Navigating the Landscape of Reproducible Research: A Predictive Modeling Approach

Akhil Pandey Akella
Northern Illinois University
Dekalb, Illinois, USA
Northwestern University
Evanston, Illinois, USA
aakella@niu.edu

Sagnik Ray Choudhury
University of North Texas
Denton, Texas, USA
sagnikrayc@gmail.com

David Koop
Northern Illinois University
Dekalb, Illinois, USA
dakoop@niu.edu

Hamed Alhoori
Northern Illinois University
Dekalb, Illinois, USA
alhoori@niu.edu

Abstract

The reproducibility of scientific articles is central to the advancement of science. Despite this importance, evaluating reproducibility remains challenging due to the scarcity of ground truth data. Predictive models can address this limitation by streamlining the tedious evaluation process. Typically, a paper's reproducibility is inferred based on the availability of artifacts such as code, data, or supplemental information, often without extensive empirical investigation. To address these issues, we utilized artifacts of papers as fundamental units to develop a novel, dual-spectrum framework that focuses on author-centric and external-agent perspectives. We used the author-centric spectrum, followed by the external-agent spectrum, to guide a structured, model-based approach to quantify and assess reproducibility. We explored the interdependencies between different factors influencing reproducibility and found that linguistic features such as readability and lexical diversity are strongly correlated with papers achieving the highest statuses on both spectrums. Our work provides a model-driven pathway for evaluating the reproducibility of scientific research.

CCS Concepts

• **Computing methodologies** → *Machine learning*; **Model development and analysis**; • **Information systems** → *Information retrieval*; **Data extraction and integration**; • **Applied computing** → Digital libraries and archives.

Keywords

Reproducibility, Scientific Data, Science of Science

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679831

ACM Reference Format:
Akhil Pandey Akella, Sagnik Ray Choudhury, David Koop, and Hamed Alhoori. 2024. Navigating the Landscape of Reproducible Research: A Predictive Modeling Approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627673.3679831

1 Introduction

The abundance of open-source libraries, version control frameworks, and publicly-available, archived datasets has made it easier than ever to ensure transparency in the scientific process. However, this increased attention on research reproducibility [16, 23] has not necessarily driven the scholarly community to implement more transparent measures to make their work fully reproducible. Instead, an inverse phenomenon is observed: surveys indicate that scientists often believe many scholarly articles are irreproducible [4], a sentiment that spans multiple fields [12].

Given the existing perception, it is crucial to develop a data-driven approach that can establish trust in the reproducibility of scientific papers. The reproducibility of research papers is a complex issue [3, 8]. For example, consider a computational paper that researchers fail to reproduce despite the publicly available code and data, possibly due to the unavailability of specific libraries used in the original code. Such a paper should not be categorized alongside those that made no effort to ensure reproducibility. Therefore, reproducibility should be viewed as a *spectrum* rather than a binary classification. By acknowledging varying degrees of reproducibility, we can elevate trust across the board and help identify common factors that contribute to reproducible research. This refined approach reduces the collective burden on conferences, journals, publishers, and the research community at large.

The initial step in constructing a reproducibility spectrum is collecting existing ground truth about signals that indicate reproducible work. This can include meta-studies confirming the reproducibility of existing research [31], citations where methodologies have been re-implemented [22], and reproducibility challenges hosted by premier conferences [2, 7, 18, 25]. While these serve as proxy measures for reproducibility, establishing definitive ground truth for the reproducibility of scholarly work is challenging and

limited to a few sources. For example, conferences such as *OOPSLA, PLDI, and ISSTA* have conducted reproducibility reviews [5] to formally evaluate software artifacts and data. The practice of evaluating artifacts was first established at *SIGMOD* 2008 [6, 24], and various sub-disciplines within the Association for Computing Machinery (*ACM*) have since adopted similar policies to audit artifacts. Collecting signals from these efforts was fundamental for establishing the *ACM* badging process. In this process, a paper may receive badges such as *Artifacts Available*, *Artifacts Evaluated-Reusable*, and *Results Reproduced*. This policy acknowledges the researchers' efforts and incentivizes reproducibility.

While efforts like ACM Badging encourage the creation of reproducible research, the current system places a significant burden on the committees that evaluate artifact availability and reproducibility. However, the specific procedures for awarding reproducibility badges can vary across venues. Moreover, much of the literature on estimating and understanding reproducibility has relied on traditional modeling [32] and cohort-based statistical analysis [26]. While valuable, these approaches cannot scale effectively – assistance of automated systems such as predictive models is needed.

In this paper, we present a predictive modeling study utilizing a novel joint spectrum on reproducibility. This spectrum consists of an author-centric framework (*A*) and an external-agent framework (*E*). The author-centric framework identifies efforts made by authors to enhance the transparency and accessibility of their papers and is composed of three categories. The external-agent framework characterizes the success of external reviewers' efforts to reproduce a paper and is composed of four categories.

In summary, our contributions are: First, we present a novel approach to characterize reproducible research. Second, we analyze various features extracted from the text and metadata of papers to understand their relevance to reproducibility. Finally, we build an interpretable model for predicting how reproducible a paper might be. Unlike the current ad-hoc method of assigning subjective scores by reviewers, our approach is more systematic and data-driven.

We acknowledge the ethical and moral implications of utilizing a predictive model to assist in evaluating the quality and reproducibility of research papers. However, our goal in this study is to provide empirical evidence to support the use of such models and to identify crucial aspects influencing a paper's reproducibility assessment. We envision that the results of these models will complement and support reviewers in navigating the landscape of reproducible research rather than replacing human judgment. The code, methods, and artifacts for our study are publicly available at the following link: [1].

2 Background and Related work

Researchers from the University of Arizona [10, 11] analyzed data on computer systems research in an attempt to measure and understand reproducibility. Although these efforts didn't generate a conclusive hypothesis, they were instrumental in initiating a process to observe the willingness of computer science researchers to share code and data. Examining the conflicting attitudes of researchers towards reproducibility [4] provided insights into the frequency of successful and unsuccessful replications at both individual and disciplinary levels. The scholarly community acknowledged the reproducibility crisis, and there has been momentum for initiatives such as creating a manifesto on reproducibility [21] and estimating reproducibility rates [9].

