Code Contribution and Scientific Authorship

Eva Maxfield Brown[a,*], Isaac Slaughter[a], Nicholas Weber[a]

[a] *University of Washington Information School,*

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

1. Introduction

Software has become integral to contemporary scientific research (Edwards et al., 2013; Mayernik et al., 2017). From scripts for analysis and visualization to infrastructure for data collection (Hasselbring et al., 2024), software and code are now essential components of modern research practice. Today, research software serves multiple critical functions: enabling reproducible experiments (Krafczyk et al., 2019; Trisovic et al., 2021), providing methodological documentation (Ram, 2013), and increasingly appearing alongside publications as an additional research artifact (Cao et al., 2023; Trujillo et al., 2022). Software's expanding role has changed how science is conducted, with many fields now depending on specialized tools to advance scientific knowledge production.

Despite the growing dependence on software, a significant recognition gap exists between software development and traditional scientific outputs. Software contributors frequently find their work relegated to acknowledgments sections rather than warranting authorship (Philippe et al., 2019), creating a disconnect between the value software provides and the credit its developers receive. This misalignment has consequences for scientific careers, as the lack of formal credit can significantly impact promotion and advancement within research organizations (Carver et al., 2022; Biagioli and Galison, 2014). The result is a system that increasingly relies on software while undervaluing those who create it, potentially discouraging talented developers from pursuing scientific careers and affecting the sustainability of research software development (Muna et al., 2016).

The scientific community has attempted to address this recognition gap through initiatives like the Contributor Roles Taxonomy (CRediT). CRediT supports more inclusive authorship criteria for specialized contribution roles, including software development (Brand et al., 2015; Li et al., 2023; Lu et al., 2019). Despite these efforts, current frameworks remain centered on traditional author lists and may perpetuate historical biases about what consitutes meaningful scientific contribution (Haeussler and Sauermann, 2013; Gøtzsche et al., 2007; Ni et al., 2021). Simultaneously, researchers have also advocated for software citation systems that better reflect the unique contributions of research software developers, though these systems often remain separate from formal authorship recognition (Merow et al., 2023; Westner et al., 2024; Katz et al., 2020).

*Corresponding author
Email address: `evamxb@uw.edu` (Eva Maxfield Brown)

The emergence of public code repositories alongside published research provides a unique opportunity to study this disconnect. Source code repositories maintain detailed records of who contributes what code and when, providing a window into the patterns of software development in scientific research. Using transaction histories from these repositories, we developed a predictive model that enables systematic matching between scientific article authors and source code developer accounts. Our predictive model enables us to create a dataset of almost 140,000 linked authors and their developer accounts to understand how code contribution is distributed across research teams and the implications of frequent code contribution on an individuals career.

We identify several patterns in the relationship between code contributions and scientific recognition by applying our predictive model across almost 140,000 paired research articles and repositories. Our analysis reveals that nearly 30% of articles have non-author code-contributors - individuals who helped create the software but received no formal authorship credit. We find that code-contributing authors are associated with modest increases in article-level impact metrics (~5.2% increase in citations per code-contributing author). However, these effects become statistically non-significant when controlling for domain, article type, and open access status. First authors are significantly more likely to be code contributors than other author positions across all conditions tested. We also document a negative relationship between coding frequency and scholarly impact: authors who contribute code more frequently show progressively lower h-indices than their non-coding peers, a pattern that persists when controlling for publication count, and author's most common author position, domain, and article type.

The primary contributions of this article are: (1) a predictive model for matching authors with developer account information that addresses challenges in identity resolution across platforms[1]; (2) a dataset of linked authors and developers for ~140,000 article-repository pairs, providing a resource for analyzing scientific software development patterns[2]; and (3) analyses that reveal insights into the software development dynamics of research teams, including patterns of recognition, impact, and career implications for code contributors. These contributions provide evidence for the ongoing discussions about scientific recognition systems and raise questions about aligning institutional incentives with the spectrum of contributions that drive modern scientific progress.

The remainder of this paper proceeds as follows. First, we review related work regarding scientific software development and the recognition of code contributors in scientific credit systems. In addition, we introduce the specific hypotheses that guide our investigation. Next, we detail our data and methods, describing how we created a dataset of linked article-repository pairs, trained and evaluated our predictive model for entity matching, and applied our model across each article-repository pair. We then present our data analysis, focusing on article-level dynamics before moving to individual-level patterns, formally accepting or rejecting each hypothesis based on our findings. We conclude by discussing the results, limitations of our work, and areas for future improvement.

2. Background

2.1. Research Team Composition and Scientific Impact

Except for foundational methodological shifts, scientific recognition systems have favored experimental and theoretical contributions more than their methodological counterparts, with experimental and theoretical articles receiving higher citation rates (Aksnes, 2006; Liu et al., 2023; Chen et al., 2024). However, computational methods have transformed researchers' work across all scientific disciplines. Modern scientific endeavors increasingly depend on sophisticated computational approaches, whether for processing large-scale

[1]Our predictive model is made available via Huggingface at https://huggingface.co/evamxb/dev-author-em-clf or via our Python package, sci-soft-models (https://github.com/evamaxfield/sci-soft-models/).

[2]Our dataset is made available via Harvard Dataverse at https://doi.org/10.7910/DVN/KPYVI1. Portions of the dataset are restricted to keep researcher and developer information private. The full dataset is available upon request via the Harvard Dataverse. For details on interacting with this dataset, see our documentation at https://github.com/evamaxfield/rs-graph.

experimental data, running complex simulations, or developing new methodological tools (Jin et al., 2015; Hampton et al., 2013; Edwards et al., 2013; Mayernik et al., 2017; Hasselbring et al., 2024).

The computational evolution in scientific practice intersects with established findings about team dynamics in both research and software development. Prior research has shown that larger and more diverse teams typically produce higher-impact scientific work (Franceschet and Costantini, 2010; Larivière et al., 2014), while in software engineering, larger development teams tend to create more reliable software with fewer defects (Wyss et al., 2023). These findings suggest that team size may be particularly important in scientific software development, where technical robustness and scientific innovation are crucial.

The unique characteristics of scientific software development - including implementing novel algorithms, requiring deep domain knowledge, and an increased emphasis on reproducibility (Muna et al., 2016; Howison and Herbsleb, 2013) - make team composition especially relevant. Larger development teams may enhance scientific impact through multiple mechanisms: they can produce more robust and generalizable software tools for methodological contributions while enabling more sophisticated computational analyses and larger-scale data processing for experimental work. Given these patterns in team dynamics, software development practices, and the computational transformation of scientific work, we propose:

H1: The number of individuals contributing code to a publication's associated repository positively correlates with the article's citation count.

