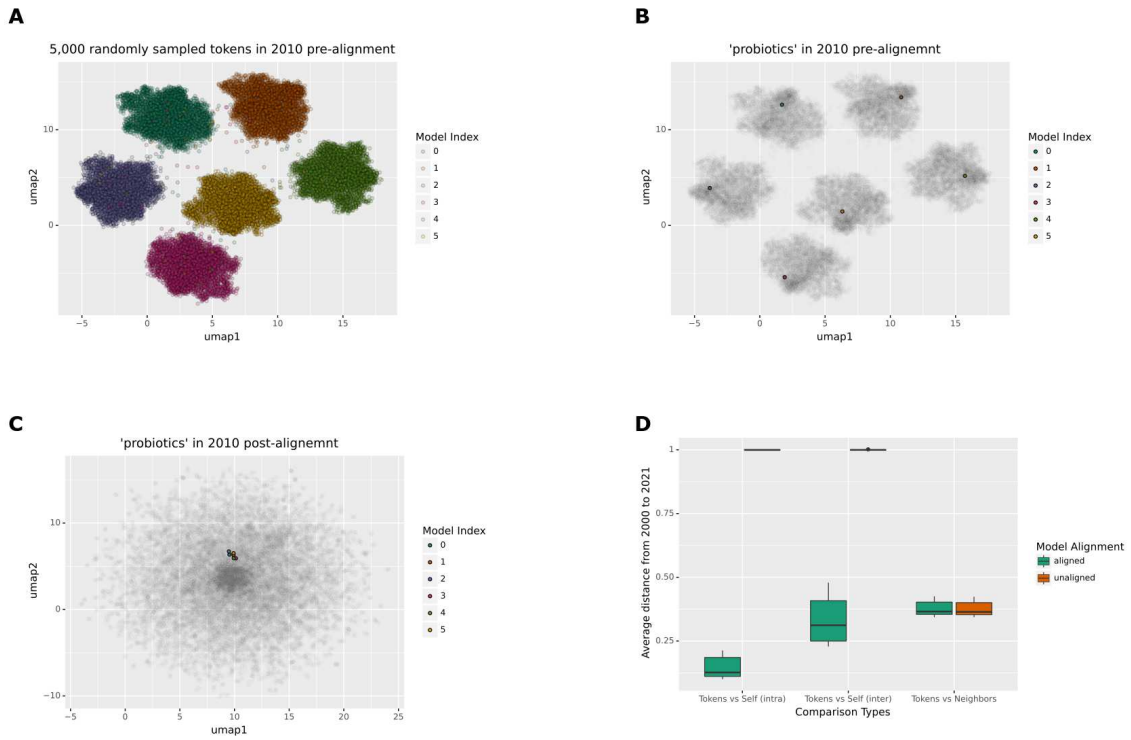


Figure 9 Confirming Alignment for Word2Vec Models

A. Without alignment, each word2vec model has its own coordinate space. This is a UMAP visualization of 5000 randomly sampled tokens from 5 distinct Word2Vec models trained on the text published in 2010. Each data point represents a token, and the color represents the respective Word2Vec model. B. The highlighted token 'probiotics' shows up in its respective clusters. Each data point represents a token, and the color represents the Word2Vec model. C. After the alignment step, the token 'probiotic' is closer in vector space. Each data point represents a token, and the color represents the different Word2Vec models. D. In the global coordinate space, token distances appear to be vastly different without alignment, but become closer upon alignment, while local distances, evaluated using neighbors, are unaffected. This boxplot shows

the average distance of 100 randomly sampled tokens shared in every year from 2000 to 2021. The x-axis shows the various groups being compared (tokens against themselves via intra-year and inter-year distances and tokens against their corresponding neighbors. The y axis shows the averaged distance for every year.

The landscape of biomedical publishing has changed rapidly during the period of our dataset. The texts for our analysis were open access manuscripts available through PubMed Central. The growth in the amount of available text and the uneven adoption of open access publishing during the interval studied was expected to induce changes in the underlying machine learning models, making comparisons more difficult. We found that the number of tokens available for model building, i.e., those in PMC OA, increased dramatically during this time (Figure 10A). This was expected to create a pattern where models trained in earlier years were more variable than those from later years simply due to the limited sample size in early years. We aimed to correct for this change in the underlying models by developing a statistic that, instead of using pairwise comparisons of token distances between individual models, integrated multiple models for each year by comparing tokens' intra- and inter-year variabilities. We defined the statistic as the ratio of the average distance between two years over the sum of the average distance within each year respectively.

The landscape of biomedical publishing has changed rapidly during the period of our dataset. The texts for our analysis were open access manuscripts available through PubMed Central. The growth in the amount of available text and the uneven adoption of open access publishing during the interval studied was expected to induce changes in the underlying machine learning models, making comparisons more difficult. We found that the number of tokens available for model building, i.e., those in PMC OA, increased dramatically during this time (Figure 10A). This was expected to create a pattern where models trained in earlier years were more variable than those from later years simply due to the limited sample size in early years. We aimed to correct for this change in the underlying models by developing a statistic that, instead of using pairwise comparisons of token distances between individual models, integrated multiple models for each

year by comparing tokens' intra- and inter-year variabilities. We defined the statistic as the ratio of the average distance between two years over the sum of the average distance within each year respectively.

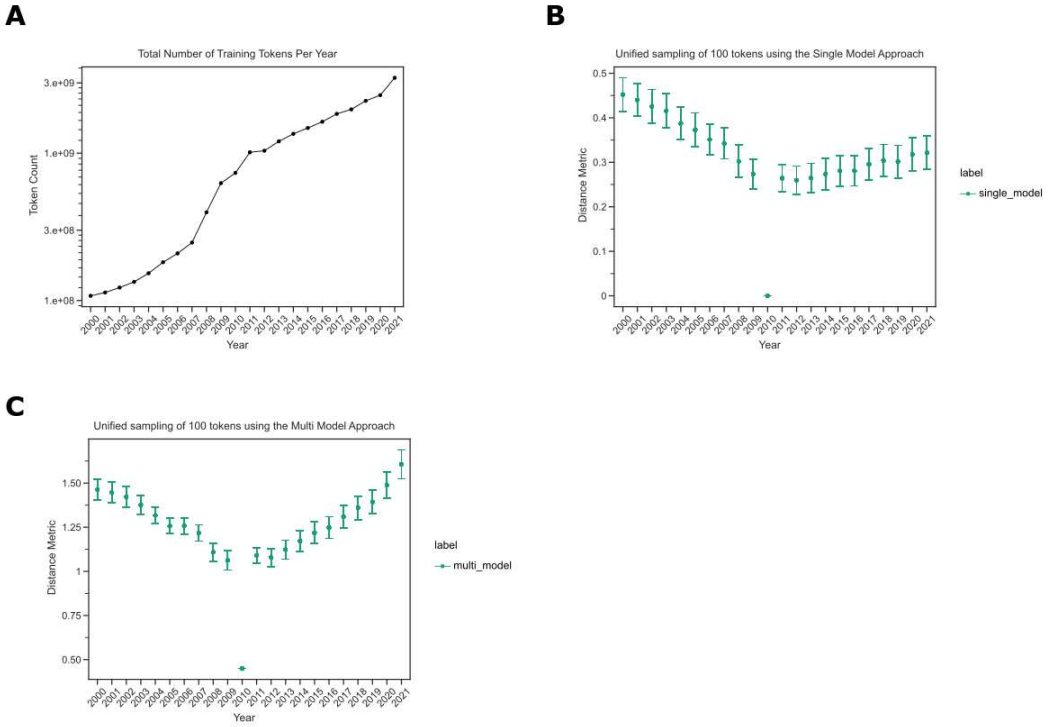


Figure 10 Examing our novel ratio metric over the years

A. The number of tokens our models have trained on increases over time. This line plot shows the number of unique tokens seen by our various machine learning models. The x-axis depicts the year and the y-axis shows the token count. B. Earlier years compared to 2010 have greater distances than later years. This confidence interval plot shows the collective distances obtained by sampling 100 tokens that are present from every year using a single model approach. The x-axis shows a given year and the y-axis shows the distance metric. C. Later years have a lower intra-distance variability compared to the earlier years. This confidence interval plot shows the collective distances obtained by sampling 100 tokens that are present from every year using our multi-model approach. The x-axis shows a given year and the y-axis shows the distance metric.

We expected most tokens to undergo minor changes from year to year, while substantial changes likely suggested model drift as opposed to true linguistic change. We measured the extent to which tokens differed from themselves using the standard single-model approach and our

integrated statistic. We filtered the token list to only contain tokens present in every year and compared their distance to the midpoint year, 2010, using the single-model and integrated-models strategies. We found that distances tended were markedly larger in the earliest years, where we expected models to be least stable, using the traditional approach (Figure 10B). The integrated model approach did not display the same pattern in the earliest years (Figure 10C). Both trends reinforce that training on smaller corpora will lead to high variation and that an integrated model strategy is needed [196]. Based on these results, we used the integrated-model strategy to calculate inter-year token distances for the remainder of this work.

Terms exhibit detectable changes in usage

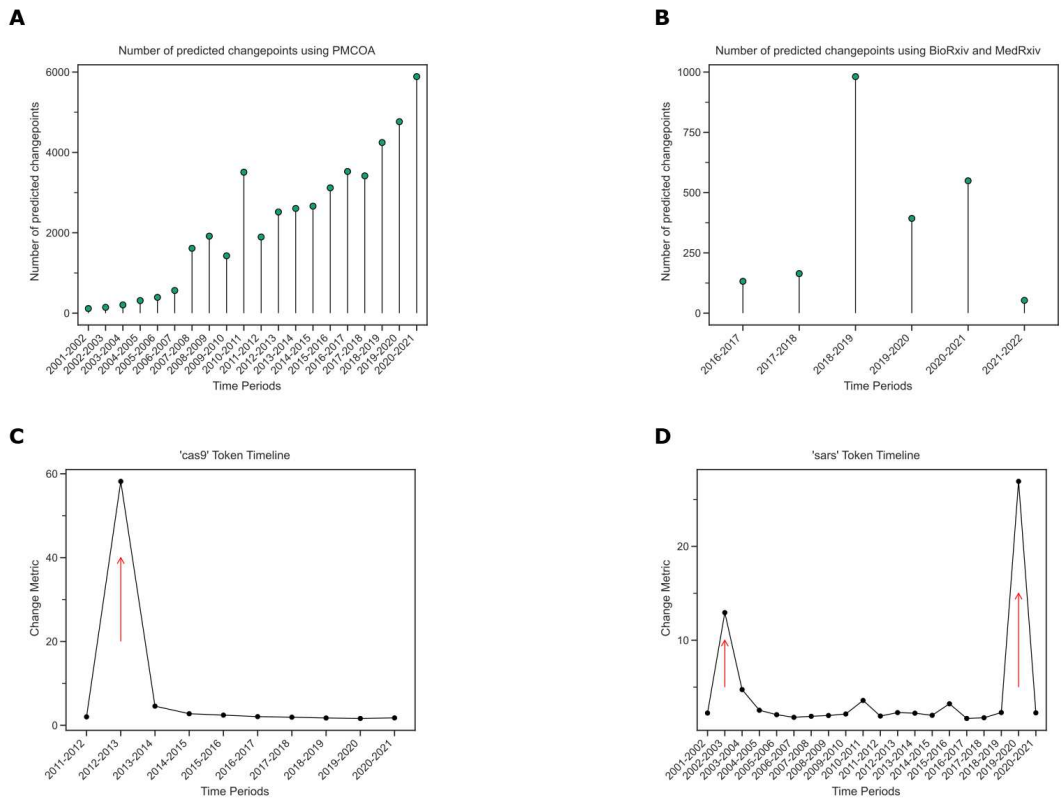


Figure 11 Reporting Detected Change points for PMCOA and bioRxiv

A. The number of changepoints increases over time in PMCOA. The x-axis shows the various time periods, while the y-axis depicts the number of detected changepoints. B. Regarding preprints, the greatest number of changepoints was during 2018-2019. The x-axis shows the

various time periods, while the y-axis depicts the number of detected changepoints. C. The token 'cas9' was detected to have a changepoint at 2012-2013. The x-axis shows the time period since the first appearance of the token, and the y-axis shows the change metric. D. 'sars' has two detected changepoints within the PMCOA corpus. The x-axis shows the time period since the first appearance of the token, and the y-axis shows the change metric.

We next sought to identify tokens that changed during the 2000-2021 interval for the text from PubMed Central's Open Access Corpus (PMCOA) and the 2015-2022 interval for our preprint corpus. We performed change point detection using the CUSUM algorithm with distances calculated with the integrated-model approach to correct for systematic differences in the underlying corpora. We found 41281 terms with a detected change point from PMCOA and 2266 terms from preprints (Figures 11A and 11B), and the vast majority (38019 for PMCOA and 2260 for preprints) had just a single change-point.

We explored individual change points. We detected one in PMCOA for 'cas9' from 2012 to 2013 (Figure 11C). Before the change point, its closest neighbors were related genetic elements (e.g., 'cas'1-3). After the change point, its closest neighbors became terms related to targeting, sgRNA, and gRNA, as well as other genome editing strategies, 'talen' and 'zfns' (Table 7). For some terms, we detected multiple change points within the studied interval. We detected change points for 'SARS' from 2002 to 2003 and 2019 to 2020 (Figure 11D), consistent with the emergences of SARS-CoV [203] and SARS-CoV-2 [204,205] as observed human pathogens. We found miscellaneous neighbors before each change point, with use consistent with the acronym for Severe Acute Respiratory Syndrome after each (Tables 8 and 9).

Out of all change points, we observed 200 tokens with at least one change point in each corpus. Only 25 of the 200 terms were detected to have simultaneous changes between the preprint and PMCOA corpora. We examined the overlap of detected change points between preprints and published articles. Many of these 25 were related to the COVID-19 pandemic (Supplementary Table 13). The complete set of detected change points is available for further analysis (see Data Availability and Software).

Table 7 The fifteen most similar neighbors to the token 'cas9' for the years 2012 and 2013.

2012	2013
cas2	sgrna
crispr1	talen
cas3	spcas9
cas1	zfns
cas10	grna
crispr3	zfn
tracrrna	dcas9
crispr	nickase
csn1	pcocas9
crispr4	crispr
cas7	sgrnas
cas6e	meganuclease
cas4	tracrrna
cse1	crispri
cas6	crrna

Table 8 The fifteen most similar neighbors to the token 'sars' for the years 2002 and 2003.

2002	2003
qsar	species_227859
herbicidal	mesh_c000657245
antiplasmodial	severe acute respiratory syndrome-related coronavirus (species_694009)
arylpiperazine	unidentified human coronavirus (species_694448)
a]pyridine	SARS1 (gene_6301)
leishmanicidal	ebola virus sp. (species_205488)
naphthyridine	pandemic

indolo[2,1	coronavirus infections (mesh_d018352)
b]quinazoline-6,12	coronavirus
nematocidal	ebola virus (species_1570291)
f]isoxazolo[2,3	severe acute respiratory syndrome (mesh_d045169)
5-(4	paramyxovirus
cholinephosphotransferase	viruse
oxovanadium(iv	drosten
catecholase	virologist

Table 9 The fifteen most similar neighbors to the token 'sars' for the years 2002 and 2003.

2019	2020
g.o.	sar
nsp13	mers
40/367	cov
lissodendoryx	sars-1
lutken	severe acute respiratory syndrome-related coronavirus (species_694009)
sarr	coronaviruse
sar	middle east respiratory syndrome-related coronavirus (species_1335626)
ophiura ophiura (species_72673)	cov.
verrill	coronavirus infections (mesh_d018352)
hirondelle	mers-
kobelt	covs
azorean	severe acute respiratory syndrome coronavirus 2 (species_2697049)
rusby	severe acute respiratory syndrome (mesh_d045169)

d’orbigny	sarscov
psychropotes longicauda	sarscov-2
(species_55639)	

The word-lapse application is an online resource for manual examination of biomedical tokens

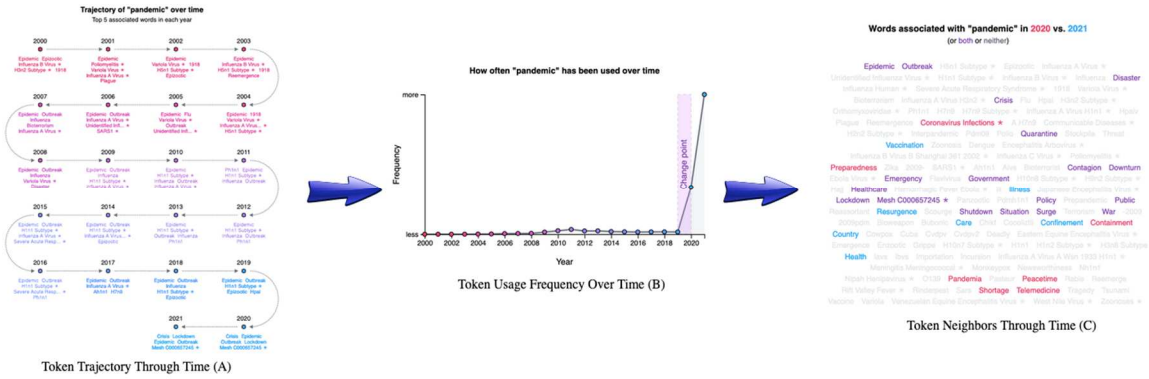


Figure 12 Walkthrough of the word-lapse manuscript

A. The trajectory visualization of the token ‘pandemic’ through time. It starts at the first mention of the token and progresses through each subsequent year. Every data point shows the top five neighbors for the respective token. B. The usage frequency of the token ‘pandemic’ through time. The x-axis shows the year, and the y-axis shows the frequency for each token. C. A word cloud visualization for the top 25 neighbors for the token ‘pandemic’ each year. This visualization highlights each neighbor from a particular year and allows for the comparison between two years. Tokens in purple are shared within both years, while tokens in red or blue are unique to their respective year.