Reproducibility has been formalized and recognized by various players involved in the scholarly publication process such as publishers, conferences, and peer reviewers. This recognition led to the establishment of funding programs such as DARPA's SCORE (Systematizing Confidence in Open Research and Evidence), which encourages researchers to develop assessment strategies to measure replication and reproduction efforts that are central to the scientific process. Additionally, many organizations introduced reproducibility checklists, most prominently ACM's rollout of Artifact Review and Badging [2] to address reproducibility and enhance research integrity across computational disciplines.

Literature that aligns with our goals for measuring and estimating reproducibility includes terminology papers [12, 14, 15, 28], statistical studies quantifying factors influencing reproducibility [26, 33], and predictive modeling studies [27, 29, 30, 32]. While these studies set an appropriate foundation, they fall short in one or more aspects to be considered conclusive in identifying reproducible works preemptively. These limitations include:

(1) **Lack of comprehensive methodology**: Most quantitative studies on reproducibility approach the analysis from a single perspective, often relying on correlations, statistical tests, predictive models, or user surveys. Identifying the reproducibility of a paper requires a comprehensive methodology capable of detecting a wide range of signals.
(2) **Potential impact on unseen data**: Understanding the reproducibility of scholarly works requires high standards of data curation. Given the limited number of works verified as reproducible, generalization becomes a challenge. It is crucial to outline the broader impact and limitations of the quantitative analysis on unseen data to validate the findings effectively.
(3) **Optimal balance on subjective vs. objective attributes**: Factors such as field of study, discipline, and venue significantly influence the structure of scientific research and the methodologies used in experiments. It is essential to strike an optimal balance between subjective and objective features when analyzing the causes of reproducible outcomes to ensure that findings about reproducibility are generalizable.

Given the significant challenges in gathering data on reproducibility, especially in computational science, our current study can serve as a primer for discussions on this topic. Building on related works [1, 26, 32], our study provides a comprehensive modeling approach to identify crucial aspects of papers that can predict whether it would be reproducible.

3 Building the Dataset

Our goal is to create a dataset that can be quantitatively analyzed in relation to artifacts and reproducibility. To achieve this, we collected papers from the ACM Digital Library because it is a singular comprehensive source with detailed information about the artifacts and reproducibility of scholarly articles. The ACM introduced the

[1] https://github.com/reproducibilityproject/NLRR/

[2] https://www.acm.org/publications/policies/artifact-review-badging

Artifact Reviewing and Badging policy, which assigns badges to indicate when publications have been successfully reproduced. These badges include:

(1) **Artifacts Available**: Assigned when papers include artifacts that have been made permanently retrievable.
(2) **Artifacts Evaluated and Functional** or **Artifacts Evaluated and Reusable**: Assigned when the artifacts have been reviewed and audited.
(3) **Results Reproduced**: Given when the primary findings of the publication have been validated and independently verified in a later investigation by a person or group other than the authors *without the use of author-supplied artifacts.*

3.1 Data Collection

Our data collection process involved the following steps:

(1) Using the ACM digital library advanced search endpoint [3] to list all scholarly articles in the ACM full-text collection that have received the *Results Reproduced* badge.
(2) Conducting separate searches using the same ACM digital library advanced search endpoint for articles with each of the following badges: *Artifacts Available*, *Artifacts Evaluated and Functional*, and *Artifacts Evaluated and Reusable*.
(3) Identifying the venues of articles with the "Results Reproduced" badge, and collecting unbadged articles from the same venues that were published in the same respective issue/year.

This resulted in an initial collection of just over three thousand badged articles. To maintain relevance, we included only papers published between 2016 and 2023, aligning with the timeframe when the ACM Badging policy was implemented. By filtering the samples based on full-text availability and publication date, we finalized a dataset of 2,659 articles. These articles were categorized as either *Artifacts Available*, *Artifacts Evaluated & Functional*, *Artifacts Evaluated & Reusable*, *Results Reproduced*, or *Unbadged*. *Unbadged* refers to papers from the same venues and years as *Results Reproduced* papers that were manually collected and included because the authors chose not to submit them for artifact & reproducibility evaluation.

The distribution of badges and the overlap between categories are illustrated in Fig. 1. Interestingly, many badged articles have multiple badge combinations. Fig. 1 shows that articles with the badges *Artifacts Available*, and *Artifacts Evaluated & Functional* have the largest intersection with 786 articles. In contrast, only 2 articles have all the badges. Furthermore, most reproducible articles tend to overlap with the *Artifacts Available* and *Artifacts Evaluated & Functional* categories. Noticeably, the Unbadged set has the highest unique category count, with 373 articles.

4 Reproducibility Spectrum

We introduce a joint spectrum for evaluating reproducibility in scientific papers as illustrated in Fig. 2. This spectrum is a result of a data-driven, iterative development process. Initially, our concept of the reproducibility spectrum categorized works as reproducible or non-reproducible. However, this simplistic approach failed to capture the nuances of scientific papers, as highlighted in Fig. 1.

[3] https://dl.acm.org/search/advanced

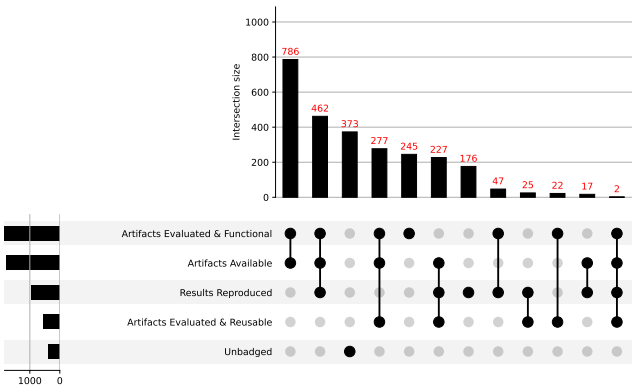

Figure 1: Visualization of badge category overlaps for the scholarly articles in our dataset.

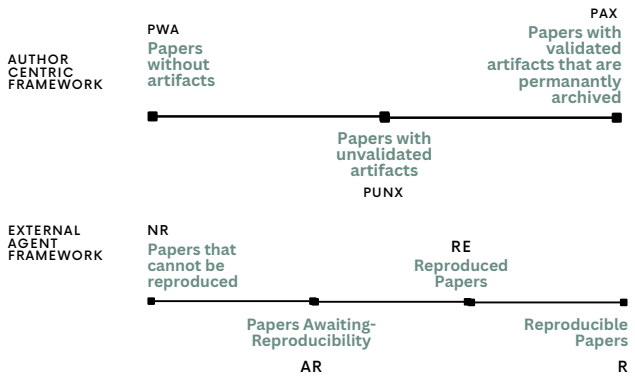

Figure 2: Joint framework to assess reproducibility levels in scientific papers.