2.2. Author Roles and Technical Contributions

Author positions in scientific publications signal specific roles and responsibilities, a relationship extensively studied through contribution role taxonomies like CRediT (Larivière et al., 2016). These studies reveal that first authors and corresponding authors, while occasionally the same individual (Chinchilla-Rodríguez et al., 2022), take on distinct responsibilities. Analyses of contribution patterns consistently show that software development, data analysis, and visualization tasks typically fall to first authors (Larivière et al., 2016; Júnior et al., 2016; Larivière et al., 2020; Sauermann and Haeussler, 2017). Meanwhile, corresponding authors, whether or not they are also first authors, often maintain responsibility for research artifacts' long-term sustainability and reuse, which we believe may include the maintenance and documentation of software tools.

Contribution records from source code repositories provide a unique method to verify these established contribution patterns. Given prior findings about the distribution of technical responsibilities within research teams, we expect these repository records to reflect similar patterns of engagement with software development:

H2a: First authors have higher code contribution rates than authors in other positions.

H2b: Corresponding authors have higher code contribution rates than non-corresponding authors.

2.3. Code Contribution and Individual Scientific Impact

Despite the increasingly central role of software in science, researchers who develop scientific software face persistent challenges in receiving formal scientific recognition for their contributions. Prior work has shown that software developers in research settings are often relegated to acknowledgment sections rather than receiving authorship credit, even when their technical contributions are fundamental to the research (Carver et al., 2022; Philippe et al., 2019).

The challenge of recognition is compounded by inconsistent practices in software citation. While researchers have tried to standardize software citation, actual citation practices remain highly variable across fields and journals (Lamprecht et al., 2020; Katz et al., 2020; Smith et al., 2016). This variability challenges researchers who maintain and update existing software packages. While creating entirely new software may lead to dedicated publications and citations, the ongoing work of maintaining, debugging, and extending existing software - often crucial for scientific progress - typically generates less visible scientific credit (Howison and Herbsleb, 2011, 2013).

These structural challenges in recognizing and citing software contributions suggest a potential misalignment between technical contributions and traditional scientific impact metrics. When researchers dedicate significant time to software development and maintenance, conventional bibliometric measures may not fully capture their contributions, regardless of the software's importance to the field. Whether through attribution practices that favor acknowledgments over authorship or citation patterns that undervalue maintenance work, multiple mechanisms could lead to lower traditional impact metrics for active code contributors. Based on these patterns in software recognition and citation, we hypothesize:

H3: The frequency with which individual researchers contribute code to their research projects is negatively correlated with their h-index.

3. Data and Methods

Our analysis examines the relationship between software development contributions and scientific credit attribution through a three-step process: (1) building a dataset of linked scientific articles and code repositories, (2) developing a predictive model to match article authors with developer accounts, and (3) analyzing patterns in these relationships.

Our dataset integrates article-repository pairs from four sources, each with explicit mechanisms for code sharing: the Journal of Open Source Software (JOSS) and SoftwareX require code repositories as part of publication, Papers with Code directly connects preprints with implementations, and Public Library of Science (PLOS) articles include mandatory data and code availability statements that we mined for repository links. We focused exclusively on GitHub-hosted repositories, which represent the predominant platform for scientific software sharing (Cao et al., 2023; Escamilla et al., 2022). For each article-repository pair, we resolved DOIs via Semantic Scholar to ensure we captured the latest version of each publication, extracted publication metadata and author metrics through OpenAlex, and collected repository information via the GitHub API. This created a comprehensive dataset with information about both the scientific and software development aspects of each project.

To match article authors with repository developer accounts, we developed a machine learning approach using transformer-based architectures. Entity matching presents unique challenges, as exact name or email matching is insufficient due to formatting variations and incomplete information (e.g., "J. Doe" vs. "Jane Doe" in publications, or use of institutional versus personal email addresses). While exact matching fails, there is typically high semantic overlap between author information and developer account details that our model can leverage. We created a gold-standard dataset of 3,000 annotated author-developer account pairs from JOSS publications, where two independent reviewers classified each pair as matching or non-matching. After systematic evaluation of three transformer architectures with various feature combinations, our optimal model (fine-tuned DeBERTa-v3-base including developer account username and display name in the training data) achieved an F1 score of 0.944, with 0.938 precision and 0.95 recall (detailed comparison available in the Appendix).

Applying our model across all article-repository pairs yielded a large-scale dataset linking scientific authors and code contributors. As shown in Table 1, this dataset contains approximately 140,000 article-repository pairs spanning multiple domains and publication types, with nearly 300,000 distinct authors and more than 150,000 developer accounts. From these, we identified almost 110,000 author-developer account relationships, creating a unique resource for investigating code contribution patterns in scientific teams. This dataset enables systematic examination of how software development work relates to scientific recognition and career trajectories. Both the dataset (https://doi.org/10.7910/DVN/KPYVI1) and model (https://huggingface.co/evamxb/dev-author-em-clf) are made publicly available to support further research.

Table 1: Counts of Article-Repository Pairs, Authors, and Developers by Data Sources, Domains, Document Types, and Access Status.

Category	**Subset**	**Article-Repository Pairs**	**Authors**	**Developers**
By Domain	Physical Sciences	116597	240536	130601
	Social Sciences	8839	29239	14023
	Life Sciences	7727	31613	12123
	Health Sciences	5176	26022	7277
By Document Type	preprint	72177	170301	87311
	research article	63528	173183	78935
	software article	2891	9294	12868
By Access Status	Open	132856	286874	147831
	Closed	5740	23668	9352
By Data Source	pwc	129615	262889	134926
	plos	6090	30233	8784
	joss	2336	7105	11362
	softwarex	555	2244	1628
Total		**138596**	**295806**	**152170**

4. Analysis of Code Contributor Authorship and Development Dynamics of Research Teams

4.1. Software Development Dynamics Within Research Teams

Understanding the composition and dynamics of software development teams provides essential context for analyzing how code contributions relate to scientific recognition and impact. To ensure reliable analysis, we focus on a subset of our article-repository pairs that meet several filtering conditions. First, we require that each article-repository pair have at least one citation, which helps ensure the research has received a basic level of engagement from the scientific community. Next, we require that repository commit activity must stop prior to 90 days past the date of article publication. Disallowing long-term projects ensures we do not include projects that may add additional code contributors later while still allowing a grace period during which developers can update repositories with additional documentation and publication information. We then subset the data to only include article-repository pairs with research teams of typical size by removing those with fewer than three authors and more than 12 authors, the 97th percentile for research team size. Finally, we filter out any author-developer pairs associated with these projects with predictive model confidence of less than 0.97 to ensure that we only include high-confidence matches[3]. This filtering process results in a dataset of 22804 article-repository pairs.

Within this filtered dataset, we categorized individuals into three groups: code-contributing authors (CC-A) who both authored papers and contributed code to associated repositories, non-code-contributing authors (NCC-A) who authored papers but showed no evidence of code contributions, and code-contributing non-authors (CC-NA) who contributed code but received no authorship recognition. This categorization revealed that papers in our dataset typically have 4.9 ± 1.9 total authors, with 1.0 ± 0.7 code-contributing authors and 3.9 ± 2.0 non-code-contributing authors. Beyond the author list, papers averaged 0.5 ± 1.7 code-contributing non-authors. Table 2 details these distributions by domain, article type, and open access status.