We constructed an online application that allows users to examine how tokens change through time. The application supports token input as text strings or as MeSH IDs, Entrez Gene IDs, and Taxonomy IDs. Users might elect to explore the term ‘pandemic’, for which we detected a change point between 2019 and 2020. Users can examine the token’s nearest neighbors through time (Figure 12A). For example, for ‘pandemic’ users can observe that the token ‘epidemic’ remains similar through time, but taxid:114727 (the H1N1 subtype of influenza) only enters the nearest neighbors with the swine flu pandemic in 2009 and that MeSH:C000657245 (COVID-19) appears

in 2020. The application also shows a frequency chart depicting how often the particular token is used each year (Figure 12B), which can be displayed as a raw count or adjusted by the total size of the corpus. When change points are detected, they are indicated on this panel (Figure 12B). The final visualization shows the union of the nearest 25 neighbors from each year ordered by the number of years that neighbor was present (Figure 12C). This visualization has a comparison function that allows users to examine differences between years. All functionalities are fully supported across the PMCOA and preprint corpora, and users can toggle between the two.

Discussion and Conclusion

Language is rapidly evolving, and the usage of words changes over time. These sorts of changes result in words assimilating new meanings or associations. A modest amount of effort has studied this trend in biomedical text, We implemented an analysis to observe how the usage of tokens changes over time using open-access biomedical corpora.

We validated that direct comparison needs a correction step such as Orthogonal Procrustes. However, even with alignment, systematic differences hidden within these corpora result in variation that needs to be corrected. We constructed a novel statistic that took the ratio of the average inter-year distance over the sum of the intra-year distances. This ratio corrected the latent variation without obstructing our ability to detect tokens that were expected to have a change point.

We perform a changepoint detection using the CUSUM algorithm to identify tokens of interest. We found tokens such as ‘cas9’, ‘pandemic’, and ‘sars’ to appear in our candidate list. These tokens were expected to appear as their changes were prominently known within the field [203,204,205,206,207]. Furthermore, we noticed many changepoints that overlapped between PMCOA and preprints were related to COVID-19. Despite our efforts, many of our detected changepoints are subject to further investigation due to the reliance on manual curation for validation. An open extension to this work would be the development of semi-automatic ways to

determine the validity of a changepoint. In addition to validation, future work could apply a similar approach to other preprint repositories such as arXiv [3] or psyArXiv [208]. Lastly, we created a website that enables a closer examination of individual tokens as they change through time.

CHAPTER 4

Expanding a Database-derived Biomedical Knowledge Graph via Multi-relation Extraction from Biomedical Abstracts

This chapter appeared as a preprint in bioRxiv with the following citation: Nicholson DN, Himmelstein DS and Greene CS Expanding a Database-derived Biomedical Knowledge Graph via Multi-relation Extraction from Biomedical Abstracts (2020) DOI:10.1101/730085.

This paper is a co-authored paper where the majority of work was performed by Nicholson DN who was advised by Greene CS and Himmelstein DS.

Introduction

Knowledge bases are essential resources that hold complex structured and unstructured information. These resources have been used to construct networks for drug repurposing discovery [209,210,211] or as a source of training labels for text mining systems [90,94,212]. Populating knowledge bases often requires highly trained scientists to read biomedical literature and summarize the results through manual curation [213]. In 2007, researchers estimated that filling a knowledge base via manual curation would require approximately 8.4 years to complete [214]. As the rate of publications increases exponentially [215], using only manual curation to populate a knowledge base has become nearly impractical.

Relationship extraction is one of several solutions to the challenge posed by an exponentially growing body of literature [213]. This process creates an expert system to automatically scan, detect, and extract relationships from textual sources. These expert systems fall into three types: unsupervised, rule-based, and supervised systems.

Unsupervised systems extract relationships without the need for annotated text. These approaches utilize linguistic patterns such as the frequency of two entities appearing in a sentence together more often than chance, commonly referred to as co-occurrence

[57,59,63,64,216,217,218,219,220]. For example, a possible system would say gene X is associated with disease Y because gene X and disease Y appear together more often than chance [57]. Besides frequency, other systems can utilize grammatical structure to identify relationships [61]. This information is modeled in the form of a tree data structure, termed a dependency tree. Dependency trees depict words as nodes, and edges represent a word's grammatical relationship with one another. Through clustering on these generated trees, one can identify patterns that indicate a biomedical relationship [61]. Unsupervised systems are desirable since they do not require well-annotated training data; however, precision may be limited compared to supervised machine learning systems.

Rule-based systems rely heavily on expert knowledge to perform relationship extraction. These systems use linguistic rules and heuristics to identify critical sentences or phrases that suggest the presence of a biomedical relationship [48,52,53,221,222,223]. For example, a hypothetical extractor focused on protein phosphorylation events would identify sentences containing the phrase "gene X phosphorylates gene Y" [53]. These approaches provide exact results, but the quantity of positive results remains modest as sentences consistently change in form and structure. For this project, we constructed our label functions without the aid of these works; however, the approaches mentioned in this section provide substantial inspiration for novel label functions in future endeavors.

Supervised systems depend on machine learning classifiers to predict the existence of a relationship using biomedical text as input. These classifiers can range from linear methods such as support vector machines [40,72] to deep learning [224,225,226,227,228,229], which all require access to well-annotated datasets. Typically, these datasets are usually constructed via manual curation by individual scientists [69,73,106,107,110] or through community-based efforts [68,230,231]. Often, these datasets are well annotated but are modest in size, making model training hard as these algorithms become increasingly complex.

Distant supervision is a paradigm that quickly sidesteps manual curation to generate large training datasets. This technique assumes that positive examples have been previously established in selected databases, implying that the corresponding sentences or data points are also positive [94]. The central problem with this technique is that generated labels are often of low quality, resulting in many false positives [232]. Despite this caveat there have been notable effort using this technique [62,99,233].

Data programming is one proposed solution to amend the false positive problem in distant supervision. This strategy combines labels obtained from distant supervision with simple rules and heuristics written as small programs called label functions [234]. These outputs are consolidated via a noise-aware model to produce training labels for large datasets. Using this paradigm can dramatically reduce the time required to obtain sufficient training data; however, writing a helpful label function requires substantial time and error analysis. This dependency makes constructing a knowledge base with a myriad of heterogenous relationships nearly impossible as tens or hundreds of label functions are necessary per relationship type.

This paper seeks to accelerate the label function creation process by measuring how label functions can be reused across different relationship types. We hypothesized that sentences describing one relationship type might share linguistic features such as keywords or sentence structure with sentences describing other relationship types. If this hypothesis were to, one could drastically reduce the time needed to build a relation extractor system and swiftly populate large databases like Hetionet v1. We conducted a series of experiments to estimate how label function reuse enhances performance over distant supervision alone. We focused on relationships that indicated similar types of physical interactions (i.e., Gene-binds-Gene and Compound-binds-Gene) and two more distinct types (i.e., Disease-associates-Gene and Compound-treats-Disease).

Methods and Materials

Hetionet

Hetionet v1 [211] is a heterogeneous network that contains pharmacological and biological information. This network depicts information in the form of nodes and edges of different types. Nodes in this network represent biological and pharmacological entities, while edges represent relationships between entities. Hetionet v1 contains 47,031 nodes with 11 different data types and 2,250,197 edges that represent 24 different relationship types (Figure 13). Edges in Hetionet v1 were obtained from open databases, such as the GWAS Catalog [235], Human Interaction database [236] and DrugBank [237]. For this project, we analyzed performance over a subset of the Hetionet v1 edge types: disease associates with a gene (DaG), compound binds to a gene (CbG), compound treating a disease (CtD), and gene interacts with gene (GiG) (bolded in Figure 13).

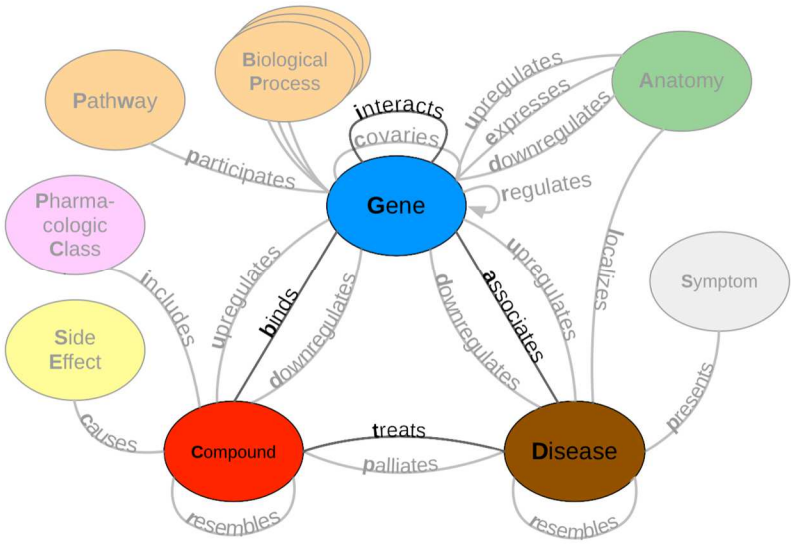


Figure 13 Metagraph of Hetionet

A metagraph (schema) of Hetionet v1 where biomedical entities are represented as nodes and the relationships between them are represented as edges. We examined performance on the highlighted subgraph; however, the long-term vision is to capture edges for the entire graph.

Dataset

We used PubTator Central [128] as input to our analysis. PubTator Central provides MEDLINE abstracts that have been annotated with well-established entity recognition tools including Tagger One [188] for disease, chemical and cell line entities, tmVar [191] for genetic variation tagging, GNormPlus [189] for gene entities and SR4GN [190] for species entities. We downloaded PubTator Central on March 1, 2020, at which point it contained approximately 30,000,000 documents. After downloading, we filtered out annotated entities that were not contained in Hetionet v1. We extracted sentences with two or more annotations and termed these sentences as candidate sentences. We used the Spacy's English natural language processing (NLP) pipeline (en_core_web_sm) [56] to generate dependency trees and parts of speech tags for every extracted candidate sentence. Each candidate sentence was stratified by their corresponding abstract ID to produce a training set, tuning set, and a testing set. We used random assortment to assign dataset labels to each abstract. Every abstract had a 70% chance of being labeled training, 20% chance of being labeled tuning, and 10% chance of being labeled testing. Despite the power of data programming, all text mining systems need to have ground truth labels to be well-calibrated. We hand-labeled five hundred to a thousand candidate sentences of each edge type to obtain a ground truth set (Table 10).

Table 10 Statistics of Candidate Sentences.

We sorted each abstract into a training, tuning and testing set. Numbers in parentheses show the number of positives and negatives that resulted from the hand-labeling process.

Relationship	Train	Tune	Test
Disease-associates-Gene (DaG)	2.49 M	696K (397+, 603-)	348K (351+, 649-)
Compound-binds-Gene (CbG)	2.4M	684K (37+, 463-)	341k (31+, 469-)
Compound-treats-Disease (CtD)	1.5M	441K (96+, 404-)	223K (112+, 388-)
Gene-interacts-Gene (GiG)	11.2M	2.19M (60+, 440-)	1.62M (76+, 424-)

Label Functions for Annotating Sentences

The challenge of having too few ground truth annotations is familiar to many natural language processing applications, even when unannotated text is abundant. Data programming circumvents this issue by quickly annotating large datasets using multiple noisy signals emitted by label functions [234]. Label functions are simple pythonic functions that emit: a positive label (1), a negative label (0), or abstain from emitting a label (-1). These functions can use different approaches or techniques to emit a label; however, these functions can be grouped into simple categories discussed below. Once constructed, these functions are combined using a generative model to output a single annotation. This single annotation is a consensus probability score bounded between 0 (low chance of mentioning a relationship) and 1 (high chance of mentioning a relationship). We used these annotations to train a discriminative model for the final classification step.

Label Function Categories

Label functions can be constructed in various ways; however, they also share similar characteristics. We grouped functions into databases and text patterns. The majority of our label functions fall into the text pattern category (Supplemental Table 11). Further, we described each label function category and provided an example that refers to the following candidate sentence: “PTK6 may be a novel therapeutic target for pancreatic cancer”.

Databases: These label functions incorporate existing databases to generate a signal, as seen in distant supervision [94]. These functions detect if a candidate sentence’s co-mention pair is present in a given database. Our label function emits a positive label if the pair is present and abstains otherwise. If the pair is not present in any existing database, a separate label function emits a negative label. We used a separate label function to prevent a label imbalance problem, which can occur when a single function labels every possible sentence despite being correct or not. If this problem isn’t handled correctly, the generative model could become biased and only emit one prediction (solely positive or solely negative) for every sentence.

$$\Lambda_{\{DB\}}(D, G) = \begin{cases} 1, & (D, G) \in DB \\ 0, & otherwise \end{cases}$$

$$\Lambda_{\{\neg DB\}}(D, G) = \begin{cases} -1, & (D, G) \notin DB \\ 0, & otherwise \end{cases}$$

Text Patterns: These label functions are designed to use keywords or sentence context to generate a signal. For example, a label function could focus on the number of words between two mentions and emit a label if two mentions are too close. Alternatively, a label function could focus on the parts of speech contained within a sentence and ensures a verb is present. Besides parts of speech, a label function could exploit dependency parse trees to emit a label. These trees are akin to the tree data structure where words are nodes and edges are how each word modifies each other. Label functions that use these parse trees will test if the generated tree matches a pattern and emits a positive label if true. For our analysis, we used previously identified patterns designed for biomedical text to generate our label functions [61].

$$\Lambda_{\{TP\}}(D, G) = \begin{cases} 1, & \text{"target"} \in Candidate\ Sentence \\ -1, & otherwise \end{cases}$$

$$\Lambda_{\{TP\}}(D, G) = \begin{cases} 1, & \text{"VB"} \notin pos(Candidate\ Sentence) \\ -1, & otherwise \end{cases}$$

$$\Lambda_{\{TP\}}(D, G) = \begin{cases} 1, & dep(Candidate\ Sentence) \in Cluster\ Theme \\ -1, & otherwise \end{cases}$$

Each text pattern label function was constructed via manual examination of sentences within the training set. For example, using the candidate sentence above, one would identify the phrase “novel therapeutic target” and incorporate this phrase into a global list that a label function would use to check if present in a sentence. After initial construction, we tested and augmented the label function using sentences in the tune set. We repeated this process for every label function in our repertoire.

Table 11 The distribution of each label function per relationship.

Relationship	Databases (DB)	Text Patterns (TP)
DaG	7	30
CtD	3	22
CbG	9	20
GiG	9	28

Training Models

Generative Model

The generative model is a core part of this automatic annotation framework. It integrates multiple signals emitted by label functions to assign each candidate sentence the most appropriate training class. This model takes as input a label function output in the form of a matrix where rows represent candidate sentences, and columns represent each label function (Λ^{nxm}). Once constructed, this model treats the true training class (Y) as a latent variable and assumes that each label function is independent of one another. Under these two assumptions, the model finds the optimal parameters by minimizing a loglikelihood function marginalized over the latent training class.

$$\hat{\theta} = argmin_{\theta} \sum_{Y} -log(P_{\theta}(\Lambda, Y))$$

Following optimization, the model emits a probability estimate that each sentence belongs to the positive training class. At this step, each probability estimate can be discretized via a chosen threshold into a positive or negative class. We used a threshold of 0.5 for discretizing our training classes within our analysis. For more information on how the likelihood function is constructed and minimized, refer to [238].

Discriminative Model

The discriminative model is the final step in this framework. This model uses training labels generated from the generative model combined with sentence features to classify the presence of a biomedical relationship. Typically, the discriminative model is a neural network. We used BioBERT [227], a BERT [239] model trained on all papers and abstracts within Pubmed Central [135], as our discriminative model. BioBERT provides its own set of word embeddings, dense vectors representing words that models such as neural networks can use to construct sentence features. We downloaded a pre-trained version of this model using huggingface's transformer python package [240] and fine-tuned it using our generated training labels. Our fine-tuning approach involved freezing all downstream layers except for the classification head of this model. Next, we trained this model for 10 epochs using the Adam optimizer [241] with huggingface's default parameter settings and a learning rate of 0.001.