Our data collection showed a much more complex landscape with interesting sub-categories of papers. This process revealed the importance of artifacts as a critical unit for assessing reproducibility. We finally constructed a version of the spectrum that is composed of an author-centric framework and an external agent framework. The author-centric framework focuses on the quality and availability of artifacts provided by the authors. It recognizes the varying degrees of effort authors put into making their work reproducible. The external-agent framework captures the external validation of a paper's reproducibility based on the available artifacts. By separating these aspects, we were able to represent the multifaceted nature of reproducibility in scientific publications.

4.1 Author-Centric Framework

The author-centric framework broadly captures the varying degrees of effort and commitment authors invest to facilitate reproducibility. The labels within this framework includes $\mathbf{A_i}$:

- $\mathbf{A_{PWA}}$: Papers without artifacts
- $\mathbf{A_{PUNX}}$: Papers with unvalidated artifacts
- $\mathbf{A_{PAX}}$ Papers with validated artifacts that are permanently archived

The key difference between $\mathbf{A_{PUNX}}$ and $\mathbf{A_{PAX}}$ is that $\mathbf{A_{PAX}}$ includes validation of the archived artifacts. While $\mathbf{A_{PUNX}}$ may include artifacts that are either archival or non-archival, the crucial difference between it and $\mathbf{A_{PAX}}$ is that they are not validated. Therefore, papers with validated artifacts that are permanently archived are considered the highest standard on our spectrum since they represent papers where authors took the most proactive measures to facilitate reproducibility evaluations.

We used the ACM badge (or the absence of one) to assign labels in the author-centric framework as follows:

- A_{PWA}: For all unbadged papers.
- A_{PUNX}: For papers with either the *Artifacts Available* or *Artifacts Evaluated & Functional* badge.
- A_{PAX}: For papers with the *Artifacts Evaluated & Reusable* badge, indicating the highest effort towards permanently archiving the paper's artifacts.

4.2 External Agent Framework

The external-agent framework presents the reproducibility evaluation status of a paper based on the information available for an independent team to assess and validate the original study's findings. This framework categorizes papers into the following units E_i on the spectrum:

- $\mathbf{E_{NR}}$: Papers that cannot be reproduced
- $\mathbf{E_{AR}}$: Papers awaiting-reproducibility
- $\mathbf{E_{Re}}$: Reproduced papers
- $\mathbf{E_{R}}$: Reproducible papers

There are several points to notice. First, E_{NR} papers lack any artifacts or supplemental information necessary for initiating reproducibility evaluation. Second, papers that are classified as Reproduced E_{Re} or Reproducible E_R have obtained their status through voluntary submission of artifacts to an evaluation committee. There is an important distinction between papers labeled E_{Re} and those labeled E_R, which is based on the archival nature of the artifacts and the reproducibility status. If a paper has $A_{PAX} \cap E_{Re}$, then it is considered E_R. In contrast, if a paper has been reproduced by any independent team, the assumption of its reproducibility status captured by E_{Re} is based on trust in the independent team's evaluation. In the external-agent framework, we assign labels as follows:

- E_{NR}: For all unbadged papers that cannot be reproduced due to a lack of available artifacts.
- E_{AR}: For papers that have artifacts but that have not yet been reproduced.
- E_{Re}: For *Results Reproduced* papers that do not have permanently archived artifacts.
- E_R: For papers that have both the *Results Reproduced* badge and the *Artifacts Evaluated & Reusable* badge.

Moving toward the rightmost end of either spectrum reflects a higher level of effort by the authors. At the same time, the ACM badges have interesting intersections as shown in Fig. 1. Specifically, a paper with a "Results Reproduced" badge need not have the artifacts available.

Table 1: Features with their respective categories.

Feature	Category
Number of Algorithms (X_1)	Structural
Number of Equations (X_2)	Structural
Google Scholar citations (X_3)	Scholarly
Availability of reproducibility checklist (X_4)	Venue
Mandatory artifact submission for papers (X_5)	Venue
Reproducibility Awards (X_6)	Venue
Author Correspondence for Reproducibility (X_7)	Venue
Mention of Zenodo Artifacts (X_8)	Artifact
Mention of GitHub Code Repository (X_9)	Artifact
Mention on Papers With Code GitHub Repository (X_{10})	Artifact
Mention on Papers With Code Datasets (X_{11})	Artifact
Mention on Papers With Code Methods (X_{12})	Artifact
Median Readability (X_{13})	Linguistic
Measure of lexical textual diversity (X_{14})	Linguistic
Availability of Funding source (X_{15})	Miscellaneous
Availability of Supplemental information (X_{16})	Miscellaneous

5 Pre-processing and Observations

Previous work [26] suggests including a wide range of both subjective and objective features to predict reproducibility, whereas the deep learning model from [32] focuses exclusively on the representational power of full-text embeddings. We selected a combination of Structural, Scholarly, Venue, Artifact, Linguistic, and Miscellaneous features, as detailed in Table 1. The Structural, Scholarly, and Linguistic features are numerical, whereas Venue and Miscellaneous features are categorical.

The metadata for each paper was collected from the ACM Digital Library website using a customized web scraper written in Python using the packages Selenium[4], and BeautifulSoup[5]. Additionally, we gathered complete metadata for all articles in our dataset using Allen AI's Academic Graph API (1.0)[6]. We utilized a similar web scraper to gather citations for each paper from Google Scholar, covering citations up to the end of 2023. To gather Miscellaneous and Venue features, we examined the individual article webpages. Miscellaneous features include details about funding and additional supplemental information such as videos, slides, and screen recordings. The Structural and Linguistic features were derived using the full texts of the article, which were processed by passing the PDFs through Allen AI's Science Parse[7].

"Readability" is a linguistic concept that measures how easily a reader can understand a written text. It considers the complexity of vocabulary, sentence structures, and overall text composition. The Median Readability was calculated in two steps. First, we used Python's Textstat [8] package to compute various readability metrics, including the Flesch Reading Ease Score, SMOG Index, Coleman-Liau Index, Automated Readability Index, Dale-Chall Readability Score, Linsear Write Formula, and Fog Scale (Gunning FOG Formula). Then, we calculated a weighted normalized score (ranging from 0 to 1) using the hypothetical minimum and maximum values for all these measures and took the median. Lexical diversity,

[4]https://pypi.org/project/selenium/
[5]https://pypi.org/project/beautifulsoup4/
[6]https://api.semanticscholar.org/api-docs/graph
[7]https://pypi.org/project/science-parse-api/
[8]https://pypi.org/project/textstat/

which reflects the variety and richness of the vocabulary used in a text, was quantified using the Measure of Textual Lexical Diversity (MTLD) [20]. As with readability, we employed the Textstat package to calculate this measure across the full text of each document.