Perhaps most striking is our finding that 6586 papers (28.9%) have at least one code contributor who did not receive authorship recognition. Within this substantial subset of papers, we found an average of 1.6 ± 2.9

[3]Figure 3 shows the distribution of predictive model confidence scores for author-developer pairs to justify this threshold. We chose the 0.97 threshold to ensure that we only include high-confidence matches while retaining a large proportion of the data.

unrecognized code contributors per paper. On average, only one code-contributing author per paper aligns with previous research by Larivière et al. (2020), showing that technical tasks like data curation, formal analysis, visualization, and software development typically fall to first authors. However, our finding that over a quarter of papers have unrecognized code contributors suggests a more complex dynamic between software development and authorship recognition.

Table 2: Mean and Standard Deviation of Non-Code-Contributing Authors (NCC-A), Code-Contributing Authors (CC-A), and Code-Contributing Non-Authors (CC-NA) Research Team Members by Domain, Article Type, and Open Access Status. Only includes research teams from article-repository pairs with a most recent commit no later than 90 days after publication and excludes research teams in the top 3% of total author sizes.

Control	Subset	Total Authors	NCC-A	CC-A	CC-NA
OA Status	Closed	5.1 ± 1.9	4.1 ± 1.9	1.0 ± 0.7	0.6 ± 2.1
	Open	4.9 ± 1.9	3.9 ± 2.0	1.0 ± 0.7	0.4 ± 1.7
Domain	Health Sciences	6.1 ± 2.5	5.1 ± 2.6	0.9 ± 0.6	0.4 ± 1.2
	Life Sciences	5.2 ± 2.1	4.2 ± 2.2	1.0 ± 0.7	0.4 ± 1.2
	Physical Sciences	4.8 ± 1.8	3.8 ± 1.9	1.0 ± 0.7	0.5 ± 1.8
	Social Sciences	4.5 ± 1.7	3.5 ± 1.8	1.1 ± 0.7	0.3 ± 1.1
Article Type	preprint	4.8 ± 1.8	3.8 ± 1.9	1.0 ± 0.7	0.6 ± 2.2
	research article	4.9 ± 1.9	3.9 ± 2.0	1.0 ± 0.7	0.4 ± 1.6
	software article	4.7 ± 1.9	3.2 ± 1.9	1.5 ± 1.4	1.0 ± 1.1
Overall		**4.9 ± 1.9**	**3.9 ± 2.0**	**1.0 ± 0.7**	**0.5 ± 1.7**

When examining these patterns over time and across different team sizes (Figure 1), we found that the number of code-contributing authors and unrecognized contributors has remained relatively stable. This stability over time suggests that while the exclusion of code contributors from authorship is not worsening, it represents a persistent feature of scientific software development rather than a historical artifact or transition period in research practices. Similarly, the number of code-contributing non-authors remains constant even as team size grows, indicating that larger research teams do not necessarily adopt more inclusive authorship practices for code contributors, despite representing broader collaborative efforts.

4.1.1. Modeling Article Citations

Building upon previous work examining the effects of team size and team diversity on scientific impact and software quality (see Section 2), we investigate how the number of code contributors within a research team may be associated with an article's research impact. We hypothesized that more code contributors might signal greater technical complexity in research, which may be associated with higher citation counts as the community builds upon more technically sophisticated works.

Using our filtered dataset of article-repository pairs, we conducted multiple regression analyses to examine these relationships while controlling for various factors. Without controlling for domain, open access, or article type differences (Table 5), our analysis revealed a modest positive association between the number of code contributing authors and article citations, with each code-contributing author associated with a 5.2% increase in article citations ($p < 0.001$).

When controlling for article type (Table 8), we observed divergent patterns between preprints and research articles. For preprints, each code-contributing non-author was associated with a statistically significant 3.1% decrease in citations ($p < 0.005$). In contrast, research articles showed more positive associations: we found a significant positive relationship between code-contributing authors and citations ($p < 0.001$), though we cannot estimate the precise magnitude due to the non-significant main effect in the model. Additionally, each code-contributing non-author was associated with a 0.7% increase in expected citations for research articles ($p < 0.001$).

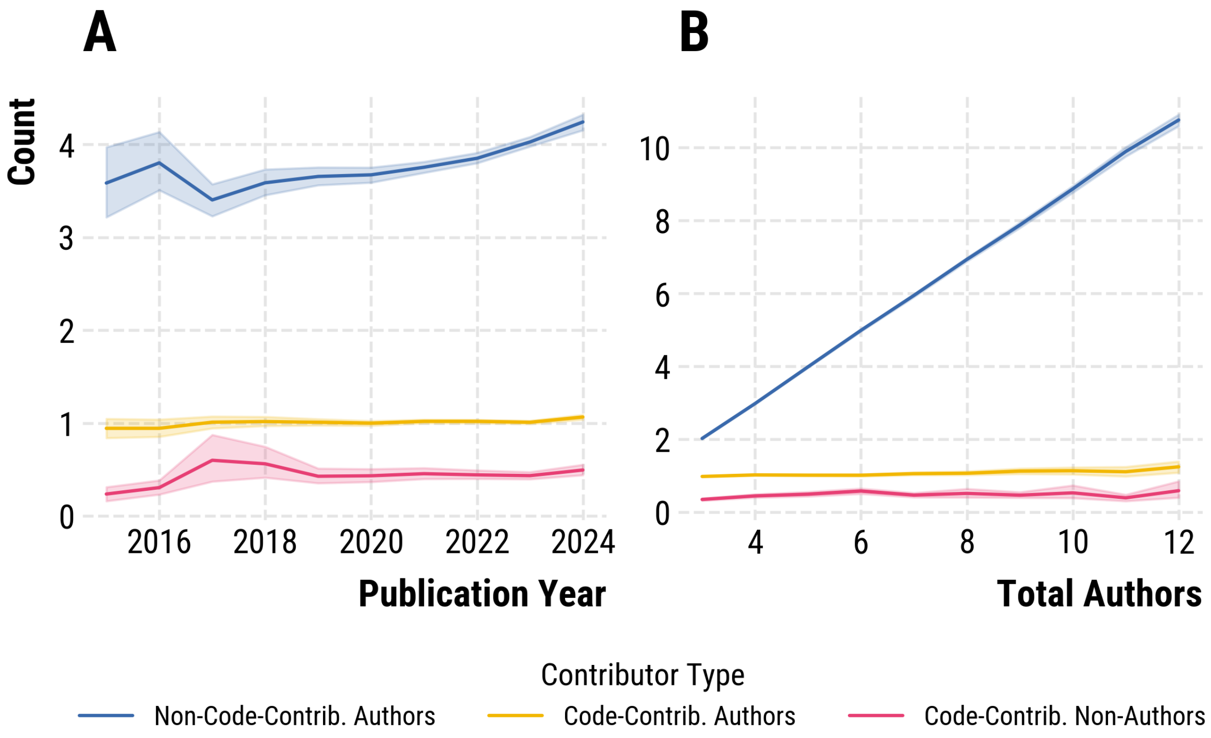

Figure 1: Average number of contributors per article, by contribution type along with A) the year the article was published, and B) the total number of authors included on the article. Only includes research teams from article-repository pairs with a most recent commit no later than 90 days after publication and excludes research teams in the top 3% of total author sizes for publication years with 50 or more articles. Shaded areas show the 95% confidence interval for the mean.