Experimental Design

Reusing label functions across edge types would substantially reduce the number of label functions required to extract multiple relationships from biomedical literature. We first established a baseline by training a generative model using only distant supervision label functions designed for the target edge type (see Supplemental Methods). Then we compared the baseline model with models that incorporated a set number of text pattern label functions. Using a sampling with replacement approach, we sampled these text pattern label functions from three different groups: within edge types, across edge types, and from a pool of all label functions. We compared within-edge-type performance to across-edge-type and all-edge-type performance. We sampled a fixed number of label functions for each edge type consisting of five evenly spaced numbers between one and the total number of possible label functions. We repeated this sampling process 50 times for each point. Furthermore, we also trained the discriminative model using annotations from the generative model trained on edge-specific label functions at each point. We report the performance of both models in terms of the area under the receiver operating characteristic curve

(AUROC) and the area under the precision-recall curve (AUPR). Ensuing model evaluations, we quantified the number of edges we could incorporate into Hetionet v1. We used our best performing discriminative model to score every candidate sentence within our dataset and grouped candidates based on their mention pair. We took the max score within each candidate group, and this score represents the probability of the existence of an edge. We established edges using a cutoff score that produced an equal error rate between the false positives and false negatives. Lastly, we report the number of preexisting edges we could recall and the number of novel edges we can incorporate.

Results

Generative Model Using Randomly Sampled Label Functions

Creating label functions is a labor-intensive process that can take days to accomplish. We sought to accelerate this process by measuring how well label functions can be reused. We evaluated this by performing an experiment where label functions are sampled on an individual (edge vs. edge) level and a global (collective pool of sources) level. We observed that performance increased when edge-specific label functions were added to an edge-specific baseline model, while label function reuse usually provided less benefit (AUROC Figure 14, AUPR Supplemental Figure 23). The quintessential example of this overarching trend is the Compound-treats-Disease (CtD) edge type, where edge-specific label functions consistently outperformed transferred label functions. However, there is evidence that label function transferability may be feasible for selected edge types and label function sources. Performance increases as more Gene-interacts-Gene (GiG) label functions are incorporated into the Compound-binds-Gene (CbG) baseline model and vice versa. This trend suggests that sentences for GiG and CbG may share similar linguistic features or terminology that allows for label functions to be reused, which could relate to both describing physical interaction relationships. Perplexingly, edge-specific Disease-associates-

Gene (DaG) label functions did not improve performance over label functions drawn from other edge types. Overall, only CbG and GiG showed significant signs of reusability. This pattern suggests that label function transferability may be possible for these two edge types.

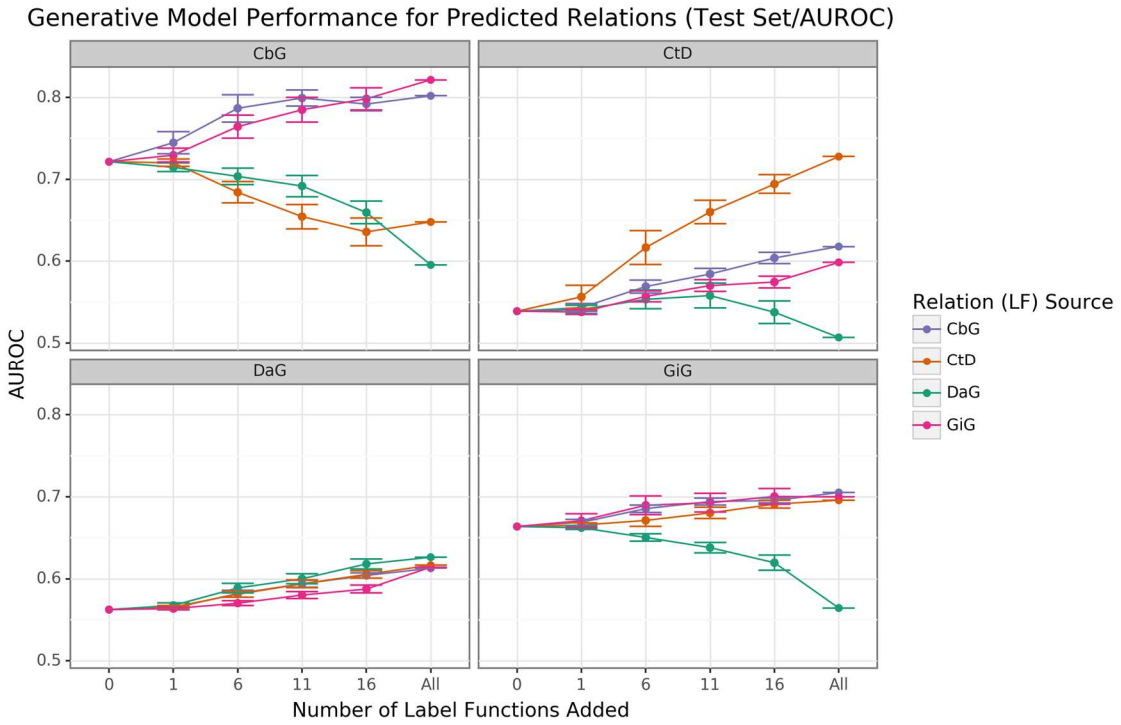


Figure 14 Generative Model Performance for Predicted Relations AUROC

Edge-specific label functions perform better than edge-mismatch label functions, but certain mismatch situations show signs of successful transfer. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound-treats-Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the receiver operating curve (AUROC). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

We found that sampling from all label function sources at once usually underperformed relative to edge-specific label functions (Figure 15 and Supplemental Figure 24). The gap between edge-specific sources and all sources widened as we sampled more label functions. CbG is a prime example of this trend (Figure 15 and Supplemental Figure 24), while CtD and GiG show a similar

but milder trend. DaG was the exception to the general rule. The pooled set of label functions improved performance over the edge-specific ones, which aligns with the previously observed results for individual edge types (Figure 14). When pooling all label functions, the decreasing trend supports the notion that label functions cannot simply transfer between edge types (exception being CbG on GiG and vice versa).

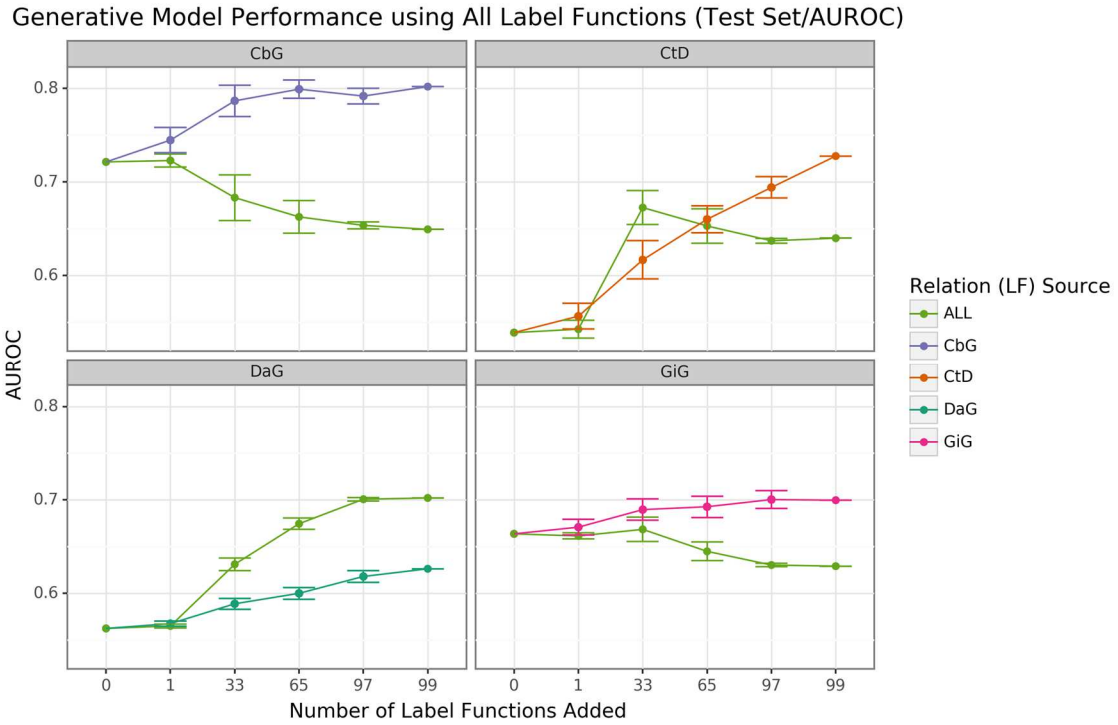


Figure 15 Generative Model Performance using All Label Functions

Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound-treats-Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the receiver operating curve (AUROC). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Discriminative Model Performance

The discriminative model is intended to augment performance over the generative model by incorporating textual features together with estimated training labels. We found that the discriminative model generally outperformed the generative model with respect to AUROC as more edge-specific label functions were incorporated (Figure 16). Regarding AUPR, this model outperformed the generative model for the DaG edge type. At the same time, it had close to par performance for the rest of the edge types (Supplemental Figure 25). The discriminative model's performance was often poorest when very few edge-specific label functions were incorporated into the baseline model (seen in DaG, CbG, and GiG). This example suggests that training generative models with more label functions produces better outputs for training for discriminative models. CtD was an exception to this trend, where the discriminative model outperformed the generative model at all sampling levels in regards to AUROC. We observed the opposite trend with the CbG edges as the discriminative model was always worse or indistinguishable from the generative model. Interestingly, the AUPR for CbG plateaus below the generative model and decreases when all edge-specific label functions are used (Supplemental Figure 25). This trend suggests that the discriminative model might have predicted more false positives in this setting. Overall, incorporating more edge-specific label functions usually improved performance for the discriminative model over the generative model.

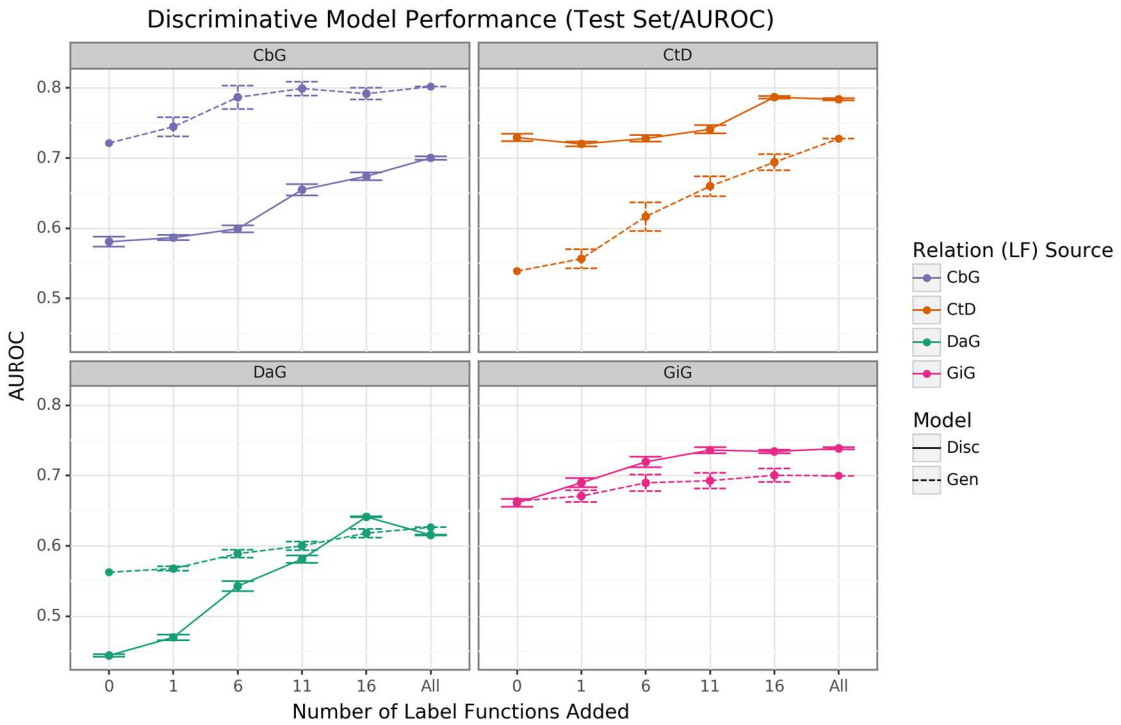


Figure 16 Discriminative Model Performance AUROC

The discriminative model usually improves faster than the generative model as more edge-specific label functions are included. The line plot headers represent the specific edge type the discriminative model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the baseline model (the point at 0). The y axis shows the area under the receiver operating curve (AUROC). Each data point represents the average of 3 sample runs for the discriminator model and 50 sample runs for the generative model. The error bars represent each run's 95% confidence interval. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Text Mined Edges Can Expand a Database-derived Knowledge Graph

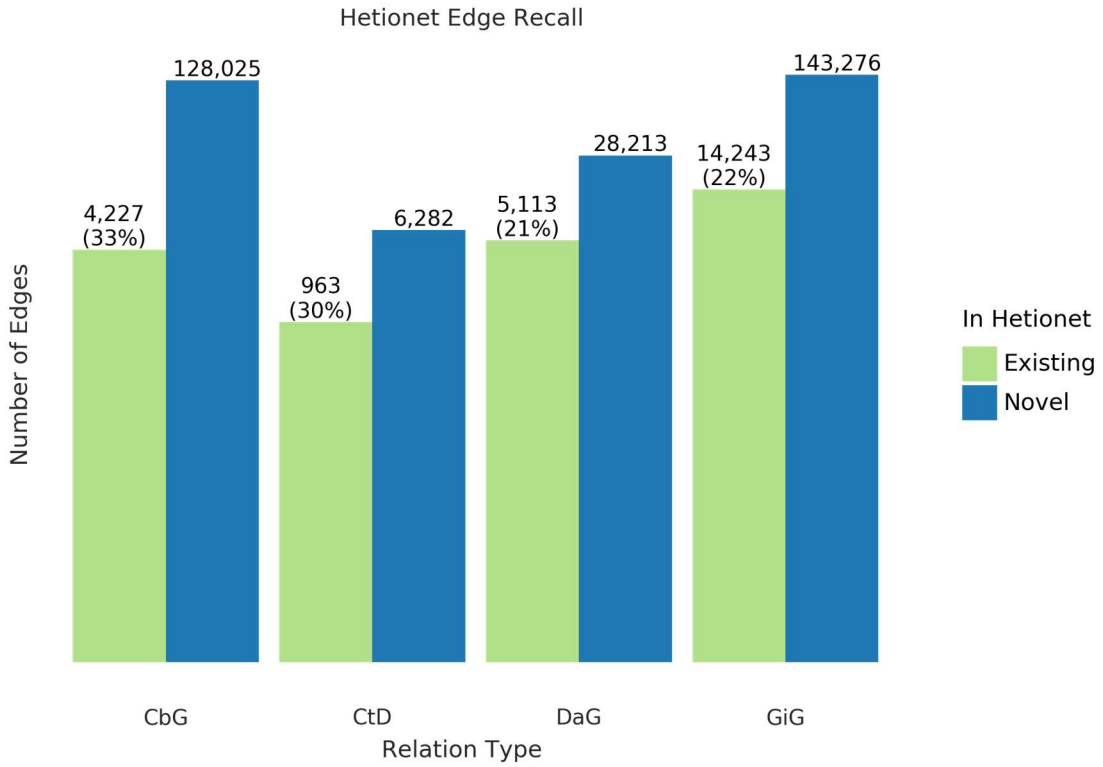


Figure 17 Edge Recall for Hetionet

Text-mined edges recreate a substantial fraction of an existing knowledge graph and include new predictions. This bar chart shows the number of edges we can successfully recall in green and indicates the number of new edges in blue.

The recall for the Hetionet v1 knowledge graph is shown as a percentage in parentheses. For example, for the Compound-treats-Disease (CtD) edge, our method recalls 30% of existing edges and can add 6,282 new ones.

One of the goals of our work is to measure the extent to which learning multiple edge types could construct a biomedical knowledge graph. Using Hetionet v1 as an evaluation set, we measured this framework's recall and quantified the number of edges that may be incorporated with high confidence. Overall, we were able to recall about thirty percent of the preexisting edges for all edge types (Figure 17) and report our top ten scoring sentences for each edge type in

Supplemental Table 14. Our best recall was with the CbG edge type, where we retained 33% of preexisting edges. In contrast, we only recalled close to 30% for CtD, while the other two categories achieved a recall score close to 22%. Despite the modest recall level, the amount of novel edge types remains elevated. This notion highlights that Hetionet v1 is missing a compelling amount of biomedical information, and relationship extraction is a viable way to close the information gap.