Feature	Statistic	p-value
Median Readability	0.952613	1.565888e-28
Number of Algorithms	0.554016	1.731140e-63
Number of Equations	0.244451	8.599141e-74
Google Scholar citations	0.100468	2.239210e-77
Measure of lexical textual diversity	0.855591	4.103254e-44

Table 2: Shapiro-Wilk Test for assessing the normality of numerical features in scholarly papers.

We observed several interesting patterns building our dataset. First, 1.76% of articles have a dataset mentioned on PapersWithCode, and 1.01% reference a method on PapersWithCode. Additionally, 9.43% of articles have an official GitHub repository linked to the paper on PapersWithCode. This information was gathered by cross-referencing Arxiv IDs and paper titles with the PapersWithCode API [9]. Further textual analysis revealed that 41.03% of the articles mention a GitHub repository in the full text (excluding the "References" section). We also found that 16% of the articles reference Zenodo in the full text, pointing to artifacts related to the study. Moreover, 32.49% of the articles provide supplemental information such as code, audio, or video files on the ACM Digital Library. Finally, 50.1% of the articles mention funding sources, with the National Science Foundation, Engineering and Physical Sciences Research Council, and Deutsche Forschungsgemeinschaft being the most frequently cited agencies.

Feature	Statistic	p-value
Median Readability	4.990988	6.862850e-03
Number of Algorithms	36.773371	1.764600e-16
Number of Equations	5.258889	5.255375e-03
Google Scholar citations	1.714010	1.803412e-01
Measure of lexical textual diversity	1.290552	2.752913e-01

Table 3: Levene's Test for Homogeneity of Variances grouped by the author-centric framework.

6 Statistical tests

The foundation of our predictive modeling study is based on a statistical analysis of the numerical features $\mathbf{X}$ outlined in Table 1. This analysis involved conducting tests for normalization and variance of groups using the Shapiro-Wilk test and Levene's test, followed by a significance test using the Kruskal-Wallis test. Together, these tests ensure the statistical robustness of our feature set by verifying the assumptions of normality and homogeneity of variance, which are important for selecting appropriate predictive models. Additionally, these tests assisted us in discerning significant differences in features observed in both frameworks across different groups of scholarly papers. Finally, these tests guided our choices to pick predictive models that are well-suited to the data distribution.

[9]https://paperswithcode.com/api/v1/docs/

Feature	Statistic	p-value
Median Readability	4.153057	6.039707e-03
Number of Algorithms	29.537040	8.830013e-19
Number of Equations	6.959335	1.158253e-04
Google Scholar citations	4.195924	5.690491e-03
Measure of lexical textual diversity	0.283903	8.370575e-01

Table 4: Levene's Test for Homogeneity of Variances grouped by the external-agent framework.

The results of the Shapiro-Wilk test for assessing the normality of distributions and Levene's test for evaluating variance across groups are presented in Tables 2 to 4. The Shapiro-Wilk test results indicate that the p values from Table 2 are < 0.05, and we reject the null hypothesis that these features are normally distributed. This is an important observation to guide our choices in selecting predictive models such as Random Forest and Decision Trees. Tree-based models perform well in utilizing non-normal features with inequalities in variance when predicting the target variable. This can be evidenced from our results when comparing models built with the feature set $\mathbf{X}$ both in Table 7, and Table 8. Additionally, this suggests that parametric models like Multi-layer Perceptrons or Logistic Regression would only be advantageous if feature scaling is applied ($\mathbf{X_{scaled}}$) to normalize the features.

Feature	Statistic	p-value
Median Readability	693.261011	2.885920e-151
Number of Algorithms	43.248067	4.062576e-10
Number of Equations	15.267781	4.837751e-04
Google Scholar citations	35.751811	1.724221e-08
Measure of lexical textual diversity	94.078257	3.725342e-21

Table 5: Kruskal-Wallis test on the author-centric framework.

The results from Levene's test for homogeneity of variances in the author-centric framework Table 3, and the external-agent framework Table 4 indicate that all features, except lexical diversity, show statistically significant differences in the non-homogenous nature of features across groups. The significant results from Levene's test in both frameworks for several features (particularly readability, algorithms, and equations) suggest that these features differ not just in their average values but also in their variability among different categories of papers. This could have implications for how these features influence the artifact and reproducibility assessments in scholarly papers in our dataset.

Feature	Statistic	p-value
Median Readability	697.771459	6.386612e-151
Number of Algorithms	54.607980	8.324174e-12
Number of Equations	28.063838	3.521685e-06
Google Scholar citations	142.053160	1.363764e-30
Measure of lexical textual diversity	108.002775	2.952022e-23

Table 6: Kruskal-Wallis test on external-agent framework.

We used significance tests such as the Kruskal-Wallis test to make statistical inferences about the variability of feature values across

papers grouped by the author-centric and external-agent frameworks. Since our numerical features are not normally distributed, it is suitable to use a non-parametric test like Kruskal-Wallis. The results from Table 5 and Table 6 indicate significant differences for all the numerical features across groups of papers in both frameworks. The low p values (< 0.05) suggest that these features are valuable for predictive models, as their variability can help distinguish papers from different parts of the spectrum. In summary, these results support our intuition that structural, linguistic, and scholarly features are useful for predicting artifact quality and reproducibility assessment status.

7 Predictive Models

Our goal is to build interpretable predictive models to estimate the reproducibility of scientific research. We develop two distinct multi-class predictive models, ϕ_{author} and $\phi_{external}$, to predict the label (e.g., E_R) of a paper in the author-centric and external-agent frameworks. We experimented with several predictive models. The results from the Shapiro test in Table 2 indicated that tree-based models such as Gradient Boosting, AdaBoost, Random Forest, and Decision Tree algorithms were more suitable. Non-parametric models such as Logistic Regression and Neural Networks were also used after applying a simple feature scaling technique using the mean and standard deviation.