Based on these findings, we ***partially accept*** our hypothesis (*H1*) that "the number of individuals contributing code to a publication's associated repository positively correlates with the article's citation count." Several important nuances qualify this acceptance: the relationship is statistically significant but modest in magnitude and differs substantially between research articles (positive association) and preprints (negative association for non-author code contributors). These variations suggest that the relationship between code contributions and citation impact is context-dependent and more complex than initially hypothesized.

4.2. Characteristics of Scientific Code Contributors

4.2.1. Author Positions of Code Contributing Authors

Building upon previous work examining the relationship between authorship position and research contributions, we investigate how author position may relate to code contribution patterns. We hypothesized that first authors, traditionally contributing the bulk of intellectual and experimental work, are most likely to contribute code to a project. In contrast, middle and last authors often provide oversight and guidance and would be less likely to contribute code.

To analyze these patterns within our previously filtered dataset of article-repository pairs, we conducted Chi-square tests of independence between author position and code contribution status. These tests revealed significant associations between author position and likelihood of code contribution overall and when controlling for research domain, article type, and open access status (all $p < 0.01$), indicating that the proportion of authors contributing code differs significantly based on author position. Following these significant associations, we examined the specific proportions across positions (Table 9): 67.4% of first authors contributed code to their projects, compared to only 8.5% of middle authors and 8.7% of last authors. The differences in these proportions remained statistically significant across all tested scenarios, regardless of research domain, article type, or open access status.

Based on these findings, we ***accept*** our hypothesis (*H2a*) that "first authors have higher code contribution rates than authors in other positions." The data demonstrates that the proportion of first authors who contribute code (67.4%) is significantly higher than the proportion of both middle authors (8.5%) and last authors (8.7%). This relationship remains robust and statistically significant across all tested conditions, including variations in research domain, article type, and open access status, indicating a fundamental connection between authorship position and technical contribution in scientific research.

4.2.2. Corresponding Status of Code Contributing Authors

Building upon our analysis of author position, we next examine how corresponding author status relates to code contribution patterns. We hypothesized that corresponding authors, who traditionally maintain research artifacts and serve as primary points of contact, would be more likely to contribute code compared to non-corresponding authors, as this role often involves responsibility for project resources and materials.

To analyze these relationships within our filtered dataset of article-repository pairs, we conducted Chi-square tests of independence between corresponding author status and code contribution status. Our analysis revealed patterns contrary to our initial hypothesis. The proportion of code contributors was low among both groups, with only 29.5% of corresponding authors and 19.9% of non-corresponding authors contributing code to their projects. Further examination (Table 10) showed that this pattern holds across nearly all conditions, with only two notable exceptions: corresponding authors in software-focused articles and those in closed-access publications showed no significant difference in their proportion of code contributors compared to non-corresponding authors in those categories.

Based on these findings, we ***reject*** our hypothesis (*H2b*) that "corresponding authors have higher code contribution rates than non-corresponding authors." Contrary to our expectations, our analysis revealed that the proportion of code contributors among corresponding authors (29.5%) did not significantly differ from the proportion among non-corresponding authors (19.9%). This pattern of similar proportions remained consistent across most studied conditions, with only limited exceptions in software-focused articles and closed-access publications.

4.2.3. Modeling Author H-Index

Building upon previous work examining career implications for researchers who prioritize software development (see Section 2), we investigated how varying levels of code contribution relate to scholarly impact through h-index metrics. To ensure a robust analysis, we applied several key data filtering steps. We only included researchers with at least three publications in our dataset, removed those with more than three developer account associations, and used each researcher's most common domain, article type, and author position, with ties broken by the most recent occurrence. We removed h-index outliers by excluding researchers below the bottom 3rd and above the top 97th percentiles. Finally, we removed any author-developer-account pairs with a predictive model confidence of less than 0.97.

We categorized researchers' coding contributions into mutually exclusive groups: non-coders (no code contributions), any coding (code contribution in less than half of article-repository pairs), majority coding (code contribution in at least half, but not all, article-repository pairs), and always coding (code contribution in every article-repository pair).

Figure 2 shows the distribution of author h-indices across these coding frequency groups, grouped by author position, publication type, and research domain.

Our analysis revealed a consistent and statistically significant negative relationship between code contribution frequency and h-index across multiple analytical controls. Our initial uncontrolled analysis (Table 11) indicates increasingly adverse h-index effects as researcher coding frequency increases. Compared to non-coding authors, researchers were associated with progressively lower h-indices: occasional code contributors showed a ~27.3% lower h-index ($p < 0.001$), majority code contributors demonstrated a ~53.5% lower h-index ($p < 0.001$), and always coding authors exhibited a ~62.1% lower h-index ($p < 0.001$).

When controlling for author position (Table 12), we found a general pattern of reduced h-indices with increased code contribution, with one notable exception. Occasional coding first authors were associated with a ~14.9% higher h-index ($p < 0.001$), while always coding first authors saw a ~21.6% reduction compared to non-coding first authors ($p < 0.001$). For middle and last authors, the pattern was more consistently negative. Middle authors who occasionally coded showed a ~26.6% lower h-index ($p < 0.001$), and those always coding demonstrated a ~52.9% lower h-index ($p < 0.001$). Similarly, last authors who occasionally coded experienced a ~13.1% lower h-index ($p < 0.001$), with always coding authors showing a ~45.7% lower h-index ($p < 0.001$).

When controlling for research domain (Table 13), majority coding scientists showed significant h-index reductions across all domains. Health sciences researchers saw the most dramatic reduction at ~76.5% ($p < 0.001$), followed by physical sciences at ~52.6% ($p < 0.001$), social sciences at ~51.4% ($p < 0.001$), and life sciences at ~47.1% ($p < 0.001$).

Analyzing by common article type (Table 14) revealed similar patterns. For authors primarily publishing preprints, the h-index reductions were substantial: ~25.6% for occasional coding, ~53.5% for majority coding, and ~62.9% for always coding authors. Authors primarily publishing software articles showed slightly better but still significant reductions: ~33.1% for majority coding and ~33.0% for always coding authors.