Discussion and Conclusions

Filling out knowledge bases via manual curation can be an arduous and erroneous task [213]. Using manual curation alone becomes impractical as the rate of publications continuously increases. Data programming is a paradigm that uses label functions to speed up the annotation process and can be used to solve this problem. However, creating useful label functions is an obstacle to this paradigm, which takes considerable time. We tested the feasibility of re-using label functions to reduce the number of label functions required for strong prediction performance. Our sampling experiment revealed that adding edge-specific label functions is better than adding off-edge label functions. An exception to this trend is using label functions designed from conceptually related edge types (using GiG label functions to predict CbG sentences and vice versa). Furthermore, broad edge types such as DaG did not follow this trend as we found this edge to be agnostic to any tested label function source. One possibility for this observation is that the “associates” relationship is a general concept that may include other concepts such as Disease (up/down) regulating a Gene (examples highlighted in our annotated sentences). The discriminator model did not have an apparent positive or negative effect on performance; however, we noticed that performance heavily depended on the annotations provided by the generative model. This pattern suggests a focus on label function construction and generative model training may be key steps to focus on in future work. Although we found that label

functions cannot be re-used across all edge types with the standard task framing, strategies like multitask [101] or transfer learning [97] may make multi-label-function efforts more successful.

CHAPTER 5

Written communication is a fundamental part of the life science community as it enables the widespread sharing of research findings. Through textual analysis, we can attain a higher understanding of life science research. This thesis was centered on performing multiple textual analyses using published and pre-published papers to better grasp how language changes within the field.

Chapter 2 concentrated on analyzing preprints and exploring how their textual content changed when subjected to the peer-review process. We found that most changes between preprints and their published counterparts were mainly stylistic. This trend suggests that output from the peer-review process is modest text changes at best, which has also been reinforced by other studies [21,241]. We found that most preprints are eventually published, which had been confirmed by previous endeavors [118], Furthermore, we established a new lower bound on the number of preprints published. However, the true proportion of published preprints remains to be seen as there were missing links within bioRxiv and many published papers were behind paywalls. As published papers become more available through open access efforts, it will be interesting to see an updated version of published preprints. Overall, preprints are being increasingly integrated into the life science community and might become valuable resources for other avenues for textual analysis, such as text mining.

Chapter 3 examined how the meaning and associations of words change over time within biomedical preprints and published. These types of changes are called semantic shifts, and we took a novel approach to model these changes. We confirmed that Word2Vec models need a correction step to enable model comparison. Despite the correction, we took a multi-model approach to account for residual variation after alignment. We performed changepoint detection and found over 43,000 different candidates that may have changed their meaning. In our candidate list, we found tokens such as “pandemic”, “sars” and “cas9” which are known positive

results [242,243]. Despite this confirmation, most of our change point list remains for future investigation as this process heavily relies on manual curation and expert knowledge. An extension to this chapter would be to explore intuitive ways to validate these findings. For example, one approach would be connecting published papers to these potential token candidates. Also, as time progresses, it will be interesting to see which tokens gain a change point.

Chapter 4 explored the paradigm of weak supervision and measured the extent to which label sources could be re-used across Hetionet edge types. We used four different relationship types, Compound-binds-Gene (CbG), Gene-interacts-Gene (GiG), Disease-associates-Gene (DaG), and Compound-treats-Disease (CtD). We found that label sources didn't transfer well across our selected relationship types, suggesting that the language used to describe each edge type is distinct. An exception to this trend was Compound-binds-Gene (CbG) and Gene-interacts-Gene (GiG). There was noticeable transferability, suggesting that scientists use similar language to describe both edge types. We also found that the discriminator model didn't significantly impact prediction performance, suggesting that most endeavors would prosper from focusing on refining the generative model's annotations. Furthermore, future endeavors could prosper more by focusing on mining one relationship at a time. Conversely, other endeavors could use techniques such as multi-task, transfer, or semi-supervised learning.

Overall, the effort performed in this thesis is just the beginning of textual analysis as a whole. The main contributions were using preprints and published papers to assist the life science community in ascertaining the language and research trends contained in these resources. Moving forward, it will be exciting to see what extensions will arise from this work.

APPENDIX A

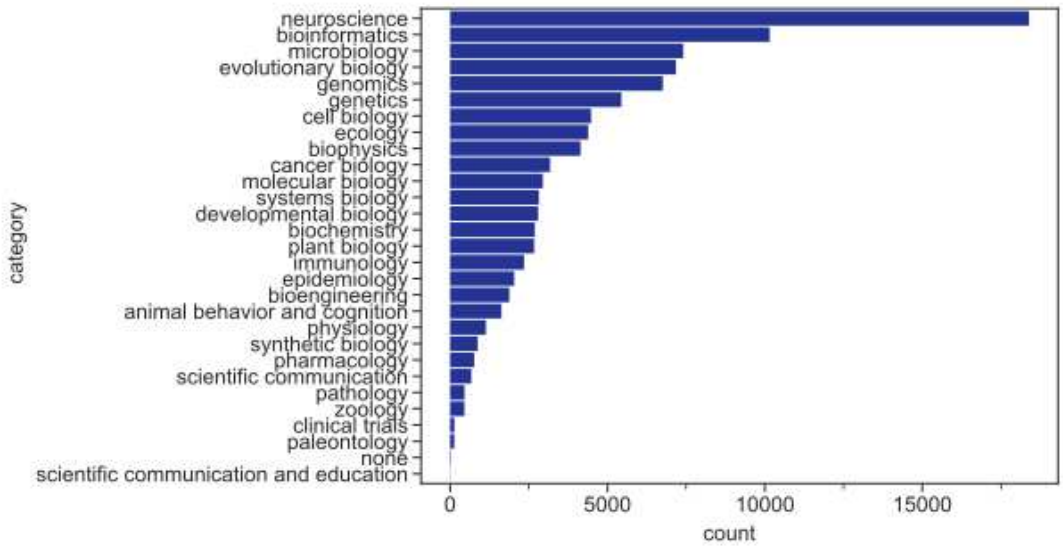


Figure 18 Document category count for bioRxiv

Neuroscience and bioinformatics are the two most common author-selected topics for bioRxiv preprints. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s2.

Table 12 The top 100 frequently occurring tokens across our three corpora

bioRxiv Tokens	PMCOA Tokens	NYTAC Tokens
‘et’	‘\\u2009’	‘said’
‘al’	‘\xa0’	‘mr.’
‘cell’	‘\t\t\t\t’	‘ ’
‘cells’	‘et’	‘–’
‘1’	‘1’	‘new’
‘different’	‘cells’	‘new’
‘2’	‘al’	‘like’
‘high’	‘cell’	‘year’
‘genes’	‘patients’	‘years’
‘gene’	‘study’	‘united’

‘3’	‘2’	‘ms.’
‘specific’	‘different’	‘today’
‘figure’	‘high’	‘york’
‘single’	‘3’	‘old’
‘non’	‘\\u2013’	‘american’
‘5’	‘significant’	‘yesterday’
‘\\u201d’	‘10’	‘time’
‘\\u201c’	‘5’	‘lead’
‘data’	‘significantly’	‘people’
‘10’	‘group’	‘dr.’
‘4’	‘4’	‘years’
‘significant’	‘non’	‘york’
‘\\u2019’	‘compared’	‘week’
‘found’	‘\\u201c’	‘officials’
‘protein’	‘\\u201d’	‘ago’
‘model’	‘found’	‘including’
‘performed’	‘performed’	‘10’
‘figure’	‘specific’	‘people’
‘analysis’	‘respectively’	‘high’
‘study’	‘\\u200a’	‘john’
‘genetic’	‘showed’	‘public’
‘significantly’	‘analysis’	‘good’
‘species’	‘including’	‘political’
‘low’	‘low’	‘1’
‘human’	‘higher’	‘said’
‘time’	‘clinical’	‘president’

‘including’	‘results’	‘year’
‘respectively’	‘groups’	‘national’
‘time’	‘shown’	‘second’
‘compared’	‘time’	‘million’
‘previously’	‘\xb0’	‘university’
‘results’	‘total’	‘recent’
‘shown’	‘treatment’	‘small’
‘fig’	‘protein’	‘percent’
‘multiple’	‘additional’	‘2’
‘large’	‘studies’	‘long’
‘similar’	‘genes’	‘far’
‘\\u2013’	‘positive’	‘big’
‘higher’	‘figure’	‘major’
‘expression’	‘cells’	‘later’
‘expression’	‘gene’	‘west’
‘samples’	‘data’	‘great’
‘i.e.’	‘anti’	‘30’
‘fig’	‘previous’	‘little’
‘individual’	‘data’	‘million’
‘\xb0’	‘addition’	‘3’
‘dna’	‘human’	‘mrs.’
‘average’	‘health’	‘states’
‘supplementary’	‘observed’	‘says’
‘previous’	‘according’	‘according’
‘total’	‘single’	‘late’
‘showed’	‘reported’	‘young’

‘data’	‘previously’	‘away’
‘observed’	‘mice’	‘life’
‘functional’	‘20’	‘american’
‘number’	‘\\u2003’	‘month’
‘based’	‘6’	‘large’
‘\\u2018’	‘c’	‘company’
‘small’	‘study’	‘way’
‘cells’	‘control’	‘black’
‘positive’	‘similar’	‘early’
‘conditions’	‘studies’	‘east’
‘20’	‘expression’	‘real’
‘data’	‘data’	‘3’
‘regions’	‘time’	‘11’
‘data’	‘30’	‘state’
‘proteins’	‘fig’	‘20’
‘new’	‘95’	‘world’
‘mice’	‘\\u2019’	‘net’
‘relative’	‘model’	‘j.’
‘addition’	‘levels’	‘street’
‘6’	‘primary’	‘end’
‘neurons’	‘samples’	‘think’
‘studies’	‘large’	‘day’
‘c’	‘small’	‘long’
‘cells’	‘lower’	‘state’
‘100’	‘ ’	‘david’
‘function’	‘increased’	‘best’

‘activity’	‘100’	‘robert’
‘highly’	‘patients’	‘local’
‘experimental’	‘based’	‘city’
‘standard’	‘figure’	‘million’
‘30’	‘blood’	‘5’
‘levels’	‘50’	‘earns’
‘brain’	‘effect’	‘st.’
‘rna’	‘normal’	‘president’
‘models’	‘standard’	‘world’
‘identified’	‘conditions’	‘nearly’
‘binding’	‘level’	‘4’
‘50’	‘important’	‘home’

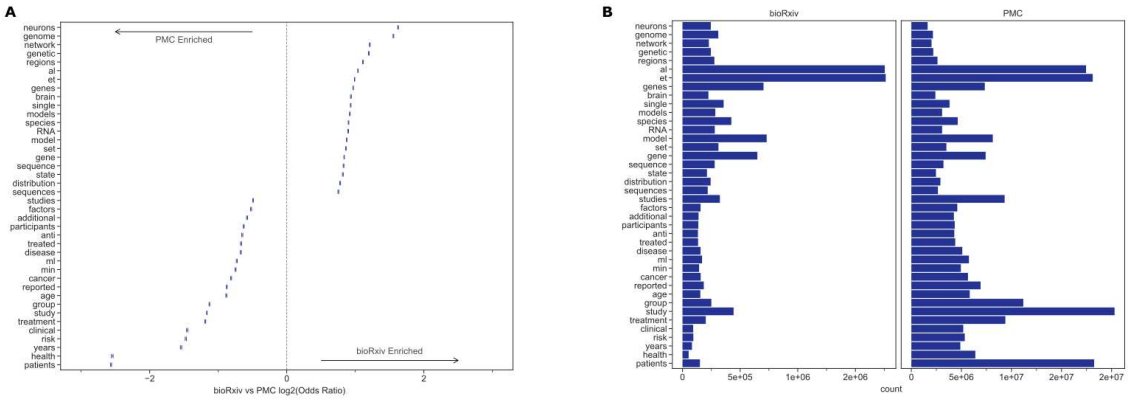


Figure 19 Individual Token Analysis for bioRxiv vs PMCOA Special Characters Removed

A. *The significant differences in token frequencies for the corpora appear to be driven by the fields with the highest uptake of bioRxiv, as terms from neuroscience and genomics are relatively more abundant in bioRxiv. We plotted the 95% confidence interval for each reported token.* ***B.*** *Of the tokens that differ between bioRxiv and PMC, the most abundant in bioRxiv are “gene”, “genes” and “model” while the most abundant in PMC is “study.” Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s3.*

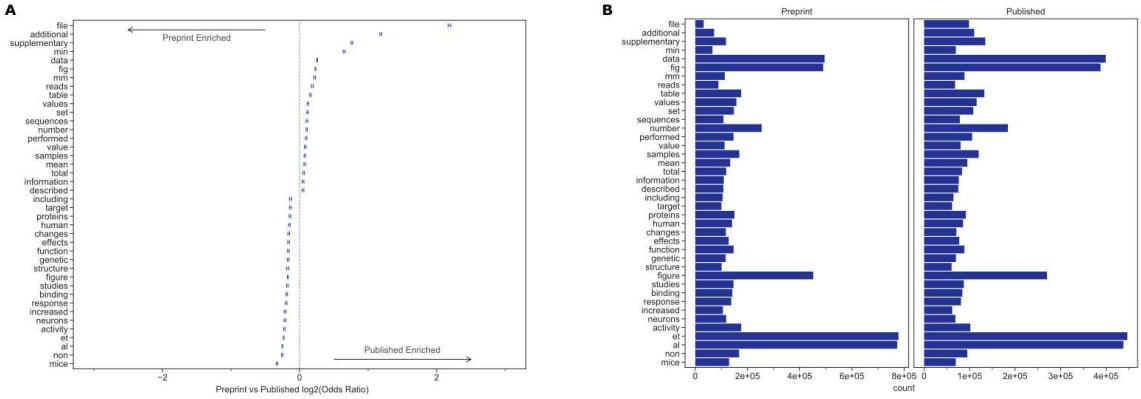


Figure 20 Individual Token Analysis for Preprints vs Their Published Counterparts (Special Characters Removed)

A. *The significant differences in token frequencies for preprints and their corresponding published version often appear to be associated with data availability and supplementary or additional materials. We plotted the 95% confidence interval for each reported token.* ***B.*** *The tokens with the largest absolute differences in abundance appear related to scientific figures and data availability. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s4.*

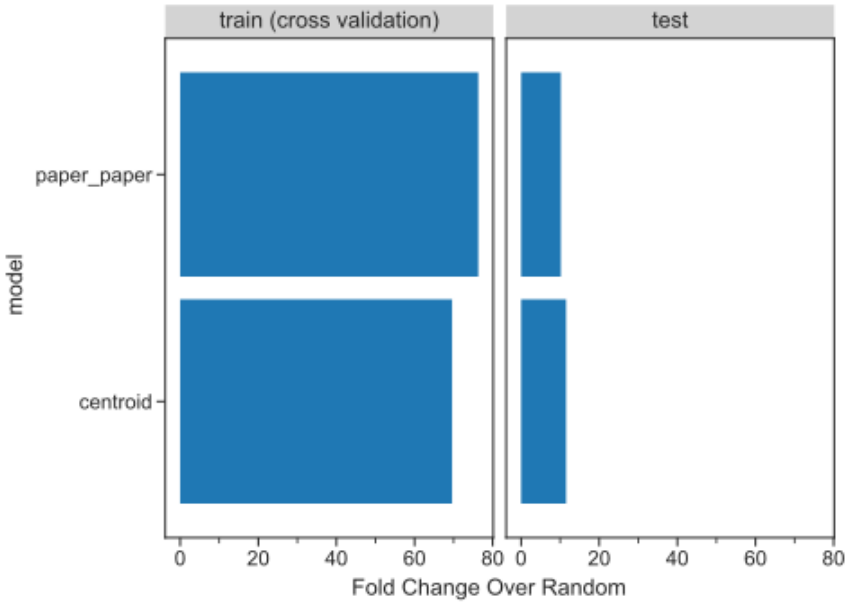


Figure 21 Machine Learning for Predicting Similar Journals

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. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s5.

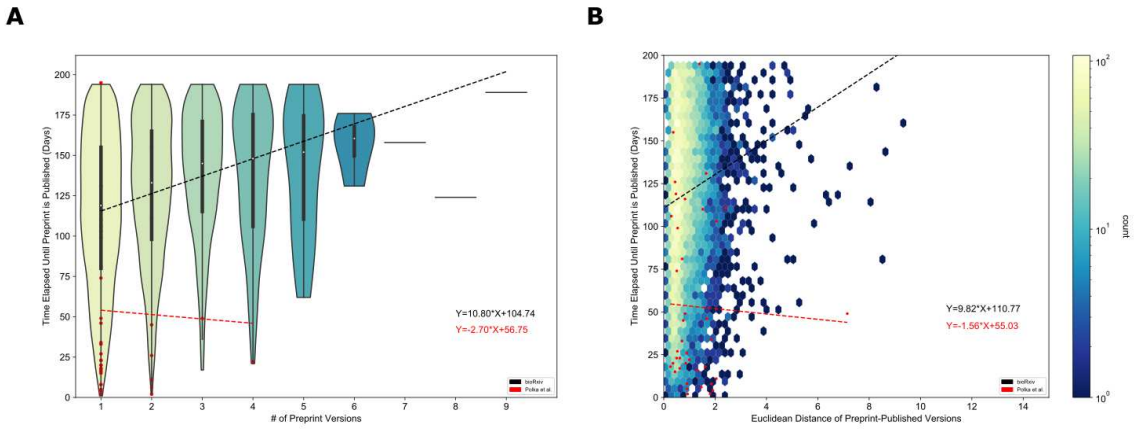


Figure 22 Time analysis for Contextualizing Preprints in Motion

A. *The Preprints in Motion were published faster than other bioRxiv preprints, and the number of versions was not associated with an increase in time to publication. The x-axis shows the number of versions of a preprint posted on bioRxiv. The y-axis indicates the number of days that elapsed between the first version of a preprint posted on bioRxiv and the date at which the peer-reviewed publication appeared. The density of observations is depicted in the violin plot with an embedded boxplot. The red dots and red regression line represent Preprints in Motion.* ***B.*** *The Preprints in Motion collection were published faster than other bioRxiv preprints, and no dependence between the amount of linguistic change and time to publish was observed. The x-axis shows the Euclidean distance between document representations of the first version of a preprint and its peer-reviewed form. The y-axis shows the number of days elapsed between the first version of a preprint posted on bioRxiv and when a preprint is published. The color bar on the right represents the density of each hexbin in this plot, where more dense regions are shown in a brighter color. The red dots and red regression line represent Preprints in Motion. Data for the information depicted in this figure are available at https://github.com/greenelab/annorxiver/blob/master/FIGURE_DATA_SOURCE.md#figure-s6.*

APPENDIX B

Supplemental Tables

Table 13 The intersection of changepoints found between published papers and preprints.