The remarkable effectiveness of feature representations from large language model embeddings cannot be overstated. By using document representations from text-embedding models such as Davinci from OpenAI, and SPECTER and Longformer from AllenAI, we can capture the full semantic context of scholarly texts. Since scholarly documents often exceed the maximum sequence length allowed by these models, we split the documents and took the average of the embeddings as the final representation. We used two models for these representations: 1. A **VanillaNN**, which is a linear classifier, and 2. An **MLP** (multi-layer perceptron) with a hidden layer.

7.1 Results for Author-Centric Framework

We evaluate the effectiveness of our predictive models for the author-centric framework labels using classification metrics such as accuracy and F1 scores. The results are presented in Table 7. As mentioned in Section 4, the ϕ_{author} models predict one of three labels: A_{PWA} (papers without artifacts), A_{PUNX} (papers with artifacts that aren't permanently archived), and A_{PAX} (papers with artifacts that are permanently archived). While it might seem that extracting artifact locations from paper texts would make a predictive model unnecessary, our experiments show that features designed to extract such information are not the best predictors. This highlights that predicting artifact availability or quality is a more challenging task than it appears.

The tree-based models, including Gradient Boosting, AdaBoost, Random Forest, and Decision Tree, demonstrate strong performance on the original feature set $\mathbf{X}$, with accuracy scores ranging from 78% to 83% and macro-averaged F1 scores between 66 % and 74 %. These results demonstrate the effectiveness of machine learning algorithms in distinguishing between papers with different levels of artifact availability, which is a critical aspect of reproducibility. In particular, the high F1 scores for A_{PUNX} and A_{PWA} indicate that these models are able to accurately differentiate between papers with and without permanently archived artifacts. On the other hand, non-parametric models like Logistic Regression and VanillaNN applied to the scaled feature set $\mathbf{X}_{\text{scaled}}$ show relatively weaker performance, which may be attributed to the loss of information during feature scaling. Finally, models leveraging text embeddings show promising results, particularly the MLP model with the ADA-002 embeddings, which achieves an accuracy score of 85% and a macro-averaged F1 score of 77%.

7.2 Results for External-Agent Framework

The results for models predicting the external agent framework labels are presented in Table 8. As mentioned in Section 4, the models here predict one of four labels: E_{NR} (papers that cannot be reproduced), E_{AR} (papers awaiting reproducibility), E_{Re} (reproduced papers), and E_R (reproducible papers). Overall, the best-performing model is an MLP that uses Longformer embeddings, which achieved the highest accuracy of 79%, along with comparably high F1 overall scores, and individual class-specific scores. However, parametric models that used scaled features $\mathbf{X}_{\mathbf{scaled}}$ demonstrated minimal predictive advantage of representational learning models.

The tree-based models, such as Gradient Boosting, Random Forest, and Decision Tree, continued to perform well with accuracy scores ranging from 69% to 75% and macro-averaged F1 scores between 67% and 72%. Although these models are effective in distinguishing between papers that cannot be reproduced (E_{NR}) and papers awaiting reproducibility (E_{AR}), improvements are needed in predicting reproduced E_{Re} and reproducible E_R papers. The key takeaway from Table 8 is the superior performance of models using embeddings ($\mathbf{X}_{\mathbf{emb}}$), particularly those based on Longformer and ADA-002, compared to both basic models ($\mathbf{X}$) and those using scaled features ($\mathbf{X}_{\mathbf{scaled}}$). Although this suggests that the semantic understanding provided by these embeddings is crucial in discerning subtle differences in paper statuses related to reproducibility, further investigation about reliability and robustness in predictions is necessary to fully understand model confidence (Section 7.4).

7.3 Important features for ϕ_{author} and $\phi_{external}$

One of the contributions of this study is the identification of features that correlate well with the reproducibility of a paper. As shown in Table 7 and Table 8, the Random Forest model consistently performs best in terms of both accuracy and overall F1 score across both frameworks. As a result, we selected this model for further analysis in the feature importance study. We collected the Gini impurity importance for all features in the Random Forest model (in both frameworks) and ranked them in Fig. 3. Linguistic measures such as readability and lexical diversity strongly influence the predictive outcomes of the models. Intuitively, clarity in language and thoroughness in explaining concepts (modeled through readability and lexical diversity features) should neither be correlated with the quality of artifacts nor should it affect the reproducibility status of a paper. However, the influence of these features on the predictive models, especially Random Forest, suggests otherwise. This surprising finding has also been observed in previous studies [13, 17, 19].

Model	Acc	$F_1(A_{PWA})$	$F_1(A_{PUNX})$	$F_1(A_{PAX})$	$F_1(macroavg)$	$F_1(weightedavg)$
X						
Gradient Boosting	0.83	0.82	0.89	0.52	0.74	0.82
AdaBoost	0.78	0.77	0.86	0.34	0.66	0.76
Random Forest	0.83	0.75	0.90	**0.57**	0.74	0.82
Decision Tree	0.79	0.74	0.87	0.53	0.71	0.79
$\mathbf{X}_{\text{scaled}}$						
Logistic Regression	0.71	0.14	0.84	0.37	0.45	0.66
VanillaNN	0.78	0.66	0.86	0.54	0.69	0.78
$\mathbf{X}_{\text{emb}}$						
SimpleNN - $\mathbf{X}_{\text{emb(ADA-002)}}$	0.80	0.76	0.86	0.36	0.67	0.77
SimpleNN - $\mathbf{X}_{\text{emb(SPECTER)}}$	0.68	0.32	0.83	0.26	0.47	0.65
SimpleNN - $\mathbf{X}_{\text{emb(Longformer)}}$	0.83	0.97	0.89	0.08	0.65	0.67
MLP - $\mathbf{X}_{\text{emb(ADA-002)}}$	0.81	0.83	0.88	0.51	0.74	0.81
MLP - $\mathbf{X}_{\text{emb(SPECTER)}}$	0.68	0.29	0.82	0.33	0.48	0.66
MLP - $\mathbf{X}_{\text{emb(Longformer)}}$	**0.85**	**0.97**	**0.90**	0.43	**0.77**	**0.83**

Table 7: Evaluation metrics for models predicting the author-centric framework labels.