Based on these findings, we ***accept*** our hypothesis (*H3*) that "the frequency with which individual researchers contribute code to their research projects is negatively correlated with their h-index." Our analysis demonstrates a clear and statistically significant negative relationship between coding frequency and scholarly impact as measured by the researcher's h-index. This relationship was robust across multiple analytical controls, including author position, research domain, and article type. These results are particularly striking because each of our models includes publication count as an input feature, suggesting that these h-index reductions persist even when accounting for total research output.

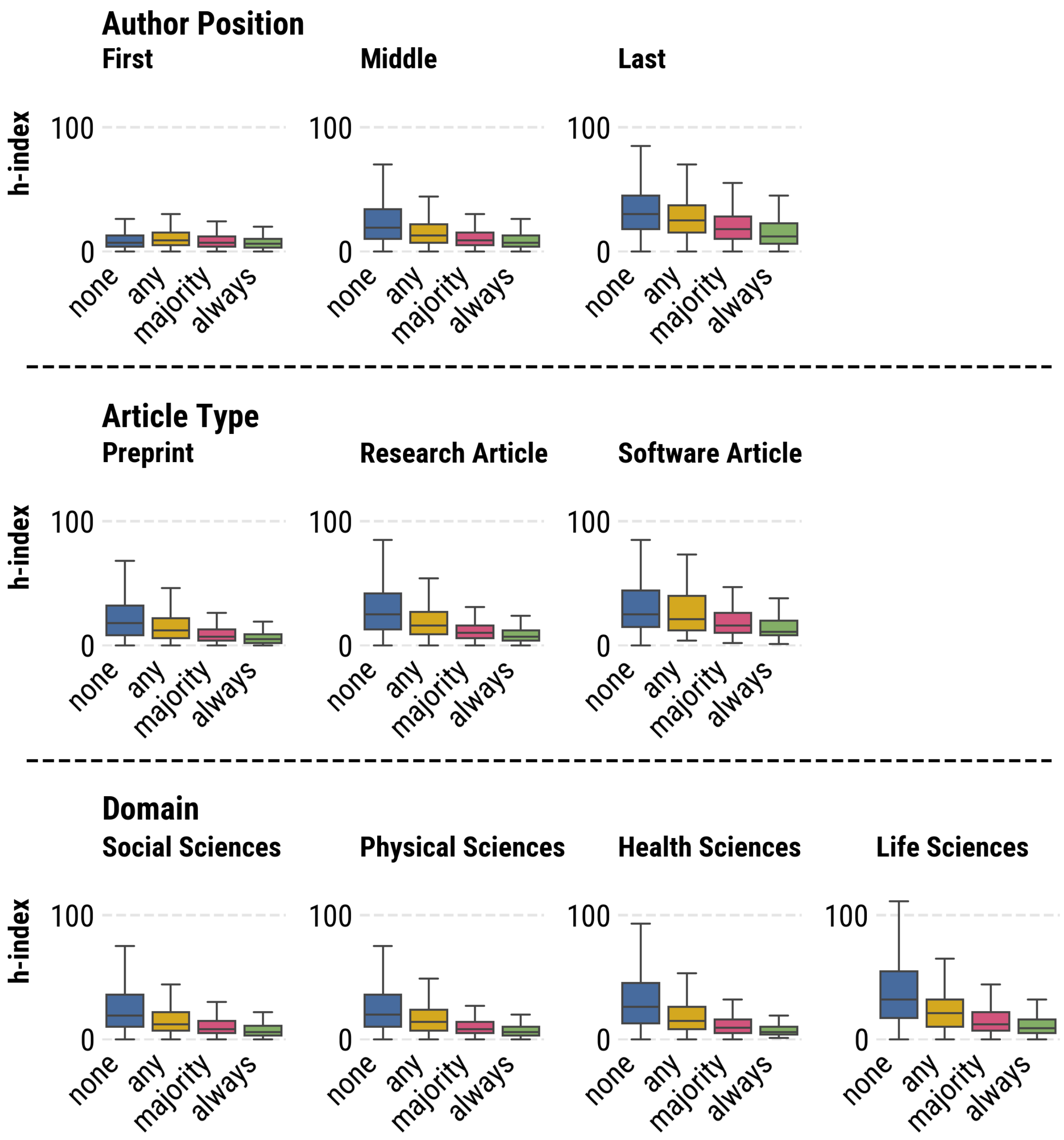

Figure 2: Distribution of author h-index by coding frequency across three key publication factors. Results are grouped by each author's most frequent: (1) position in publication bylines (first, middle, or last), (2) publication type (preprint, research article, or software article), and (3) research domain (Social Sciences, Physical Sciences, Health Sciences, or Life Sciences). Within each subplot, h-indices are divided by the author's coding frequency: 'none' (no coding in any of their publications), 'any' (coding in at least one but fewer than half of their publications), 'majority' (coding in at least half but not all of their publications), and 'always' (coding in each of their publications). Authors are only included if they have three or more publications within our dataset and are associated with no more than three developer accounts, with each association having a predicted model confidence of at least 97%.

5. Discussion

Our analysis reveals significant disparities in the recognition of software contributions to scientific research, with nearly 30% of articles having non-author code-contributors who received no formal authorship credit. This persistent pattern suggests a systemic disconnect between software development and scientific recognition systems, reflecting challenges in how scientific contributions are valued and credited. This exclusion reflects what Shapin (1989) observed about scientific authority—the selective attribution of technical work as either genuine knowledge or mere skill significantly impacts who receives formal recognition. These findings further support previous research by Philippe et al. (2019) and Carver et al. (2022) documenting the frequent relegation of software contributors to either acknowledgment sections or receiving no credit at all, rather than authorship positions, despite the increasingly central role of software in scientific inquiry. The stability of this pattern over time indicates that this phenomenon has embedded itself in scientific software development rather than representing a transitional phase, raising questions about scientific labor and how reward structures integrate technical contributions.

Our finding that, on average, article-repository pairs have only a single code contributor mirrors prior work from Färber (2020). Further, the distribution of code contributions across author positions provides context to the hierarchical organization of scientific work. First authors emerge as significantly more likely to contribute code with 67.4% of all first authors in our dataset contributing code. Middle and last authors, meanwhile, were statistically significantly less likely to contribute code, with only 8.5% of middle authors and 8.7% of last authors acting as code-contributing members of the research team. Corresponding authors were similarly less likely than expected to be code contributors, as we found that within our dataset, corresponding authors were code contributors 29.5% of the time. These patterns align with traditional scientific labor distribution where first authors typically handle technical aspects of research while middle and last authors are likely specialist contributors or provide guidance and oversight (Larivière et al., 2020; Sauermann and Haeussler, 2017). However, our data did not support our initial hypothesis that corresponding authors would also be more likely to contribute code due to their shared responsibility for the long-term maintenance of research artifacts. This finding suggests a potential strict division between project management responsibilities and direct technical engagement with software development.