Token	Changepoint
lockdown	2019-2020
2021	2020-2021
distancing	2019-2020
2019	2018-2019
ace2	2019-2020
pandemic	2019-2020
2020	2019-2020
coronavirus	2019-2020
bcl2a1	2018-2019
peak3	2020-2021
3.6.2	2019-2020
quarantine	2019-2020
cobl	2020-2021
injectrode	2020-2021
nrc3	2020-2021
4.0.5	2020-2021
TMPRSS2 (gene_7113)	2019-2020
n262	2019-2020
bin1	2017-2018
n3c	2020-2021
tip1	2020-2021

omicron	2020-2021
pangolin	2019-2020
adrn	2020-2021
seir	2019-2020

APPENDIX C

Supplementary Figures

Generative Model Using Randomly Sampled Label Functions

Individual Sources

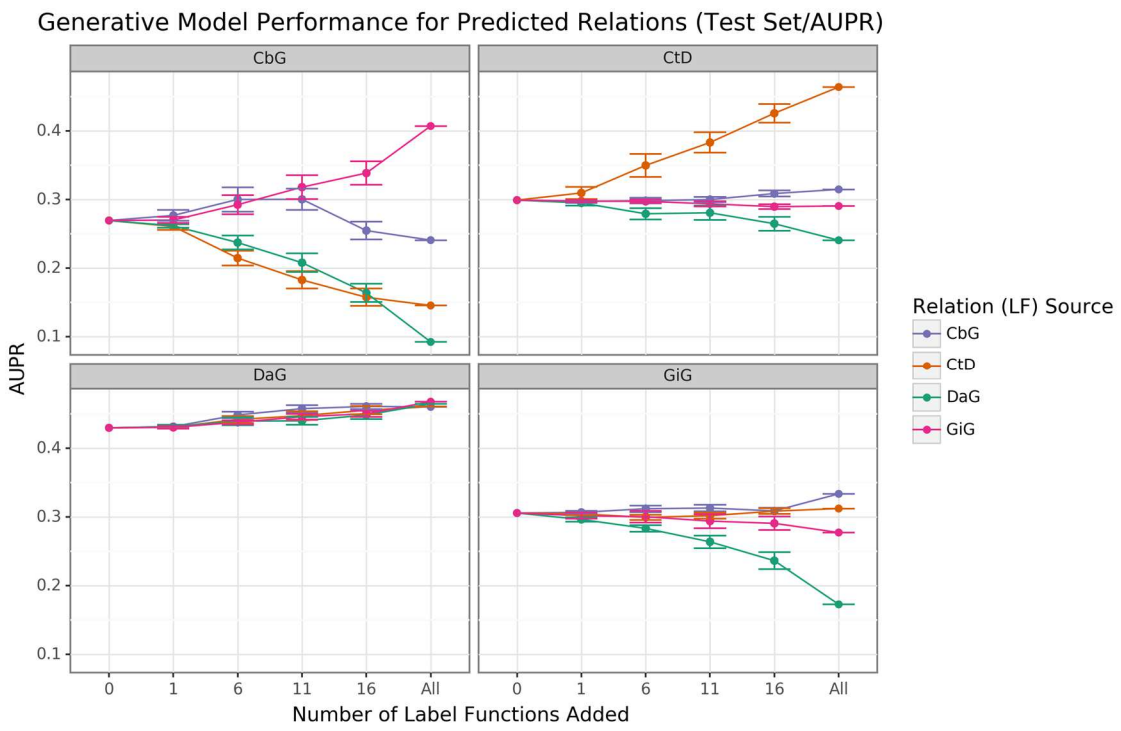


Figure 23 Generative Model Performance for Predicted Relations AUPR

Edge-specific label functions improve performance over edge-mismatch label functions. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the precision-recall curve (AUPR). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Collective Pool of Sources

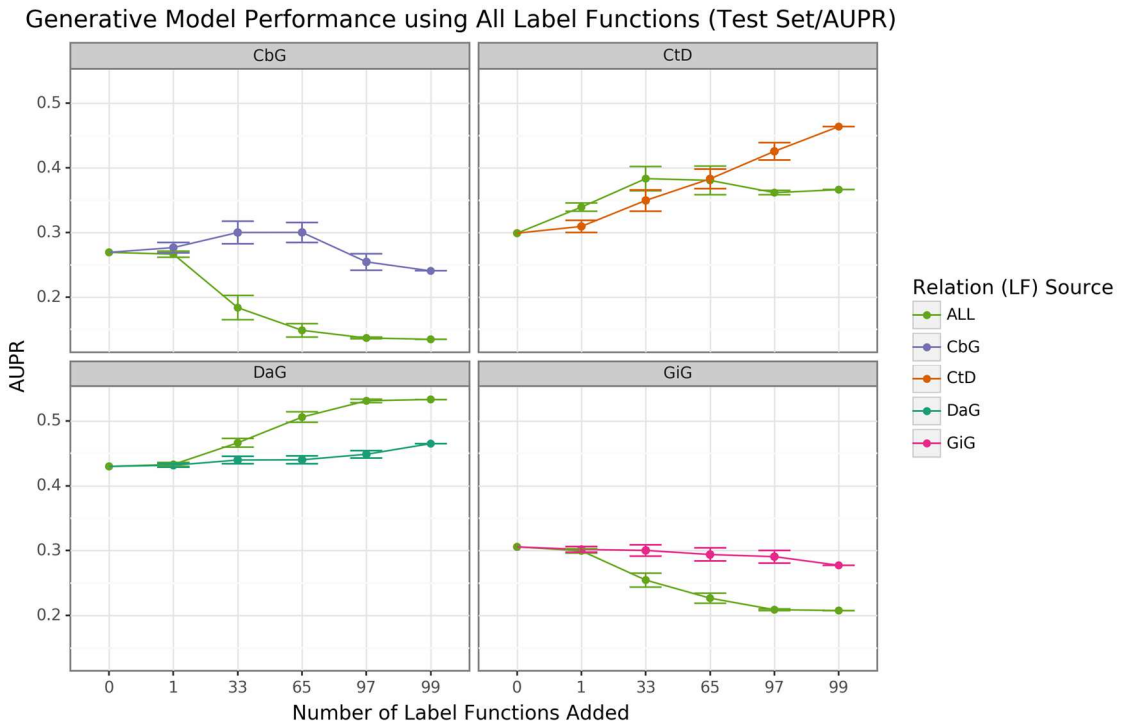


Figure 24 Generative Model Performance using All Label Functions (AUPR)

Using all label functions generally hinders generative model performance. Each line plot header depicts the edge type the generative model is trying to predict, while the colors represent the source of label functions. For example, orange represents sampling label functions designed to predict the Compound treats Disease (CtD) edge type. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the database-only baseline model (the point at 0). The y-axis shows the area under the precision-recall curve (AUPR). Each point on the plot shows the average of 50 sample runs, while the error bars show the 95% confidence intervals of all runs. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Discriminative Model Performance

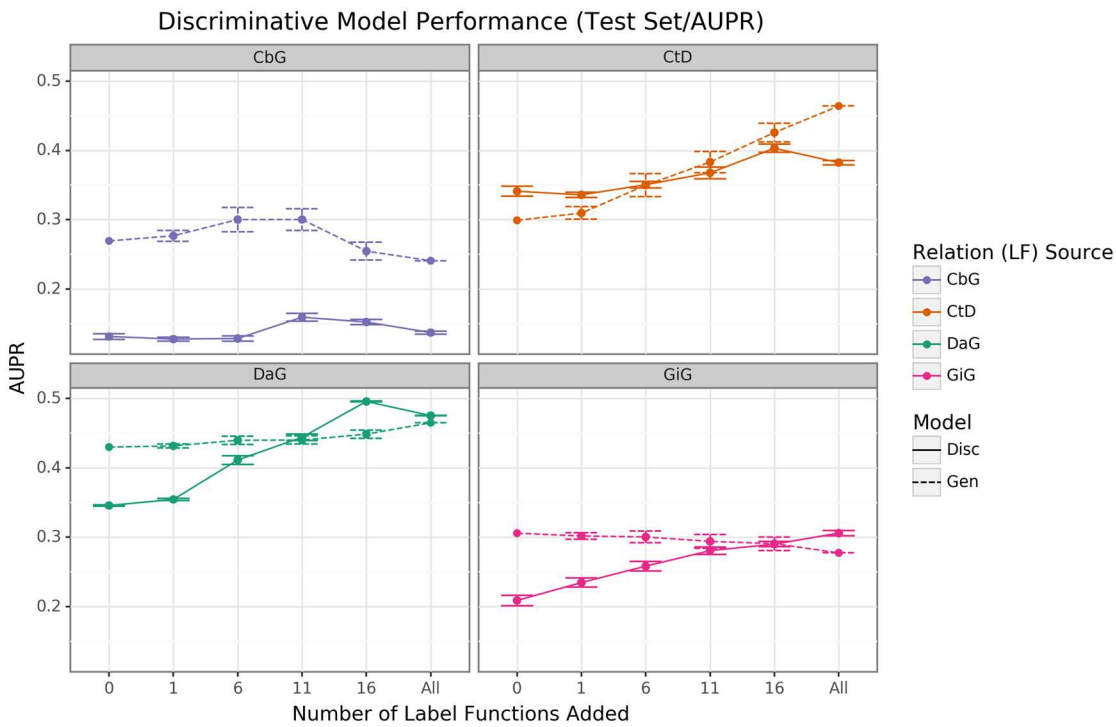


Figure 25 Discriminator Model Performance in AUPR

The discriminator model improves performance as the number of edge-specific label functions is added to the baseline model. The line plot headers represent the specific edge type the discriminator model is trying to predict. The x-axis shows the number of randomly sampled label functions incorporated as an addition to the baseline model (the point at 0). The y axis shows the area under the precision-recall curve (AUPR). Each data point represents the average of 3 sample runs for the discriminator model and 50 sample runs for the generative model. The error bars represent each run's 95% confidence interval. The baseline and "All" data points consist of sampling from the entire fixed set of label functions.

Supplemental Tables

Table 14 Top Ten Sentences for Each Edge Type

Contains the top ten predictions for each edge type. Highlighted words represent entities mentioned within the given sentence.

Edge	Source Node	Target Node	Generative Model	Discriminative Model Prediction	Number of	In Hetionet	Text

Type			Prediction		Sentences		
DaG	hematologic cancer	STMN1	1.000	0.979	83	Novel	the stathmin1 mrna expression level in de novo al patient be high than that in healthy person (p < 0.05) , the stathmin1 mrna expression level in relapse patient with al be high than that in de novo patient (p < 0.05) , and there be no significant difference of stathmin1 mrna

							expression between patient with aml and patient with all .
Da G	breast cancer	INSIG2	1.000	0.979	4	Novel	in analysis of idc cell , the level of insig2 mrna expression be significantly high in late - stage patient than in early - stage patient .
Da G	lung cancer	GNAO1	1.000	0.979	104	Novel	high numb expression be associate with favorable prognosis in patient with lung adenocarcino ma , but not in those with

							squamous cell carcinoma .
Da G	breast cancer	TTF1	1.000	0.977	88	Novel	significant ttf-1 overexpression be observe in adenocarcinomas harbor egfr mutation (p = 0.008) , and no or significantly low level expression of ttf-1 be observe in adenocarcinomas harbor kras mutation (p = 0.000) .
Da G	breast cancer	BUB1B	1.000	0.977	13	Novel	elevated bubr1 expression be associate with poor survival in early stage

							breast cancer patient .
Da G	Alzheimer's disease	SERPIN A3	1.000	0.977	182	Existi ng	a common polymorphism within act and il-1beta gene affect plasma level of act or il-1beta , and ad patient with the act t , t or il-1beta t , t genotype show the high level of plasma act or il-1beta , respectively .
Da G	esophageal cancer	TRAF6	1.000	0.976	15	Novel	expression of traf6 be highly elevated in esophageal cancer tissue , and patient with high traf6 expression

							have a significantly short survival time than those with low traf6 expression .
Da G	hypertension	TBX4	1.000	0.975	146	Novel	the proportion of circulate th1 cell and the level of t - bet , ifng mrna be increase in ht patient , the expression of ifng - as1 be upregulated and positively correlate with the proportion of circulate th1 cell or t - bet , and ifng expression , or serum level of anti -

							thyroglobulin antibody / thyroperoxidase antibody in ht patient .
DaG	breast cancer	TP53	1.000	0.975	3481	Existing	hormone receptor status rather than her2 status be significantly associate with increase ki-67 and p53 expression in triple - negative breast carcinoma , and high expression of ki-67 but not p53 be significantly associate with axillary nodal

							metastasis in triple - negative and high - grade non - triple - negative breast carcinoma .
DaG	esophageal cancer	COL17A1	1.000	0.975	32	Novel	high cd147 expression in patient with esophageal cancer be associate with bad survival outcome and common clinicopathological indicator of poor prognosis .
CtD	Docetaxel	prostate cancer	0.996	0.964	5614	Existing	docetaxel and atrasentan versus docetaxel and placebo for

							man with advanced castration - resistant prostate cancer (swog s0421) : a randomised phase 3 trial
CtD	E7389	breast cancer	0.999	0.957	862	Novel	clinical effect of prior trastuzumab on combination eribulin mesylate plus trastuzumab as first - line treatment for human epidermal growth factor receptor 2 positive locally recurrent or metastatic

							breast cancer : result from a phase ii , single - arm , multicenter study
CtD	Zoledronate	bone cancer	0.996	0.955	226	Novel	zoledronate in combination with chemotherapy and surgery to treat osteosarcoma (os2006) : a randomised , multicentre , open - label , phase 3 trial .
CtD			0.878	0.954	484	Existing	the role of ixazomib as an augment conditioning therapy in salvage autologous stem cell

							transplant (asct) and as a post - asct consolidation and maintenance strategy in patient with relapse multiple myeloma (accord [uk - mra myeloma xii] trial) : study protocol for a phase iii randomise controlled trial
CtD	Topotecan	lung cancer	1.000	0.954	315	Existing	combine chemotherapy with cisplatin , etoposide , and irinotecan versus topotecan alone as

							second - line treatment for patient with sensitive relapse small - cell lung cancer (jcog0605) : a multicentre , open - label , randomised phase 3 trial .
CtD	Epirubicin	breast cancer	0.999	0.953	2147	Existing	accelerate versus standard epirubicin follow by cyclophosphamide , methotrexate , and fluorouracil or capecitabine as adjuvant therapy for breast cancer

							in the randomised uk tact2 trial (cruk/05/19) : a multicentre , phase 3 , open - label , randomise , control trial
CtD	Paclitaxel	breast cancer	1.000	0.952	10255	Existing	sunitinib plus paclitaxel versus bevacizumab plus paclitaxel for first - line treatment of patients with advanced breast cancer : a phase iii , randomized , open - label trial
CtD	Anastrozole	breast cancer	0.996	0.952	2364	Existing	a european organisation for research

							and treatment of cancer randomize , double - blind , placebo - control , multicentre phase ii trial of anastrozole in combination with gefitinib or placebo in hormone receptor - positive advanced breast cancer (nct00066378) .
CtD	Gefitinib	lung cancer	1.000	0.950	11860	Existing	gefitinib versus placebo as maintenance therapy in patient with locally

							advanced or metastatic non - small - cell lung cancer (inform ; c - tong 0804) : a multicentre , double - blind randomise phase 3 trial .
CtD	Docetaxel	prostate cancer	1.000	0.949	5614	Existing	ipilimumab versus placebo after radiotherapy in patient with metastatic castration - resistant prostate cancer that have progress after docetaxel chemotherapy (ca184 - 043) : a multicentre

							, randomised , double - blind , phase 3 trial
CtD	Sulfamethazine	lung cancer	0.611	0.949	4	Novel	tmp / smz (320/1600 mg / day) treatment be compare to placebo in a double - blind , randomized trial in patient with newly diagnose small cell carcinoma of the lung during the initial course of chemotherapy with cyclophosphamide , doxorubicin ,

							and etoposide .
Cb G	D-Tyrosine	EGFR	0.601	0.876	3423	Novel	amphiregulin (ar) and heparin - binding egf - like growth factor (hb - egf) bind and activate the egfr while heregulin (hrg) act through the p185erbb-2 and p180erbb-4 tyrosine kinase .
Cb G	Phosphonotyrosine	ANK3	0.004	0.865	1	Novel	at least two domain of p85 can bind to ank3 , and the interaction involve the p85 c - sh2 domain be

							find to be phosphotyrosine - independent .
CbG	Adenosine	ABCC8	0.891	0.860	353	Novel	sulfonylurea act by inhibition of beta - cell adenosine triphosphate - dependent potassium (k(atp)) channel after bind to the sulfonylurea subunit 1 receptor (sur1) .
CbG	D-Tyrosine	AREG	0.891	0.857	22	Novel	amphiregulin (ar) and heparin - binding egf - like growth factor (hb - egf) bind and

							activate the egfr while heregulin (hrg) act through the p185erbb-2 and p180erbb-4 tyrosine kinase .
Cb G	D-Tyrosine	EGF	0.602	0.856	389	Novel	upon activation of the receptor for the epidermal growth factor (egfr) , sprouty2 undergoe phosphorylatio n at a conserve tyrosine that recruit the src homology 2 domain of c - cbl .