Model	Acc	$F_1(E_{NR})$	$F_1(E_{AR})$	$F_1(E_{Re})$	$F_1(E_R)$	$F_1(macroavg)$	$F_1(weightedavg)$
X							
Gradient Boosting	0.73	0.81	0.78	0.60	0.66	0.71	0.73
AdaBoost	0.57	0.72	0.59	0.24	0.60	0.54	0.59
Random Forest	0.75	0.74	0.81	**0.63**	0.68	0.72	0.75
Decision Tree	0.69	0.77	0.76	0.57	0.58	0.67	0.69
$\mathbf{X}_{\text{scaled}}$							
Logistic Regression	0.55	0.07	0.66	0.15	0.53	0.35	0.50
VanillaNN	0.70	0.69	0.78	0.60	0.59	0.66	0.70
$\mathbf{X}_{\text{emb}}$							
SimpleNN - $\mathbf{X}_{\text{emb(ADA-002)}}$	0.75	0.79	0.81	0.44	0.68	0.68	0.74
SimpleNN - $\mathbf{X}_{\text{emb(SPECTER)}}$	0.57	0.30	0.70	0.38	0.54	0.48	0.57
SimpleNN - $\mathbf{X}_{\text{emb(Longformer)}}$	0.73	0.97	0.77	0.13	0.59	0.62	0.70
MLP - $\mathbf{X}_{\text{emb(ADA-002)}}$	0.74	0.83	0.81	0.52	0.63	0.70	0.74
MLP - $\mathbf{X}_{\text{emb(SPECTER)}}$	0.54	0.35	0.68	0.40	0.47	0.47	0.55
MLP - $\mathbf{X}_{\text{emb(Longformer)}}$	**0.79**	**0.97**	**0.82**	0.60	**0.70**	**0.77**	**0.79**

Table 8: Evaluation metrics for models predicting the external agent labels.

Among the top five features, we also observe the importance of citations and other venue-based features in both models. Citations act as a latent variable connecting a scholarly paper's impact and credibility. Highly cited papers might be considered more reproducible due to peer validation, but results from [32] suggest there is more room for introspection. The justification for having venue-based features, such as Reproducibility Awards, is to assess if such a variable serves the purpose of motivating authors to put more effort into making the artifacts available and consequently voluntarily opting in for reproducibility evaluation. Other categorical features that measure connections to references of supplemental information either within a paper or external sites such as Zenodo, Github, and PapersWithCode appear to have relatively lower rankings. Direct references to repositories where code and artifacts are stored are expected to be significant, given their role in facilitating artifact evaluation and reproducibility. However, the lower Gini importance suggests that additional factors are influencing the outcomes. Further research and experimentation are needed to uncover more latent variables within both frameworks.

7.4 Model confidence for ϕ_{author} and $\phi_{external}$

Understanding the confidence of predictive models is critical for establishing reliability. The confidence calibration curves for our models in are shown in Figs. 4 and 5. The bigger plots on the left show model confidence curves with mean predicted probabilities $\hat{p}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \hat{p}_{ik}$ on the x axis, and fraction of positives observed through an indicator function $\mathbb{I}(y_{test,i} = k)$ on the y-axis, which evaluates whether the predicted category k aligns with the actual category of each paper. In other words, these plots visualize the fraction of papers correctly identified within each category as a function of the predicted probabilities, allowing us to assess the calibration of the models across different categories of papers. The smaller plots on the right side of the confidence curves are histograms that show the overall distribution of predictive probabilities for each category of papers. These plots are useful for understanding the distribution of confidence the models have in their predictions.

Fig. 4 suggests that in the author-centric framework, a Random Forest model is reliable *only* when it predicts if papers have permanently archived artifacts. Also, the mean predicted probabilities in the range 0.2–0.6 suggest it is not confident in predicting A_{PWA}, or

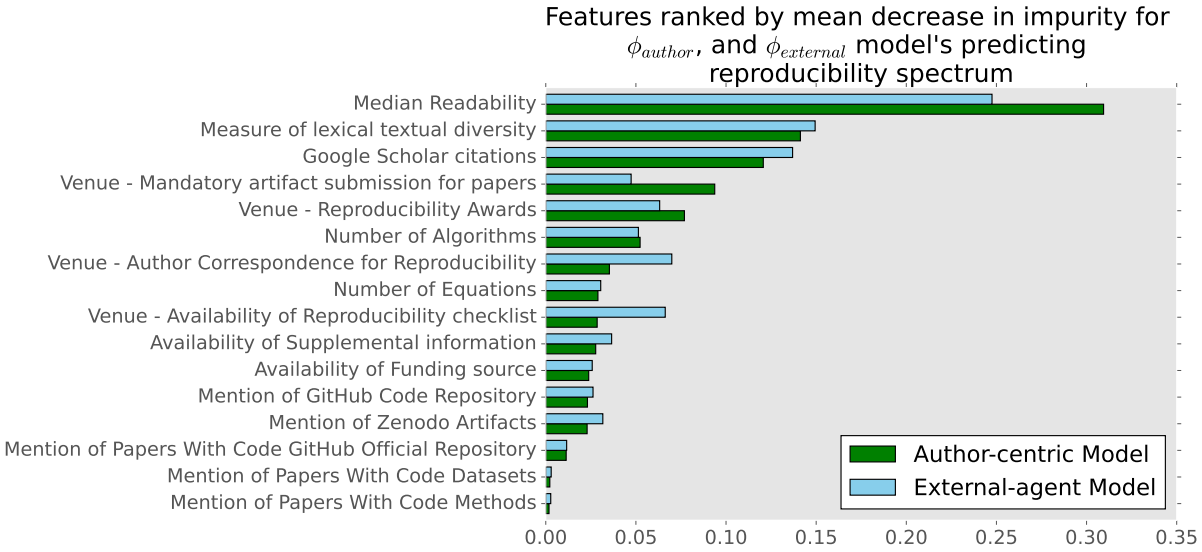

Figure 3: Most important features for predicting labels in the author-centric, and external-agent frameworks.