The modest citation advantage associated with code-contributing authors (5.2% increase in citations per code-contributing author) stands in contrast with the significant negative relationship between coding frequency and an individual's scholarly impact (h-index). This misalignment between technical contributions and scientific recognition creates an asymmetrical relationship in which software development may enhance research impact but potentially penalizes individual careers. The progressive reduction in h-index as coding frequency increases indicates a cumulative disadvantage for frequent code contributors. This pattern persists even when controlling for publication count, suggesting issues in how software contributions are valued relative to other scientific outputs. These findings echo concerns raised by Muna et al. (2016) about the sustainability of research software development and highlight how current reward structures may discourage talented developers from pursuing scientific careers.

Software development represents a form of scholarly labor that has become increasingly essential to modern research yet remains incompletely integrated into formal recognition systems. Similar to the high proportion of articles with authors who made data curation contributions towards research observed by Larivière et al. (2020), our finding that a quarter of papers have unacknowledged code contributors highlights a labor role that is simultaneously common and undervalued. The prevalence of code contributions across domains demonstrates the importance of this work to contemporary research. However, the persistent exclusion of contributors from authorship suggests that researchers continue to classify software development as technical support rather than intellectual contribution. This classification may reflect disciplinary traditions that privilege certain forms of scholarly production despite the growing recognition that software represents a legitimate research output (Katz et al., 2020). The tension between software's importance and contributors' recognition status raises questions about how we define, value, and reward different forms of scientific labor in an increasingly computational research landscape.

5.1. Limitations

Our data collection approach introduces several methodological constraints that should be considered when interpreting these results. By focusing exclusively on GitHub repositories, we likely miss contributions stored on alternative platforms such as GitLab, Bitbucket, or institutional repositories, potentially skewing our understanding of contribution patterns. As Trujillo et al. (2022), Cao et al. (2023), and Escamilla et al. (2022) have all noted, while GitHub is the predominate host of scientific software, significant portions of research code exist on other platforms. Additionally, our reliance on public repositories means we cannot account for private repositories or code that were never publicly shared, potentially underrepresenting sensitive research areas or proprietary methods.

Our predictive modeling approach for matching authors with developer accounts presents additional limitations. The model's performance can be affected by shorter names where less textual information is available for matching, potentially creating biases against researchers from cultures with shorter naming conventions. Organization accounts used for project management pose particular challenges for accurate matching, and while we implemented filtering mechanisms to minimize their impact, some misclassifications may persist. Furthermore, our approach may not capture all code contributors if multiple individuals developed code. However, only one uploaded it to a repository, creating attribution artifacts that may systematically underrepresent specific contributors, particularly junior researchers or technical staff who may not have direct repository access.

Our analytical approach required substantial data filtering to ensure reliable results, introducing potential selection biases in our sample. By focusing on article-repository pairs with commit activity no later than 90 days past the date of article publication and at least three authors and less than 12 authors, we may have systematically excluded certain types of research projects, particularly those with extended development timelines or extensive collaborations. Our categorization of coding status (non-coder, any coding, majority coding, always coding) necessarily simplifies complex contribution patterns. It does not account for code contributions' quality, complexity, or significance. Additionally, our reliance on OpenAlex metadata introduces certain limitations to our analysis. While OpenAlex provides good overall coverage, it lags behind proprietary databases in indexing references and citations. The lag in OpenAlex data may affect our citation-based analyses and the completeness of author metadata used in our study (Alperin et al., 2024).

5.2. Future Work

Future technical improvements may enhance our understanding of the relationship between software development and scientific recognition systems. Expanding analysis beyond GitHub to include other code hosting platforms would provide a more comprehensive understanding of scientific software development practices across domains and institutional contexts. More sophisticated entity-matching techniques could improve author-developer account identification, particularly for cases with limited information or common names. Developing more nuanced measures of code contribution quality and significance beyond binary contribution identification would better capture the true impact of technical contributions to research. These methodological advances would enable more precise tracking of how code contributions translate—or fail to translate—into formal scientific recognition, providing clearer evidence for policy interventions.

Our findings point to several directions for future research on the changing nature of scientific labor and recognition. Longitudinal studies tracking how code contribution patterns affect career trajectories would provide valuable insights into the long-term impacts of the observed h-index disparities and whether these effects vary across career stages. Comparative analyses across different scientific domains could reveal discipline-specific norms and practices around software recognition, potentially identifying models that more equitably credit technical contributions. Qualitative studies examining how research teams make authorship decisions regarding code contributors would complement our quantitative findings by illuminating the social and organizational factors influencing recognition practices. Additionally, to better understand corresponding authors' role in maintaining research artifacts, future work could remove the 90-day post-publication commit activity filter to examine long-term sustainability actions. However, this approach would need to address the introduction of contributors unrelated to the original paper.

Despite their growing importance, the persistent underrecognition of software contributions suggests a need for structural interventions in how we conceptualize and reward scientific work. Building upon efforts like CRediT (Brand et al., 2015), future work should investigate potential policy changes to better align institutional incentives with the diverse spectrum of contributions that drive modern scientific progress. However, as the example of CRediT demonstrates, even well-intentioned taxonomies may reproduce existing hierarchies or create new forms of inequality if they fail to address underlying power dynamics in scientific communities. The challenge is not merely technical but social: creating recognition systems that simultaneously support innovation, ensure appropriate credit, maintain research integrity, and foster equitable participation in an increasingly computational scientific enterprise.

6. References

Aksnes, D.W., 2006. Citation rates and perceptions of scientific contribution. J. Assoc. Inf. Sci. Technol. 57, 169–185.

Alperin, J.P., Portenoy, J., Demes, K., Larivière, V., Haustein, S., 2024. An analysis of the suitability of openalex for bibliometric analyses. arXiv preprint arXiv:2404.17663 .

Biagioli, M., Galison, P., 2014. Scientific authorship: Credit and intellectual property in science. Routledge.

Brand, A., Allen, L., Altman, M., Hlava, M., Scott, J., 2015. Beyond authorship: Attribution, contribution, collaboration, and credit. Learned Publishing 28.

Cao, H., Dodge, J., Lo, K., McFarland, D.A., Wang, L.L., 2023. The rise of open science: Tracking the evolution and perceived value of data and methods link-sharing practices. ArXiv abs/2310.03193.

Carver, J.C., Weber, N., Ram, K., Gesing, S., Katz, D.S., 2022. A survey of the state of the practice for research software in the united states. PeerJ Computer Science 8.

Chen, L., lan Ding, J., Song, D., Qu, Z., 2024. Exploring scientific contributions through citation context and division of labor. ArXiv abs/2410.13133.

Chinchilla-Rodríguez, Z., Costas, R., Larivíere, V., Robinson-García, N., Sugimoto, C.R., 2022. The relationship between corresponding authorship and author position, in: Proceedings of the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022), pp. 7–9.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Edwards, P.N., Jackson, S.J., Chalmers, M.K., Bowker, G.C., Borgman, C.L., Ribes, D., Burton, M., Calvert, S., 2013. Knowledge infrastructures: Intellectual frameworks and research challenges .