Cb G	D-Tyrosine	CSF1	0.101	0.854	106	Novel	as a member of the subclass iii family of receptor tyrosine kinase , kit be closely relate to the receptor for platelet derive growth factor alpha and beta (pdgf - a and b) , macrophage colony stimulate factor (m - csf) , and flt3 ligand .
Cb G	D-Tyrosine	ERBB4	0.101	0.848	115	Novel	the efgr family be a group of four structurally similar

							tyrosine kinase (egfr , her2 / neu , erbb-3 , and erbb-4) that dimerize on bind with a number of ligand , include egf and transform growth factor alpha .
Cb G	D-Tyrosine	EGFR	0.969	0.848	3423	Novel	the epidermal growth factor receptor be a member of type - -pron- growth factor receptor family with tyrosine kinase activity that be activate follow the binding of

							multiple cognate ligand .
Cb G	D-Tyrosine	VAV1	0.601	0.842	187	Novel	stimulation of quiescent rodent fibroblast with either epidermal or platelet - derive growth factor induce an increase affinity of vav for cbl - b and result in the subsequent formation of a vav - dependent trimeric complex with the ligand - stimulate tyrosine

							kinase receptor .
Cb G	Tretinoin	RORB	0.601	0.840	7	Novel	the retinoid z receptor beta (rzr beta) , an orphan receptor , be a member of the retinoic acid receptor (rar)/thyroid hormone receptor (tr) subfamily of nuclear receptor .
Cb G	L-Tryptophan	TACR1	0.891	0.839	4	Novel	these result suggest that the tryptophan and quinuclidine series of nk-1 antagonist bind to similar bind site on

							the human nk-1 receptor .
GiG	CYSLTR2	CYSLTR2	0.967	0.564	37	Novel	the bind pocket of cyslt2 receptor and the proposition of the interaction mode between cyslt2 and hami3379 be identify .
GiG	RXRA	PPARA	1.000	0.563	143	Novel	after bind ligand , the ppar - y receptor heterodimerize with the rxr receptor .
GiG	RXRA	RXRA	0.824	0.551	1101	Existing	nuclear hormone receptor , for example , bind either as homodimer or

							as heterodimer with retinoid x receptor (rxr) to half - site repeat that be stabilize by protein - protein interaction mediate by residue within both the dna- and ligand - bind domain .
GiG	ADRBK1	ADRA2A	0.822	0.543	3	Novel	mutation of these residue within the holo - alpha(2a)ar diminish grk2-promoted phosphorylation of the receptor as well as the ability of the

							kinase to be activate by receptor binding .
GiG	ESRRA	ESRRA	0.001	0.531	308	Existing	the crystal structure of the ligand bind domain (lbd) of the estrogen - relate receptor alpha (erralpha , nr3b1) complexe with a coactivator peptide from peroxisome proliferator - activate receptor coactivator-1alpha (pgc-1alpha) reveal a transcriptionall

							y active conformation in the absence of a ligand .
GiG	GP1BA	VWF	0.518	0.527	144	Existing	these finding indicate the novel bind site require for vwf binding of human gpibalpha .
GiG	NR2C1	NR2C1	0.027	0.522	26	Novel	the human testicular receptor 2 (tr2) , a member of the nuclear hormone receptor superfamily , have no identify ligand yet .
GiG	NCOA1	ESRRG	0.992	0.518	1	Novel	the crystal structure of the ligand bind

							domain (lbd) of the estrogen - relate receptor 3 (err3) complexe with a steroid receptor coactivator-1 (src-1) peptide reveal a transcriptionally active conformation in absence of any ligand .
GiG	PPARG	PPARG	0.824	0.504	2497	Existing	although these agent can bind and activate an orphan nuclear receptor , peroxisome proliferator - activate

							receptor gamma (ppargamma) , there be no direct evidence to conclusively implicate this receptor in the regulation of mammalian glucose homeostasis .
GiG	ESR2	ESR1	0.995	0.503	1715	Novel	ligand bind experiment with purify er alpha and er beta confirm that the two phytoestrogen be er ligand .
GiG	FGFR2	FGFR2	1.000	0.501	584	Existing	receptor modeling of kgfr be use to identify selective kgfr

tyrosine kinase (tk) inhibitor molecule that have the potential to bind selectively to the kgfr .

BIBLIOGRAPHY

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

2. **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

3. **arXiv.org: the Los Alamos National Laboratory e□print server**
Gerry McKiernan
International Journal on Grey Literature (2000-09-01) https://doi.org/fg8pw7
DOI: 10.1108/14666180010345564

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

5. **medRxiv: Navigating the New Frontier of Medical Research and Publishing**
David Petersen
Journal of Electronic Resources in Medical Libraries (2022-03-15) https://doi.org/gp8zpk
DOI: 10.1080/15424065.2022.2046229

6. **Scientific communication pathways: an overview and introduction to a symposium**
David F. Zaye, W. V. Metanomski
Journal of Chemical Information and Computer Sciences (1986-05-01) https://doi.org/bwsxhq
DOI: 10.1021/ci00050a001

7. **The trouble with medical journals**
R. Smith
Journal of the Royal Society of Medicine (2006-03-01) https://doi.org/bntcbh
DOI: 10.1258/jrsm.99.3.115 · PMID: 16508048 · PMCID: PMC1383755

8. **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

9. **The development of preprints during the COVID□19 pandemic**
Andreas Älgå, Oskar Eriksson, Martin Nordberg
Journal of Internal Medicine (2021-02-09) https://doi.org/gh9crc
DOI: 10.1111/joim.13240 · PMID: 33560546 · PMCID: PMC8014163

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

11. **Abstract**
eLife Sciences Publications, Ltd
https://doi.org/gf5cqt
DOI: 10.7554/elife.45133.001

12. **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

13. **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 (2019-01)
https://doi.org/ggz7f9
DOI: 10.1002/pra2.175

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

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

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

17. **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

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

19. **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

20. **Comparing published scientific journal articles to their pre-print versions**
Martin Klein, Peter Broadwell, Sharon E. Farb, Todd Grappone

International Journal on Digital Libraries (2018-02-05) https://doi.org/cmd5
DOI: 10.1007/s00799-018-0234-1

21. **Tracking changes between preprint posting and journal publication during a pandemic**
Liam Brierley, Federico Nanni, Jessica K. Polka, Gautam Dey, Máté Pálfy, Nicholas Fraser, Jonathon Alexis Coates
PLOS Biology (2022-02-01) https://doi.org/gpcg6r
DOI: 10.1371/journal.pbio.3001285 · PMID: 35104285 · PMCID: PMC8806067

22. **Semantic Change**
Elizabeth Closs Traugott
Oxford Research Encyclopedia of Linguistics (2017-03-29) https://doi.org/gp574c
DOI: 10.1093/acrefore/9780199384655.013.323

23. **Syntactic annotations for the google books ngram corpus**
Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, Slav Petrov
Proceedings of the acl 2012 system demonstrations (2012)

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

25. **Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English**
Mark Davies
Corpora (2012-11) https://doi.org/gf9jpj
DOI: 10.3366/cor.2012.0024

26. **Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change**
William L. Hamilton, Jure Leskovec, Dan Jurafsky
arXiv (2016) https://doi.org/gp8zpp
DOI: 10.48550/arxiv.1605.09096

27. **Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change**
William L. Hamilton, Jure Leskovec, Dan Jurafsky
Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016) https://doi.org/gfw6bv
DOI: 10.18653/v1/d16-1229 · PMID: 28580459 · PMCID: PMC5452980

28. **A distributional similarity approach to the detection of semantic change in the google books ngram corpus.**
Kristina Gulordava, Marco Baroni
Proceedings of the gems 2011 workshop on geometrical models of natural language semantics (2011)

29. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
arXiv (2014) https://doi.org/gp8zpn
DOI: 10.48550/arxiv.1411.3315

30. **A framework for analyzing semantic change of words across time**
Adam Jatowt, Kevin Duh
IEEE/ACM Joint Conference on Digital Libraries (2014-09) https://doi.org/gp8zpm
DOI: 10.1109/jcdl.2014.6970173

31. **Learning Diachronic Word Embeddings with Iterative Stable Information Alignment**
Zefeng Lin, Xiaojun Wan, Zongming Guo
Natural Language Processing and Chinese Computing (2019) https://doi.org/gp8zpg
DOI: 10.1007/978-3-030-32233-5_58

32. **Deep Neural Models of Semantic Shift**
Alex Rosenfeld, Katrin Erk
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (2018) https://doi.org/gp574f
DOI: 10.18653/v1/n18-1044

33. **Words are Malleable: Computing Semantic Shifts in Political and Media Discourse**
Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, Jaap Kamps
arXiv (2017) https://doi.org/gp8zpq
DOI: 10.48550/arxiv.1711.05603

34. **Dynamic Word Embeddings for Evolving Semantic Discovery**
Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, Hui Xiong
Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2018-02-02) https://doi.org/ghtk5v
DOI: 10.1145/3159652.3159703

35. **Reading Between the Lines: Prediction of Political Violence Using Newspaper Text**
HANNES MUELLER, CHRISTOPHER RAUH
American Political Science Review (2017-12-14) https://doi.org/gdj77d
DOI: 10.1017/s0003055417000570

36. **Detection of Emerging Drugs Involved in Overdose via Diachronic Word Embeddings of Substances Discussed on Social Media**
Austin P. Wright, Christopher M. Jones, Duen Horng Chau, R. Matthew Gladden, Steven A. Sumner
Journal of Biomedical Informatics (2021-05) https://doi.org/gp8zph
DOI: 10.1016/j.jbi.2021.103824 · PMID: 34048933

37. **How COVID-19 Is Changing Our Language : Detecting Semantic Shift in Twitter Word Embeddings**
Yanzhu Guo, Christos Xypolopoulos, Michalis Vazirgiannis
arXiv (2021) https://doi.org/gp8zpr
DOI: 10.48550/arxiv.2102.07836

38. **Semantic Changepoint Detection for Finding Potentially Novel Research Publications**
Bhavish Dinakar, Mayla R. Boguslav, Carsten Görg, Deendayal Dinakarpandian
Biocomputing 2021 (2020-11) https://doi.org/gp574d
DOI: 10.1142/9789811232701_0011

39. **LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task**
Neha Warikoo, Yung-Chun Chang, Wen-Lian Hsu
Database (2018-01-01) https://doi.org/gfhjr6
DOI: 10.1093/database/bay108 · PMID: 30346607 · PMCID: PMC6196310

40. **DTMiner: identification of potential disease targets through biomedical literature mining**
Dong Xu, Meizhuo Zhang, Yanping Xie, Fan Wang, Ming Chen, Kenny Q. Zhu, Jia Wei
Bioinformatics (2016-08-09) https://doi.org/f9nw36
DOI: 10.1093/bioinformatics/btw503 · PMID: 27506226 · PMCID: PMC5181534

41. **Exploiting graph kernels for high performance biomedical relation extraction**
Nagesh C. Panyam, Karin Verspoor, Trevor Cohn, Kotagiri Ramamohanarao
Journal of Biomedical Semantics (2018-01-30) https://doi.org/gf49nn
DOI: 10.1186/s13326-017-0168-3 · PMID: 29382397 · PMCID: PMC5791373

42. **BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction.**
Siddhartha Jonnalagadda, Graciela Gonzalez
AMIA … Annual Symposium proceedings. AMIA Symposium (2010-11-13)
https://www.ncbi.nlm.nih.gov/pubmed/21346999
PMID: 21346999 · PMCID: PMC3041388

43. **iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system.**
Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, K Vijay-Shanker
Database : the journal of biological databases and curation (2014-05-21)
https://www.ncbi.nlm.nih.gov/pubmed/24850848
DOI: 10.1093/database/bau038 · PMID: 24850848 · PMCID: PMC4028706

44. **BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences**
K. E. Ravikumar, Majid Rastegar-Mojarad, Hongfang Liu
Database (2017-01-01) https://doi.org/gf7rbx
DOI: 10.1093/database/baw156 · PMID: 28365720 · PMCID: PMC5467463

45. **A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems**
Yifan Peng, Manabu Torii, Cathy H Wu, K Vijay-Shanker
BMC Bioinformatics (2014-08-23) https://doi.org/f6rndz
DOI: 10.1186/1471-2105-15-285 · PMID: 25149151 · PMCID: PMC4262219

46. **Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system**
Catalina O. Tudor, Karen E. Ross, Gang Li, K. Vijay-Shanker, Cathy H. Wu, Cecilia N. Arighi
Database (2015-01-01) https://doi.org/gf8fpt
DOI: 10.1093/database/bav020 · PMID: 25833953 · PMCID: PMC4381107

47. **miRTex: A Text Mining System for miRNA-Gene Relation Extraction**
Gang Li, Karen E. Ross, Cecilia N. Arighi, Yifan Peng, Cathy H. Wu, K. Vijay-Shanker
PLOS Computational Biology (2015-09-25) https://doi.org/f75mwb
DOI: 10.1371/journal.pcbi.1004391 · PMID: 26407127 · PMCID: PMC4583433

48. **LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes**
Andres Cañada, Salvador Capella-Gutierrez, Obdulia Rabal, Julen Oyarzabal, Alfonso Valencia, Martin Krallinger
Nucleic Acids Research (2017-05-22) https://doi.org/gf479h
DOI: 10.1093/nar/gkx462 · PMID: 28531339 · PMCID: PMC5570141

49. **Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature**
H.-M. Müller, K. M. Van Auken, Y. Li, P. W. Sternberg
BMC Bioinformatics (2018-03-09) https://doi.org/gf7rbz
DOI: 10.1186/s12859-018-2103-8 · PMID: 29523070 · PMCID: PMC5845379

50. **DiMeX: A Text Mining System for Mutation-Disease Association Extraction**
A. S. M. Ashique Mahmood, Tsung-Jung Wu, Raja Mazumder, K. Vijay-Shanker
PLOS ONE (2016-04-13) https://doi.org/f8xktj
DOI: 10.1371/journal.pone.0152725 · PMID: 27073839 · PMCID: PMC4830514

51. **Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors**
F. Horn, A. L. Lau, F. E. Cohen
Bioinformatics (2004-01-22) https://doi.org/d7cjgj
DOI: 10.1093/bioinformatics/btg449 · PMID: 14990452

52. **Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing**
Rong Xu, QuanQiu Wang
BMC Bioinformatics (2013-06-06) https://doi.org/gb8v3k
DOI: 10.1186/1471-2105-14-181 · PMID: 23742147 · PMCID: PMC3702428

53. **RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information**
Manabu Torii, Cecilia N. Arighi, Gang Li, Qinghua Wang, Cathy H. Wu, K. Vijay-Shanker
IEEE/ACM Transactions on Computational Biology and Bioinformatics (2015-01-01) https://doi.org/gf8fpv
DOI: 10.1109/tcbb.2014.2372765 · PMID: 26357075 · PMCID: PMC4568560

54. **PKDE4J: Entity and relation extraction for public knowledge discovery**
Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, Keun Young Kang
Journal of Biomedical Informatics (2015-10) https://doi.org/f7v7jj
DOI: 10.1016/j.jbi.2015.08.008 · PMID: 26277115

55. **PhpSyntaxTree tool**
A Eisenbach, M Eisenbach
(2006)

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

57. **DISEASES: Text mining and data integration of disease–gene associations**
Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, Lars Juhl Jensen
Methods (2015-03) https://doi.org/f3mn6s
DOI: 10.1016/j.ymeth.2014.11.020 · PMID: 25484339