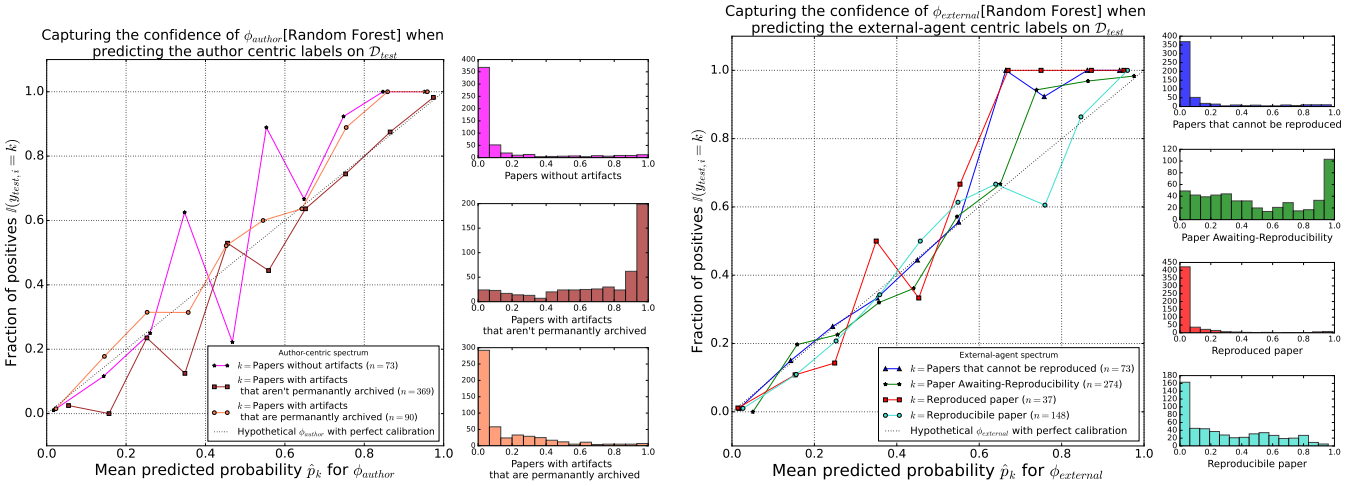

Figure 4: Confidence calibration of ϕ_{author} Random Forest model, author-centric framework (left) and ϕ_{external} Random Forest model, external-agent centric framework (right).

A_{PUNX}. The Longformer model (*MLP with longformer*, Fig. 5) shows a weakness in reliability compared to the Random Forest model. It shows consistent under- or overconfidence across the author-centric labels, especially at higher probabilities for A_{PWA}. Most importantly, Fig. 5 suggests that despite its effectiveness in evaluation metrics, the Longformer model is less effective at assessing papers without artifacts, potentially due to a lack of distinguishing features in the embeddings.

In the external-agent framework, for papers that cannot be reproduced (E_{NR}), we notice that Longformer model (Fig. 5) is extremely under confident, predicting lower probabilities than the actual outcomes. Additionally, the confidence of the Longformer, when predicting papers awaiting reproducibility (E_{AR}), or Reproduced (E_{Re}), or Reproducible (E_R) papers, is variable, especially at higher probabilities, suggesting slight inconsistencies in predictive robustness, and reliability. The Random Forest model, on the other hand (Fig. 4) shows a better alignment in predictive probabilities against the fraction of positives for E_{NR}, E_{AR}, E_{Re}, and E_R. This suggests the Random Forest model is better when compared to an MLP with Longformer representations, specifically when we talk about reliability, robustness, and consistency of the labels predicted across both frameworks. The histograms corroborate the reliability curves, indicating that the Random forest model not only predicts with high confidence but also aligns these predictions closely with

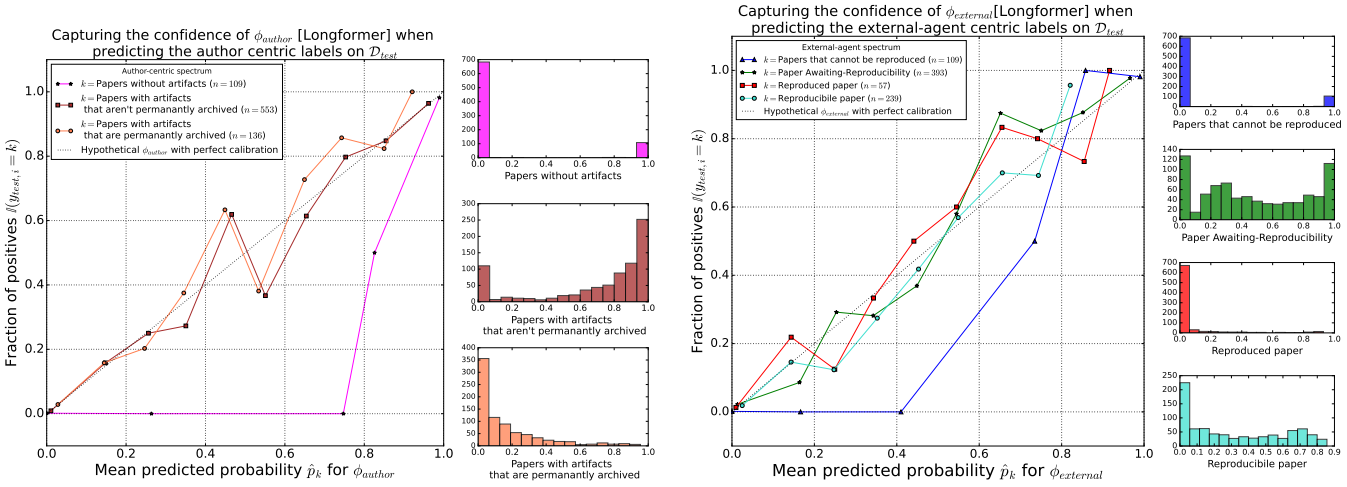

Figure 5: Confidence calibration of ϕ_{author} Longformer-MLP model, author-centric framework (left) and ϕ_{external} Longformer-MLP model, external-agent centric framework (right).

the actual outcomes, which is critical for downstream applications using predictive models for analyzing reproducibility.

8 Conclusion & Future Work

We define a spectrum to assess the reproducibility of scientific papers, collect a new dataset, and establish a framework for automatic prediction of the reproducibility of scientific papers. Our work presents a thorough analysis of predictive models that include feature importance tests and confidence calibration curves. We draw two surprising conclusions: 1. Linguistic features such as readability and lexical diversity are strong predictors for both the quality of artifacts mentioned in a paper and their reproducibility status, and 2. Neural nets built on text embeddings from large language models are accurate but not robust.

This work can be improved and extended in various ways. The predictive models can be improved, and the Neural nets can be made more robust. The unreasonable effectiveness of linguistic features can be investigated. Using a model or algorithmically-driven intelligent system to reward "reproducible" research practices, however, can be problematic, and we must have foresight in developing an approach toward quantifying reproducibility to avoid potential ethical problems. For example, suppose a model or system finds that the language of a paper positively affects its likelihood to be reproducible. It may thus penalize research simply because of the language in which the paper is written. Similarly, a model or system could identify institutions it associates with more reproducible results. Then, papers submitted from that institution might be labeled by the model as reproducible, without considering their content. Certainly, these are not outcomes we would expect or desire of such an algorithm or model.