Escamilla, E., Klein, M., Cooper, T., Rampin, V., Weigle, M.C., Nelson, M.L., 2022. The rise of github in scholarly publications, in: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A. (Eds.), Linking Theory and Practice of Digital Libraries, Springer International Publishing, Cham. pp. 187–200.

Färber, M., 2020. Analyzing the github repositories of research papers, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Association for Computing Machinery, New York, NY, USA. p. 491–492. URL: https://doi.org/10.1145/3383583.3398578, doi:10.1145/3383583.3398578.

Franceschet, M., Costantini, A., 2010. The effect of scholar collaboration on impact and quality of academic papers. J. Informetrics 4, 540–553.

Gøtzsche, P.C., Hróbjartsson, A., Johansen, H.K., Haahr, M.T., Altman, D.G., Chan, A.W., 2007. Ghost authorship in industry-initiated randomised trials. PLoS medicine 4, e19.

Haeussler, C., Sauermann, H., 2013. Credit where credit is due? the impact of project contributions and social factors on authorship and inventorship. Research Policy 42, 688–703. URL: https://www.sciencedirect.com/science/article/pii/S0048733312002259, doi:10.1016/j.respol.2012.09.009.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11, 156–162.

Hasselbring, W., Druskat, S., Bernoth, J., Betker, P., Felderer, M., Ferenz, S., Lamprecht, A.L., Linxweiler, J., Rumpe, B., 2024. Toward research software categories. URL: https://arxiv.org/abs/2404.14364, arXiv:2404.14364.

Hata, H., Guo, J.L.C., Kula, R.G., Treude, C., 2021. Science-software linkage: The challenges of traceability between scientific knowledge and software artifacts. ArXiv abs/2104.05891.

He, P., Gao, J., Chen, W., 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.

He, P., Liu, X., Gao, J., Chen, W., 2021b. Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=XPZIaotutsD.

Howison, J., Herbsleb, J.D., 2011. Scientific software production: incentives and collaboration, in: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, Association for Computing Machinery, New York, NY, USA. p. 513–522. URL: https://doi.org/10.1145/1958824.1958904, doi:10.1145/1958824.1958904.

Howison, J., Herbsleb, J.D., 2013. Incentives and integration in scientific software production. Proceedings of the 2013 conference on Computer supported cooperative work .

Jin, X., Wah, B.W., Cheng, X., Wang, Y., 2015. Significance and challenges of big data research. Big data research 2, 59–64.

Júnior, E.A.C., Silva, F.N., da Fontoura Costa, L., Amancio, D.R., 2016. Patterns of authors contribution in scientific manuscripts. J. Informetrics 11, 498–510.

Katz, D.S., Hong, N.P.C., Clark, T., Muench, A., Stall, S., Bouquin, D.R., Cannon, M., Edmunds, S.C., Faez, T., Feeney, P., Fenner, M., Friedman, M., Grenier, G., Harrison, M., Heber, J., Leary, A., MacCallum, C.J., Murray, H., Pastrana, É., Perry, K., Schuster, D.C., Stockhause, M., Yeston, J.S., 2020. Recognizing the value of software: a software citation guide. F1000Research 9.

Kelley, A., Garijo, D., 2021. A framework for creating knowledge graphs of scientific software metadata. Quantitative Science Studies 2, 1423–1446.

Krafczyk, M., Shi, A., Bhaskar, A., Marinov, D., Stodden, V., 2019. Scientific tests and continuous integration strategies to enhance reproducibility in the scientific software context, in: Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems, Association for Computing Machinery, New York, NY, USA. p. 23–28. URL: https://doi.org/10.1145/3322790.3330595, doi:10.1145/3322790.3330595.

Lamprecht, A.L., García, L.J., Kuzak, M., Martinez, C., Arcila, R., Pico, E.M.D., Angel, V.D.D., van de Sandt, S., Ison, J.C., Martínez, P.A., McQuilton, P., Valencia, A., Harrow, J.L., Psomopoulos, F., Gelpi, J.L., Hong, N.P.C., Goble, C.A., Capella-Gutiérrez, S., 2020. Towards fair principles for research software. Data Sci. 3, 37–59.

Larivière, V., Gingras, Y., Sugimoto, C.R., Tsou, A., 2014. Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology 66.

Larivière, V., Pontille, D., Sugimoto, C.R., 2020. Investigating the division of scientific labor using the contributor roles taxonomy (credit). Quantitative Science Studies , 1–18.

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., Sugimoto, C.R., 2016. Contributorship and division of labor in knowledge production. Social Studies of Science 46, 417–435. URL: https://doi.org/10.1177/0306312716650046, doi:10.1177/0306312716650046, arXiv:https://doi.org/10.1177/0306312716650046. pMID: 28948891.

Li, K., Zhang, C., Larivière, V., 2023. Are research contributions assigned differently under the two contributorship classification systems in plos one? ArXiv abs/2310.11687.

Liu, X., Zhang, C., Li, J., 2023. Conceptual and technical work: Who will disrupt science? Journal of Informetrics 17, 101432. URL: https://www.sciencedirect.com/science/article/pii/S1751157723000573, doi:10.1016/j.joi.2023.101432.

Lu, C., Zhang, Y., Ahn, Y.Y., Ding, Y., Zhang, C., Ma, D., 2019. Co-contributorship network and division of labor in individual scientific collaborations. Journal of the Association for Information Science and Technology 71, 1162 – 1178.

Mayernik, M.S., Hart, D.L., Maull, K.E., Weber, N.M., 2017. Assessing and tracing the outcomes and impact of research infrastructures. Journal of the Association for Information Science and Technology 68, 1341–1359.

Merow, C., Boyle, B.L., Enquist, B.J., Feng, X., Kass, J.M., Maitner, B.S., McGill, B., Owens, H.L., Park, D.S., Paz, A., Pinilla-Buitrago, G.E., Urban, M.C., Varela, S., Wilson, A.M., 2023. Better incentives are needed to reward academic software development. Nature Ecology & Evolution 7, 626–627.

Milewicz, R., Pinto, G., Rodeghero, P., 2019. Characterizing the roles of contributors in open-source scientific software projects, in: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), pp. 421–432. doi:10.1109/MSR.2019.00069.