58. **STRING v9.1: protein-protein interaction networks, with increased coverage and integration**
Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, Lars J. Jensen
Nucleic Acids Research (2012-11-29) https://doi.org/gf5kcd
DOI: 10.1093/nar/gks1094 · PMID: 23203871 · PMCID: PMC3531103

59. **A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts**
David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, Søren Brunak
PLOS Computational Biology (2018-02-15) https://doi.org/gcx747
DOI: 10.1371/journal.pcbi.1005962 · PMID: 29447159 · PMCID: PMC5831415

60. **STITCH 4: integration of protein–chemical interactions with user data**
Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H. Blicher, Christian von Mering, Lars J. Jensen, Peer Bork
Nucleic Acids Research (2013-11-28) https://doi.org/f5shb4
DOI: 10.1093/nar/gkt1207 · PMID: 24293645 · PMCID: PMC3964996

61. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman
Bioinformatics (2018-02-27) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

62. **CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Bioinformatics (2019-06-14) https://doi.org/gf4789
DOI: 10.1093/bioinformatics/btz490 · PMID: 31199464 · PMCID: PMC6956794

63. **A new method for prioritizing drug repositioning candidates extracted by literature-based discovery**
Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Dingcheng Li, Rashmi Prasad, Hongfang Liu
2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2015-11)
https://doi.org/gf479j
DOI: 10.1109/bibm.2015.7359766

64. **Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases**
Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, Wynand Alkema
PLoS Computational Biology (2010-09-23) https://doi.org/bhrw7x
DOI: 10.1371/journal.pcbi.1000943 · PMID: 20885778 · PMCID: PMC2944780

65. **STRING v10: protein–protein interaction networks, integrated over the tree of life**
Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, … Christian von Mering
Nucleic Acids Research (2014-10-28) https://doi.org/f64rfn
DOI: 10.1093/nar/gku1003 · PMID: 25352553 · PMCID: PMC4383874

66. **Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine**
Ayush Singhal, Michael Simmons, Zhiyong Lu
PLOS Computational Biology (2016-11-30) https://doi.org/f9gz4b
DOI: 10.1371/journal.pcbi.1005017 · PMID: 27902695 · PMCID: PMC5130168

67. **Overview of the biocreative vi chemical-protein interaction track**
Martin Krallinger, Obdulia Rabal, Saber A Akhondi, others
Proceedings of the sixth biocreative challenge evaluation workshop (2017)
https://www.semanticscholar.org/paper/Overview-of-the-BioCreative-VI-chemical-protein-Krallinger-Rabal/eed781f498b563df5a9e8a241c67d63dd1d92ad5

68. **BioCreative V CDR task corpus: a resource for chemical disease relation extraction**
Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, Zhiyong Lu
Database (2016) https://doi.org/gf5hfw
DOI: 10.1093/database/baw068 · PMID: 27161011 · PMCID: PMC4860626

69. **RelEx–Relation extraction using dependency parse trees**
K. Fundel, R. Kuffner, R. Zimmer
Bioinformatics (2006-12-01) https://doi.org/cz7q4d
DOI: 10.1093/bioinformatics/btl616 · PMID: 17142812

70. **CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations**
Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, Jong C Park
BMC Bioinformatics (2013-11-14) https://doi.org/gb8v5s
DOI: 10.1186/1471-2105-14-323 · PMID: 24225062 · PMCID: PMC3833657

71. **Text Mining for Protein Docking**
Varsha D. Badal, Petras J. Kundrotas, Ilya A. Vakser
PLOS Computational Biology (2015-12-09) https://doi.org/gcvj3b
DOI: 10.1371/journal.pcbi.1004630 · PMID: 26650466 · PMCID: PMC4674139

72. **Automatic extraction of gene-disease associations from literature using joint ensemble learning**
Balu Bhasuran, Jeyakumar Natarajan
PLOS ONE (2018-07-26) https://doi.org/gdx63f
DOI: 10.1371/journal.pone.0200699 · PMID: 30048465 · PMCID: PMC6061985

73. **Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research**
Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, Laura I Furlong
BMC Bioinformatics (2015-02-21) https://doi.org/f7kn8s
DOI: 10.1186/s12859-015-0472-9 · PMID: 25886734 · PMCID: PMC4466840

74. **Deep learning**
Ian Goodfellow, Yoshua Bengio, Aaron Courville
The MIT Press (2016)
ISBN: 0262035618, 9780262035613

75. **Deep learning**
Yann LeCun, Yoshua Bengio, Geoffrey Hinton
Nature (2015-05-27) https://doi.org/bmqp
DOI: 10.1038/nature14539 · PMID: 26017442

76. **Long Short-Term Memory**
Sepp Hochreiter, Jürgen Schmidhuber
Neural Computation (1997-11-01) https://doi.org/bxd65w
DOI: 10.1162/neco.1997.9.8.1735 · PMID: 9377276

77. **Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts**
Anne Cocos, Alexander G Fiks, Aaron J Masino
Journal of the American Medical Informatics Association (2017-02-22) https://doi.org/gbp9nj
DOI: 10.1093/jamia/ocw180 · PMID: 28339747 · PMCID: PMC7651964

78. **Semantic Relations in Compound Nouns: Perspectives from Inter-Annotator Agreement**
Prabha Yadav, Elisabetta Jezek, Pierrette Bouillon, Tiffany J Callahan, Michael Bada, Lawrence E Hunter, K Bretonnel Cohen
Studies in health technology and informatics (2017)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7781293/
PMID: 29295175 · PMCID: PMC7781293

79. **Cross-Sentence N-ary Relation Extraction with Graph LSTMs**
Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih
arXiv (2017-08-15) https://arxiv.org/abs/1708.03743

80. **Drug drug interaction extraction from biomedical literature using syntax convolutional neural network**
Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, Jian Wang
Bioinformatics (2016-07-27) https://doi.org/f9nsq7
DOI: 10.1093/bioinformatics/btw486 · PMID: 27466626 · PMCID: PMC5181565

81. **N-ary Relation Extraction using Graph State LSTM**
Linfeng Song, Yue Zhang, Zhiguo Wang, Daniel Gildea
arXiv (2018-08-29) https://arxiv.org/abs/1808.09101

82. **A neural joint model for entity and relation extraction from biomedical text**
Fei Li, Meishan Zhang, Guohong Fu, Donghong Ji
BMC Bioinformatics (2017-03-31) https://doi.org/gcgnx2
DOI: 10.1186/s12859-017-1609-9 · PMID: 28359255 · PMCID: PMC5374588

83. **The problem of learning long-term dependencies in recurrent networks**
Y. Bengio, P. Frasconi, P. Simard
IEEE International Conference on Neural Networks https://doi.org/d7zs24
DOI: 10.1109/icnn.1993.298725

84. **Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network**
Alex Sherstinsky
Physica D: Nonlinear Phenomena (2020-03) https://doi.org/ggmzpd
DOI: 10.1016/j.physd.2019.132306

85. **On the difficulty of training Recurrent Neural Networks**
Razvan Pascanu, Tomas Mikolov, Yoshua Bengio
arXiv (2013-02-19) https://arxiv.org/abs/1211.5063

86. **Revisiting Unreasonable Effectiveness of Data in Deep Learning Era**
Chen Sun, Abhinav Shrivastava, Saurabh Singh, Abhinav Gupta
arXiv (2017-08-07) https://arxiv.org/abs/1707.02968

87. **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

88. **Distributed Representations of Words and Phrases and their Compositionality**
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-10-18) https://arxiv.org/abs/1310.4546

89. **Deep learning for extracting protein-protein interactions from biomedical literature**
Yifan Peng, Zhiyong Lu
arXiv (2017-06-05) https://arxiv.org/abs/1706.01556v2

90. **Knowledge-guided convolutional networks for chemical-disease relation extraction**
Huiwei Zhou, Chengkun Lang, Zhuang Liu, Shixian Ning, Yingyu Lin, Lei Du
BMC Bioinformatics (2019-05-21) https://doi.org/gf45zn
DOI: 10.1186/s12859-019-2873-7 · PMID: 31113357 · PMCID: PMC6528333

91. **Extraction of protein–protein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings**
Sung-Pil Choi
Journal of Information Science (2016-11-01) https://doi.org/gcv8bn
DOI: 10.1177/0165551516673485

92. **Extracting chemical–protein relations with ensembles of SVM and deep learning models**
Yifan Peng, Anthony Rios, Ramakanth Kavuluru, Zhiyong Lu
Database (2018-01-01) https://doi.org/gf479f
DOI: 10.1093/database/bay073 · PMID: 30020437 · PMCID: PMC6051439

93. **Expanding a Database-derived Biomedical Knowledge Graph via Multi-relation Extraction from Biomedical Abstracts**
David N. Nicholson, Daniel S. Himmelstein, Casey S. Greene
Cold Spring Harbor Laboratory (2019-08-08) https://doi.org/gf6qxh
DOI: 10.1101/730085

94. **Distant supervision for relation extraction without labeled data**
Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky
Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09 (2009) https://doi.org/fg9q43
DOI: 10.3115/1690219.1690287

95. **Introduction to Semi-Supervised Learning**
Xiaojin Zhu, Andrew B. Goldberg
Synthesis Lectures on Artificial Intelligence and Machine Learning (2009-01) https://doi.org/bq7pm2
DOI: 10.2200/s00196ed1v01y200906aim006

96. **A Survey on Transfer Learning**
Sinno Jialin Pan, Qiang Yang
IEEE Transactions on Knowledge and Data Engineering (2010-10) https://doi.org/bc4vws
DOI: 10.1109/tkde.2009.191

97. **A survey of transfer learning**
Karl Weiss, Taghi M. Khoshgoftaar, DingDing Wang
Journal of Big Data (2016-05-28) https://doi.org/gfkr2w
DOI: 10.1186/s40537-016-0043-6

98. **Exploring Semi-supervised Variational Autoencoders for Biomedical Relation Extraction**
Yijia Zhang, Zhiyong Lu
arXiv (2019-01-18) https://arxiv.org/abs/1901.06103v1

99. **Large-scale extraction of gene interactions from full-text literature using DeepDive**
Emily K. Mallory, Ce Zhang, Christopher Ré, Russ B. Altman
Bioinformatics (2015-09-03) https://doi.org/gb5g7b
DOI: 10.1093/bioinformatics/btv476 · PMID: 26338771 · PMCID: PMC4681986

100. **Snorkel**
Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
Proceedings of the VLDB Endowment (2017-11) https://doi.org/ch44
DOI: 10.14778/3157794.3157797 · PMID: 29770249 · PMCID: PMC5951191

101. **Snorkel MeTaL**
Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, Christopher Ré
Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning (2018-06-15) https://doi.org/gf3xk7
DOI: 10.1145/3209889.3209898 · PMID: 30931438 · PMCID: PMC6436830

102. **Learning protein protein interaction extraction using distant supervision**
Philippe Thomas, Illés Solt, Roman Klinger, Ulf Leser
(2011-01)
103. **Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning**
Pengda Qin, Weiran Xu, William Yang Wang
arXiv (2018-05-28) https://arxiv.org/abs/1805.09927

104. **DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction**
Pengda Qin, Weiran Xu, William Yang Wang
arXiv (2018-05-28) https://arxiv.org/abs/1805.09929

105. **Noise Reduction Methods for Distantly Supervised Biomedical Relation Extraction**
Gang Li, Cathy Wu, K. Vijay-Shanker
BioNLP 2017 (2017) https://doi.org/ggmk8s
DOI: 10.18653/v1/w17-2323

106. **Comparative experiments on learning information extractors for proteins and their interactions**
Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, Yuk Wah Wong
Artificial Intelligence in Medicine (2005-02) https://doi.org/dhztpn
DOI: 10.1016/j.artmed.2004.07.016 · PMID: 15811782

107. **BioInfer: a corpus for information extraction in the biomedical domain**
Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, Tapio Salakoski
BMC Bioinformatics (2007-02-09) https://doi.org/b7bhhc
DOI: 10.1186/1471-2105-8-50 · PMID: 17291334 · PMCID: PMC1808065

108. **Learning language in logic - genic interaction extraction challenge**
C. Nédellec
Proceedings of the learning language in logic 2005 workshop at the international conference on machine learning (2005)

109. **Mining medline: Abstracts, sentences, or phrases?**
Jing Ding, Daniel Berleant, Dan Nettleton, Eve Syrkin Wurtele
Pacific symposium on biocomputing (2002) http://helix-web.stanford.edu/psb02/ding.pdf

110. **The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships**
Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, Laura I. Furlong
Journal of Biomedical Informatics (2012-10) https://doi.org/f36vn6
DOI: 10.1016/j.jbi.2012.04.004 · PMID: 22554700

111. **Concept annotation in the CRAFT corpus**
Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner Jr, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, Lawrence E Hunter
BMC Bioinformatics (2012-07-09) https://doi.org/gb8vdr
DOI: 10.1186/1471-2105-13-161 · PMID: 22776079 · PMCID: PMC3476437

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

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

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

115. **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/

116. **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

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

118. **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

119. **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

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

121. **Day-to-day discovery of preprint–publication links**
Guillaume Cabanac, Theodora Oikonomidi, Isabelle Boutron
Scientometrics (2021-04-18) https://doi.org/gjr9k4
DOI: 10.1007/s11192-021-03900-7 · PMID: 33897069 · PMCID: PMC8053368

122. **Preprints in motion: tracking changes between preprint posting and journal publication during a pandemic**
Liam Brierley, Federico Nanni, Jessica K Polka, Gautam Dey, Máté Pálfy, Nicholas Fraser, Jonathon Alexis Coates
Cold Spring Harbor Laboratory (2021-02-20) https://doi.org/gh5mhm
DOI: 10.1101/2021.02.20.432090

123. **Textual Analysis in Accounting and Finance: A Survey**
TIM LOUGHRAN, BILL MCDONALD
Journal of Accounting Research (2016-06-08) https://doi.org/gc3hf7
DOI: 10.1111/1475-679x.12123

124. **SciReader: A Cloud-based Recommender System for Biomedical Literature**
Priya Desai, Natalie Telis, Ben Lehmann, Keith Bettinger, Jonathan K. Pritchard, Somalee Datta
Cold Spring Harbor Laboratory (2018-05-30) https://doi.org/gkw2zw
DOI: 10.1101/333922

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

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

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

128. **PubTator central: automated concept annotation for biomedical full text articles**
Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu
Nucleic Acids Research (2019-05-22) https://doi.org/ggzfsc
DOI: 10.1093/nar/gkz389 · PMID: 31114887 · PMCID: PMC6602571

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

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

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

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

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

134. **BioRxiv Machine access and text/data mining resources** https://www.biorxiv.org/tdm

135. **PubMed Central: The GenBank of the published literature**
Richard J. Roberts
Proceedings of the National Academy of Sciences (2001-01-09) https://doi.org/bbn9k8
DOI: 10.1073/pnas.98.2.381 · PMID: 11209037 · PMCID: PMC33354

136. **For Authors - PMC** https://www.ncbi.nlm.nih.gov/pmc/about/submission-methods/

137. **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

138. **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

139. **About PMC - PMC** https://www.ncbi.nlm.nih.gov/pmc/about/intro/

140. **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 (2019-01-31) https://doi.org/ggzfsb
DOI: 10.1093/bioinformatics/btz070 · PMID: 30715220 · PMCID: PMC6748740

141. **Author Manuscripts in PMC - PMC** https://www.ncbi.nlm.nih.gov/pmc/about/authorms/
142. **CrossRef Text and Data Mining Services**
Rachael Lammey
Insights the UKSG journal (2015-07-07) https://doi.org/gg4hp9
DOI: 10.1629/uksg.233

143. **Odds Ratio**
Steven Tenny, Mary R. Hoffman
StatPearls (2022) http://www.ncbi.nlm.nih.gov/books/NBK431098/

144. **Gensim–python framework for vector space modelling**
Radim Rehurek, Petr Sojka
NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2011)

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

146. **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

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

148. **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

149. **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
Cold Spring Harbor Laboratory (2018-07-24) https://doi.org/gg4hp7
DOI: 10.1101/376665

150. **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 (2019-01-01) https://doi.org/gg4hp8
DOI: 10.1242/bio.038232 · PMID: 31164339 · PMCID: PMC6602326

151. **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

152. **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

153. **The bioRxiv Wall of Shame**
Jordan Anaya
(2018-08-03) https://medium.com/@OmnesRes/the-biorxiv-wall-of-shame-aa3d9cfc4cd7

154. **Journal/Author Name Estimator (JANE)**
Carolann Lee Curry
Journal of the Medical Library Association (2019-01-04) https://doi.org/ghjw7j
DOI: 10.5195/jmla.2019.598 · PMCID: PMC6300233

155. **Introduction — PyMuPDF 1.20.0 documentation**
https://pymupdf.readthedocs.io/en/latest/intro.html

156. **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

157. **Preprinting the COVID-19 pandemic**
Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K Polka, Máté Pálfy, Federico Nanni, Jonathon Alexis Coates
Cold Spring Harbor Laboratory (2020-05-23) https://doi.org/dxdb
DOI: 10.1101/2020.05.22.111294