Code and data artifacts are critical for reproducibility evaluation, and papers without artifacts and papers that cannot be reproduced represent a sizeable portion of scientific literature. While it can be argued that features such as the Number of Algorithms, Equations, and Reproducibility checklists are aligned more toward ACM's Badging policy, the foundational principles of reproducibility are universal and not exclusive to ACM. The structure of computational science adopted by most researchers involves artifacts. These artifacts, when made available, enable other researchers to verify, build upon, and extend the original work. This process of verification and extension, facilitated by accessible artifacts, creates a pathway for more generalizable findings. Utilizing our spectrum through the author-centric and external-agent framework for a larger multi-disciplinary study will offer valuable insights into the broader landscape of scientific research reproducibility.

Limitations: Generalizing the findings of our study to other disciplines is both data-intensive and challenging. While it is true that the composition of the ACM dataset and predictive modeling experiments cater to a specific category of computational science papers, the heuristics used to create the joint spectrum for reproducibility and the catalog of experiments we presented show a tangible pathway for expanding the study across other scientific disciplines. Despite the limitations, our work offers robust findings across the experiments, affirming the importance of "readability" for reproducibility.

Acknowledgement

This work is supported in part by NSF Grant No. 2022443. The experiments involved in the study were run on Google Cloud, and compute is supported by the Google Cloud Research Credits program with the award 274000118.

References

[1] Akhil Pandey Akella, Hamed Alhoori, and David Koop. 2022. Reproducibility Signals in Science: A preliminary analysis. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. 140–144.

[2] Jaime Arguello, Matt Crane, Fernando Diaz, Jimmy Lin, and Andrew Trotman. 2016. Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). In *ACM SIGIR Forum*, Vol. 49. ACM New York, NY, USA, 107–116.

[3] Vaibhav Bajpai, Mirja Kühlewind, Jörg Ott, Jürgen Schönwälder, Anna Sperotto, and Brian Trammell. 2017. Challenges with reproducibility. In *Proceedings of the Reproducibility Workshop*. 1–4.

[4] Monya Baker. 2016. Reproducibility crisis. *Nature* 533, 26 (2016), 353–66.

[5] Ronald F Boisvert. 2016. Incentivizing reproducibility. *Commun. ACM* 59, 10 (2016), 5–5.

[6] Fernando Chirigati, Rémi Rampin, Dennis Shasha, and Juliana Freire. 2016. Reprozip: Computational reproducibility with ease. In *Proceedings of the 2016 international conference on management of data*. 2085–2088.

[7] Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu. 2019. The SIGIR 2019 open-source IR replicability challenge (OSIRRC 2019). In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 1432–1434.

[8] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (2020), 70–79.

[9] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[10] Christian Collberg, Todd Proebsting, and Alex M Warren. 2015. Repeatability and benefaction in computer systems research. *University of Arizona TR* 14, 4 (2015), 1–68.

[11] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (feb 2016), 62–69. https://doi.org/10.1145/2812803

[12] Daniele Fanelli. 2018. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2628–2631.

[13] Marco Guerini, Alberto Pepe, and Bruno Lepri. 2012. Do linguistic style and readability of scientific abstracts affect their virality?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. 475–478.

[14] Odd Erik Gundersen. 2021. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A* 379, 2197 (2021), 20200210.

[15] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[16] Peter Ivie and Douglas Thain. 2018. Reproducibility in scientific computing. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–36.

[17] Tan Jin, Huiqiong Duan, Xiaofei Lu, Jing Ni, and Kai Guo. 2021. Do research articles with more readable abstracts receive higher online attention? Evidence from Science. *Scientometrics* 126 (2021), 8471–8490.

[18] Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. 2016. Toward reproducible baselines: The open-source IR reproducibility challenge. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*. Springer, 408–420.

[19] Ansel MacLaughlin, John Wihbey, and David Smith. 2018. Predicting news coverage of scientific articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[20] Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42, 2 (2010), 381–392.

[21] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9.

[22] Rochana R Obadage, Sarah M Rajtmajer, and Jian Wu. 2024. Can citations tell us about a paper's reproducibility? A case study of machine learning papers. *arXiv preprint arXiv:2405.03977* (2024).

[23] National Academies of Sciences, Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics, Board on Mathematical Sciences, et al. 2019. *Reproducibility and replicability in science*. National Academies Press.

[24] Mateusz Pawlik, Thomas Hütter, Daniel Kocher, Willi Mann, and Nikolaus Augsten. 2019. A link is not enough–reproducibility of data. *Datenbank-Spektrum* 19 (2019), 107–115.

[25] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research* 22, 164 (2021), 1–20.

[26] Edward Raff. 2019. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/c429429bf1f2af051f2021dc92a8ebea-Paper.pdf

[27] Edward Raff. 2023. Does the Market of Citations Reward Reproducible Work?. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*. 89–96.

[28] Edward Raff and Andrew L Farris. 2023. A siren song of open source reproducibility, examples from machine learning. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*. 115–120.

[29] Sarah Rajtmajer, Christopher Griffin, Jian Wu, Robert Fraleigh, Laxmaan Balaji, Anna Squicciarini, Anthony Kwasnica, David Pennock, Michael McLaughlin, Timothy Fritton, et al. 2022. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13218–13220.

[30] Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A Ingram, Edward A Fox, Sarah M Rajtmajer, and C Lee Giles. 2022. A study of computational reproducibility using urls linking to open access datasets and software. In *Companion Proceedings of the Web Conference 2022*. 784–788.

[31] Timothy E Sweeney, Winston A Haynes, Francesco Vallania, John P Ioannidis, and Purvesh Khatri. 2017. Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic acids research* 45, 1 (2017), e1–e1.

[32] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 20 (2020), 10762–10768.

[33] Burak Yildiz, Hayley Hung, Jesse H Krijthe, Cynthia CS Liem, Marco Loog, Gosia Migut, Frans A Oliehoek, Annibale Panichella, Przemysław Pawełczak, Stjepan Picek, et al. 2021. ReproducedPapers. org: Openly teaching and structuring machine learning reproducibility. In *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 3–11.