158. **ID Converter** https://www.ncbi.nlm.nih.gov/pmc/tools/idconv/

159. **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

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

161. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**
Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang
Bioinformatics (2019-09-10) https://doi.org/ggh5qq
DOI: 10.1093/bioinformatics/btz682 · PMID: 31501885 · PMCID: PMC7703786

162. **Unsupervised word embeddings capture latent knowledge from materials science literature**
Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, Anubhav Jain
Nature (2019-07) https://doi.org/gf4kk5
DOI: 10.1038/s41586-019-1335-8 · PMID: 31270483

163. **Distributed Representations of Sentences and Documents**
Quoc V. Le, Tomas Mikolov
arXiv (2014) https://doi.org/gp7nv9
DOI: 10.48550/arxiv.1405.4053

164. **Efficient Vector Representation for Documents through Corruption**
Minmin Chen
arXiv (2017) https://doi.org/gp7nwb
DOI: 10.48550/arxiv.1707.02377

165. **Document Network Projection in Pretrained Word Embedding Space**
Antoine Gourru, Adrien Guille, Julien Velcin, Julien Jacques
arXiv (2020) https://doi.org/gp7nv6
DOI: 10.48550/arxiv.2001.05727

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

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

168. **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

169. **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

170. **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

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

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

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

174. **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
Cold Spring Harbor Laboratory (2019-08-13) https://doi.org/gg94bf
DOI: 10.1101/733162

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

176. **A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity**
Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, Emmanuelle Charpentier
Science (2012-08-17) https://doi.org/f22dgd

DOI: 10.1126/science.1225829 · PMID: 22745249 · PMCID: PMC6286148
177. **Medical preprint server debuts**
Jocelyn Kaiser
Science (2019-06-05) https://doi.org/gpxkkf
DOI: 10.1126/science.aay2933

178. **Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change**
William L. Hamilton, Jure Leskovec, Dan Jurafsky
arXiv (2018-10-26) https://arxiv.org/abs/1605.09096

179. **Multi-Label Zero-Shot Learning via Concept Embedding**
Ubai Sandouk, Ke Chen
arXiv https://arxiv.org/abs/1606.00282

180. **Bayesian Online Changepoint Detection**
Ryan Prescott Adams, David J. C. MacKay
arXiv (2007-10-22) https://arxiv.org/abs/0710.3742

181. **Adaptive filtering and change detection**
Fredrik Gustafsson, Fredrik Gustafsson
Citeseer (2000)

182. **Tracing armed conflicts with diachronic word embedding models**
Andrey Kutuzov, Erik Velldal, Lilja Øvrelid
Proceedings of the Events and Stories in the News Workshop (2017) https://doi.org/ghx5gj
DOI: 10.18653/v1/w17-2705

183. **Words are Malleable: Computing Semantic Shifts in Political and Media Discourse**
Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, Jaap Kamps
arXiv (2017-11-16) https://arxiv.org/abs/1711.05603

184. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
arXiv (2014-11-13) https://arxiv.org/abs/1411.3315

185. **Semantic word shifts in a scientific domain**
Baitong Chen, Ying Ding, Feicheng Ma
Scientometrics (2018-07-13) https://doi.org/gd7bd7
DOI: 10.1007/s11192-018-2843-2

186. **How COVID-19 Is Changing Our Language : Detecting Semantic Shift in Twitter Word Embeddings**
Yanzhu Guo, Christos Xypolopoulos, Michalis Vazirgiannis
arXiv (2021-02-17) https://arxiv.org/abs/2102.07836

187. **Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study**
Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, Satrajit S Ghosh
Journal of Medical Internet Research (2020-10-12) https://doi.org/ghm9v2
DOI: 10.2196/22635 · PMID: 32936777 · PMCID: PMC7575341

188. **TaggerOne: joint named entity recognition and normalization with semi-Markov Models**
Robert Leaman, Zhiyong Lu
Bioinformatics (2016-06-09) https://doi.org/f855dg
DOI: 10.1093/bioinformatics/btw343 · PMID: 27283952 · PMCID: PMC5018376

189. **GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
BioMed Research International (2015) https://doi.org/gb85jb
DOI: 10.1155/2015/918710 · PMID: 26380306 · PMCID: PMC4561873

190. **SR4GN: A Species Recognition Software Tool for Gene Normalization**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
PLoS ONE (2012-06-05) https://doi.org/gpq498
DOI: 10.1371/journal.pone.0038460 · PMID: 22679507 · PMCID: PMC3367953

191. **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine**
Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu
Bioinformatics (2017-09-01) https://doi.org/gbzsmc
DOI: 10.1093/bioinformatics/btx541 · PMID: 28968638 · PMCID: PMC5860583

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

193. **Machine access and text/data mining resources | medRxiv** https://www.medrxiv.org/tdm

194. **Factors Influencing the Surprising Instability of Word Embeddings**
Laura Wendlandt, Jonathan K. Kummerfeld, Rada Mihalcea
arXiv (2020-06-05) https://arxiv.org/abs/1804.09692
DOI: 10.18653/v1/n18-1190

195. **Stability of Word Embeddings Using Word2Vec**
Mansi Chugh, Peter A. Whigham, Grant Dick
AI 2018: Advances in Artificial Intelligence (2018) https://doi.org/gpxkkc
DOI: 10.1007/978-3-030-03991-2_73

196. **Evaluating the Stability of Embedding-based Word Similarities**
Maria Antoniak, David Mimno
Transactions of the Association for Computational Linguistics (2018-12) https://doi.org/gf39k8
DOI: 10.1162/tacl_a_00008

197. **Predicting Word Embeddings Variability**
Benedicte Pierrejean, Ludovic Tanguy
Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (2018) https://doi.org/gh6qpc
DOI: 10.18653/v1/s18-2019

198. **A generalized solution of the orthogonal procrustes problem**
Peter H. Schönemann
Psychometrika (1966-03) https://doi.org/dx77sz
DOI: 10.1007/bf02289451

199. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv (2020-09-21) https://arxiv.org/abs/1802.03426

200. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
Proceedings of the 24th International Conference on World Wide Web (2015-05-18) https://doi.org/ghcv6k
DOI: 10.1145/2736277.2741627

201. **Improving semantic change analysis by combining word embeddings and word frequencies**
Adrian Englhardt, Jens Willkomm, Martin Schäler, Klemens Böhm
International Journal on Digital Libraries (2019-05-20) https://doi.org/gpxkkd
DOI: 10.1007/s00799-019-00271-6

202. **DUKweb, diachronic word representations from the UK Web Archive corpus**
Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu, Barbara McGillivray
Scientific Data (2021-10-15) https://doi.org/gqbkx4
DOI: 10.1038/s41597-021-01047-x · PMID: 34654827 · PMCID: PMC8520005

203. **SARS: clinical virology and pathogenesis**
John NICHOLLS, Xiao-Ping DONG, Gu JIANG, Malik PEIRIS
Respirology (2003-11) https://doi.org/cxjwrc
DOI: 10.1046/j.1440-1843.2003.00517.x · PMID: 15018126 · PMCID: PMC7169081

204. **Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges**
Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, Po-Ren Hsueh
International Journal of Antimicrobial Agents (2020-03) https://doi.org/ggpj9d
DOI: 10.1016/j.ijantimicag.2020.105924 · PMID: 32081636 · PMCID: PMC7127800

205. **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2***Nature Microbiology* (2020-03-02) https://doi.org/ggqj7m
DOI: 10.1038/s41564-020-0695-z · PMID: 32123347 · PMCID: PMC7095448

206. **Factors Influencing the Surprising Instability of Word Embeddings**
Laura Wendlandt, Jonathan K. Kummerfeld, Rada Mihalcea
arXiv (2018) https://doi.org/gqcn9m
DOI: 10.48550/arxiv.1804.09692

207. **PsyArXiv**
Margie Ruppel
The Charleston Advisor (2021-10-01) https://doi.org/gqbk4k
DOI: 10.5260/chara.23.2.38

208. **Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action**
Ruggero Gramatica, T. Di Matteo, Stefano Giorgetti, Massimo Barbiani, Dorian Bevec, Tomaso Aste
PLoS ONE (2014-01-09) https://doi.org/gf45zp
DOI: 10.1371/journal.pone.0084912 · PMID: 24416311 · PMCID: PMC3886994

209. **Drug repurposing through joint learning on knowledge graphs and literature**
Mona Alshahrani, Robert Hoehndorf
Cold Spring Harbor Laboratory (2018-08-06) https://doi.org/gf45zk
DOI: 10.1101/385617

210. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

211. **CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision**
Alexander Junge, Lars Juhl Jensen
Cold Spring Harbor Laboratory (2018-10-16) https://doi.org/gf45zm
DOI: 10.1101/444398

212. **Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?**
R. Winnenburg, T. Wachter, C. Plake, A. Doms, M. Schroeder
Briefings in Bioinformatics (2008-07-11) https://doi.org/bfsnwg
DOI: 10.1093/bib/bbn043 · PMID: 19060303

213. **Manual curation is not sufficient for annotation of genomic databases**
William A. Baumgartner Jr, K. Bretonnel Cohen, Lynne M. Fox, George Acquaah-Mensah, Lawrence Hunter
Bioinformatics (2007-07-01) https://doi.org/dtck86
DOI: 10.1093/bioinformatics/btm229 · PMID: 17646325 · PMCID: PMC2516305

214. **Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references**
Lutz Bornmann, Rüdiger Mutz
Journal of the Association for Information Science and Technology (2015-04-29) https://doi.org/gfj5zc
DOI: 10.1002/asi.23329

215. **PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more**
Yifeng Liu, Yongjie Liang, David Wishart
Nucleic Acids Research (2015-04-29) https://doi.org/f7nzn5
DOI: 10.1093/nar/gkv383 · PMID: 25925572 · PMCID: PMC4489268

216. **The research on gene-disease association based on text-mining of PubMed**
Jie Zhou, Bo-quan Fu
BMC Bioinformatics (2018-02-07) https://doi.org/gf479k
DOI: 10.1186/s12859-018-2048-y · PMID: 29415654 · PMCID: PMC5804013

217. **Analyzing a co-occurrence gene-interaction network to identify disease-gene association**
Amira Al-Aamri, Kamal Taha, Yousof Al-Hammadi, Maher Maalouf, Dirar Homouz
BMC Bioinformatics (2019-02-08) https://doi.org/gf49nm
DOI: 10.1186/s12859-019-2634-7 · PMID: 30736752 · PMCID: PMC6368766

218. **COMPARTMENTS: unification and visualization of protein subcellular localization evidence**
J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, L. J. Jensen
Database (2014-02-25) https://doi.org/btbm
DOI: 10.1093/database/bau012 · PMID: 24573882 · PMCID: PMC3935310

219. **Comprehensive comparison of large-scale tissue expression datasets**
Alberto Santos, Kalliopi Tsafou, Christian Stolte, Sune Pletscher-Frankild, Seán I. O'Donoghue, Lars Juhl Jensen
PeerJ (2015-06-30) https://doi.org/f3mn6p
DOI: 10.7717/peerj.1054 · PMID: 26157623 · PMCID: PMC4493645

220. **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text**
Yael Garten, Russ B Altman
BMC Bioinformatics (2009-02) https://doi.org/df75hq
DOI: 10.1186/1471-2105-10-s2-s6 · PMID: 19208194 · PMCID: PMC2646239

221. **PPInterFinder—a mining tool for extracting causal relations on human proteins from literature**
Kalpana Raja, Suresh Subramani, Jeyakumar Natarajan
Database (2013-01-01) https://doi.org/gf479b
DOI: 10.1093/database/bas052 · PMID: 23325628 · PMCID: PMC3548331

222. **PKDE4J: Entity and relation extraction for public knowledge discovery.**
Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, Keun Young Kang
Journal of biomedical informatics (2015-08-12) https://www.ncbi.nlm.nih.gov/pubmed/26277115
DOI: 10.1016/j.jbi.2015.08.008 · PMID: 26277115

223. **Extracting chemical–protein relations using attention-based neural networks**
Sijia Liu, Feichen Shen, Ravikumar Komandur Elayavilli, Yanshan Wang, Majid Rastegar-Mojarad, Vipin Chaudhary, Hongfang Liu
Database (2018-01-01) https://doi.org/gfdz8d
DOI: 10.1093/database/bay102 · PMID: 30295724 · PMCID: PMC6174551
224. **Deep learning in neural networks: An overview**
Jürgen Schmidhuber
Neural Networks (2015-01) https://doi.org/f6v78n
DOI: 10.1016/j.neunet.2014.09.003 · PMID: 25462637

225. **Probing Biomedical Embeddings from Language Models**
Qiao Jin, Bhuwan Dhingra, William W. Cohen, Xinghua Lu
arXiv (2019-04-05) https://arxiv.org/abs/1904.02181

226. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**
Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang
arXiv (2019-10-21) https://arxiv.org/abs/1901.08746
DOI: 10.1093/bioinformatics/btz682

227. **Attention Is All You Need**
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv (2017-12-07) https://arxiv.org/abs/1706.03762

228. **Chemical–gene relation extraction using recursive neural network**
Sangrak Lim, Jaewoo Kang
Database (2018-01-01) https://doi.org/gdss6f
DOI: 10.1093/database/bay060 · PMID: 29961818 · PMCID: PMC6014134

229. **Overview of the biocreative vi chemical-protein interaction track**
Martin Krallinger, Obdulia Rabal, Saber A Akhondi, others
Proceedings of the sixth biocreative challenge evaluation workshop (2017)
https://www.semanticscholar.org/paper/Overview-of-the-BioCreative-VI-chemical-protein-Krallinger-Rabal/eed781f498b563df5a9e8a241c67d63dd1d92ad5

230. **Comparative analysis of five protein-protein interaction corpora**
Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, Tapio Salakoski
BMC Bioinformatics (2008-04) https://doi.org/fh3df7
DOI: 10.1186/1471-2105-9-s3-s6 · PMID: 18426551 · PMCID: PMC2349296

231. **Revisiting distant supervision for relation extraction**
Tingsong Jiang, Jing Liu, Chin-Yew Lin, Zhifang Sui
Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018) (2018-05) https://aclanthology.org/L18-1566

232. **Distant Supervision for Large-Scale Extraction of Gene–Disease Associations from Literature Using DeepDive**
Balu Bhasuran, Jeyakumar Natarajan
International Conference on Innovative Computing and Communications (2018-11-20) https://doi.org/gf5hfv
DOI: 10.1007/978-981-13-2354-6_39

233. **Data Programming: Creating Large Training Sets, Quickly**
Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré
arXiv (2018-12-10) https://arxiv.org/abs/1605.07723

234. **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)**
Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, … Helen Parkinson
Nucleic Acids Research (2016-11-29) https://doi.org/f9v7cp
DOI: 10.1093/nar/gkw1133 · PMID: 27899670 · PMCID: PMC5210590

235. **A Proteome-Scale Map of the Human Interactome Network**
Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, … Marc Vidal
Cell (2014-11) https://doi.org/f3mn6x
DOI: 10.1016/j.cell.2014.10.050 · PMID: 25416956 · PMCID: PMC4266588

236. **DrugBank 5.0: a major update to the DrugBank database for 2018**
David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, … Michael Wilson
Nucleic Acids Research (2017-11-08) https://doi.org/gcwtzk
DOI: 10.1093/nar/gkx1037 · PMID: 29126136 · PMCID: PMC5753335

237. **Snorkel: rapid training data creation with weak supervision**
Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
The VLDB Journal (2019-07-15) https://doi.org/ghbw5f
DOI: 10.1007/s00778-019-00552-1 · PMID: 32214778 · PMCID: PMC7075849

238. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
arXiv (2018) https://doi.org/hm65
DOI: 10.48550/arxiv.1810.04805

239. **Transformers: State-of-the-Art Natural Language Processing**
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, … Alexander M. Rush
Association for Computational Linguistics (2020-10)
https://www.aclweb.org/anthology/2020.emnlp-demos.6

240. **Adam: A Method for Stochastic Optimization**
Diederik P. Kingma, Jimmy Ba
arXiv (2017-01-31) https://arxiv.org/abs/1412.6980

241. **Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature**
Clarissa F. D. Carneiro, Victor G. S. Queiroz, Thiago C. Moulin, Carlos A. M. Carvalho, Clarissa B. Haas, Danielle Rayêe, David E. Henshall, Evandro A. De-Souza, Felippe E. Amorim, Flávia Z. Boos, … Olavo B. Amaral
Research Integrity and Peer Review (2020-12) https://doi.org/gjfn8t
DOI: 10.1186/s41073-020-00101-3 · PMID: 33292815 · PMCID: PMC7706207

242. **Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever**
Clare Watson
Nature Medicine (2022-01) https://doi.org/gn7nwd
DOI: 10.1038/s41591-021-01654-6 · PMID: 35031791

243. **The COVID-19 pandemic**
Marco Ciotti, Massimo Ciccozzi, Alessandro Terrinoni, Wen-Can Jiang, Cheng-Bin Wang, Sergio Bernardini
Critical Reviews in Clinical Laboratory Sciences (2020-07-09) https://doi.org/gg8kxx
DOI: 10.1080/10408363.2020.1783198 · PMID: 32645276