Muna, D., Alexander, M., Allen, A., Ashley, R., Asmus, D., Azzollini, R., Bannister, M., Beaton, R., Benson, A., Berriman, G.B., Bilicki, M., Boyce, P., Bridge, J., Cami, J., Cangi, E., Chen, X., Christiny, N., Clark, C., Collins, M., Comparat, J., Cook, N., Croton, D., Davids, I.D., Éric Depagne, Donor, J., dos Santos, L.A., Douglas, S., Du, A., Durbin, M., Erb, D., Faes, D., Fernández-Trincado, J.G., Foley, A., Fotopoulou, S., Frimann, S., Frinchaboy, P., Garcia-Dias, R., Gawryszczak, A., George, E., Gonzalez, S., Gordon, K., Gorgone, N., Gosmeyer, C., Grasha, K., Greenfield, P., Grellmann, R., Guillochon, J., Gurwell, M., Haas, M., Hagen, A., Haggard, D., Haines, T., Hall, P., Hellwing, W., Herenz, E.C., Hinton, S., Hlozek, R., Hoffman, J., Holman, D., Holwerda, B.W., Horton, A., Hummels, C., Jacobs, D., Jensen, J.J., Jones, D., Karick, A., Kelley, L., Kenworthy, M., Kitchener, B., Klaes, D., Kohn, S., Konorski, P., Krawczyk, C., Kuehn, K., Kuutma, T., Lam, M.T., Lane, R., Liske, J., Lopez-Camara, D., Mack, K., Mangham, S., Mao, Q., Marsh, D.J.E., Mateu, C., Maurin, L., McCormac, J., Momcheva, I., Monteiro, H., Mueller, M., Munoz, R., Naidu, R., Nelson, N., Nitschelm, C., North, C., Nunez-Iglesias, J., Ogaz, S., Owen, R., Parejko, J., Patrício, V., Pepper, J., Perrin, M., Pickering, T., Piscionere, J., Pogge, R., Poleski, R., Pourtsidou, A., Price-Whelan, A.M., Rawls, M.L., Read, S., Rees, G., Rein, H., Rice, T., Riemer-Sørensen, S., Rusomarov, N., Sanchez, S.F., Santander-García, M., Sarid, G., Schoenell, W., Scholz, A., Schuhmann, R.L., Schuster, W., Scicluna, P., Seidel, M., Shao, L., Sharma, P., Shulevski, A., Shupe, D., Sifón, C., Simmons, B., Sinha, M., Skillen, I., Soergel, B., Spriggs, T., Srinivasan, S., Stevens, A., Streicher, O., Suchyta, E., Tan, J., Telford, O.G., Thomas, R., Tonini, C., Tremblay, G., Tuttle, S., Urrutia, T., Vaughan, S., Verdugo, M., Wagner, A., Walawender, J., Wetzel, A., Willett, K., Williams, P.K.G., Yang, G., Zhu, G., Zonca, A., 2016. The astropy problem. URL: https://arxiv.org/abs/1610.03159, arXiv:1610.03159.

Ni, C., Smith, E., Yuan, H., Larivière, V., Sugimoto, C.R., 2021. The gendered nature of authorship. Science Advances 7, eabe4639. URL: https://www.science.org/doi/abs/10.1126/sciadv.abe4639, doi:10.1126/sciadv.abe4639, arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abe4639.

Philippe, O., Hammitzsch, M., Janosch, S., van der Walt, A., van Werkhoven, B., Hettrick, S., Katz, D.S., Leinweber, K., Gesing, S., Druskat, S., Henwood, S., May, N.R., Lohani, N.P., Sinha, M., 2019. softwaresaved/international-survey: Public release for 2018 results. URL: https://doi.org/10.5281/zenodo.2585783, doi:10.5281/zenodo.2585783.

Ram, K., 2013. Git can facilitate greater reproducibility and increased transparency in science. Source code for biology and medicine 8, 1–8.

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. URL: https://arxiv.org/abs/1908.10084.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108.

Sauermann, H., Haeussler, C., 2017. Authorship and contribution disclosures. Science Advances 3, e1700404. URL: https://www.science.org/doi/abs/10.1126/sciadv.1700404, doi:10.1126/sciadv.1700404, arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.1700404.

Shapin, S., 1989. The invisible technician. American Scientist 77, 554–563. URL: http://www.jstor.org/stable/27856006.

Sharma, N.K., Ayyala, R., Deshpande, D., Patel, Y., Munteanu, V., Ciorba, D., Bostan, V., Fiscutean, A., Vahed, M., Sarkar, A., et al., 2024. Analytical code sharing practices in biomedical research. PeerJ Computer Science 10, e2066.

Smith, A.M., Katz, D.S., Niemeyer, K.E., 2016. Software citation principles. PeerJ Comput. Sci. 2, e86.

Stankovski, A., Garijo, D., 2024. Repofrompaper: An approach to extract software code implementations from scientific publications, in: NSLP.

Stodden, V., Guo, P., Ma, Z., 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. PloS one 8, e67111.

Trisovic, A., Lau, M.K., Pasquier, T., Crosas, M., 2021. A large-scale study on research code quality and execution. Scientific Data 9.

Trujillo, M.Z., Hébert-Dufresne, L., Bagrow, J., 2022. The penumbra of open source: projects outside of centralized platforms are longer maintained, more academic and more collaborative. EPJ Data Science 11, 31.

Westner, B.U., McCloy, D.R., Larson, E., Gramfort, A., Katz, D.S., Smith, A.M., invited co signees, 2024. Cycling on the freeway: The perilous state of open source neuroscience software. ArXiv abs/2403.19394.

Wyss, E., De Carli, L., Davidson, D., 2023. (nothing but) many eyes make all bugs shallow, in: Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses, Association for Computing Machinery, New York, NY, USA. p. 53–63. URL: https://doi.org/10.1145/3605770.3625216, doi:10.1145/3605770.3625216.

7. Appendix

7.1. Extended Data and Methods

7.1.1. Building a Dataset of Linked Scientific Articles and Code Repositories

The increasing emphasis on research transparency has led many journals and platforms to require or recommend code and data sharing (Stodden et al., 2013; Sharma et al., 2024), creating traceable links between publications and code. These explicit links enable systematic study of both article-repository and author-developer account relationships (Hata et al., 2021; Kelley and Garijo, 2021; Stankovski and Garijo, 2024; Milewicz et al., 2019).

Our dataset collection process leveraged four sources of linked scientific articles and code repositories, each with specific mechanisms for establishing these connections:

1. **Public Library of Science (PLOS)**: We extracted repository links from PLOS articles' mandatory data and code availability statements.
2. **Journal of Open Source Software (JOSS)**: JOSS requires explicit code repository submission and review as a core part of its publication process.
3. **SoftwareX**: Similar to JOSS, SoftwareX mandates code repositories as a publication requirement.
4. **Papers with Code**: This platform directly connects machine learning preprints with their implementations.

To create a comprehensive and analyzable dataset, we enriched these article-repository pairs with metadata from multiple sources. We utilized the Semantic Scholar API for DOI resolution to ensure we found the latest version of each article. This resolution step was particularly important when working with preprints, as journals may have published these papers since their inclusion in the Papers with Code dataset. Using Semantic Scholar, we successfully resolved 56.3% of all DOIs within our dataset[4].

We then utilized the OpenAlex API to gather detailed publication metadata, including:

- Publication characteristics (open access status, domain, publication date)
- Author details (name, author position, corresponding author status)
- Article- and individual-level metrics (citation counts, FWCI, h-index)

Similarly, the GitHub API provided comprehensive information for source code repositories:

[4]Broken out by dataset source, we resolved 2.1% of all PLOS DOIs, 4.0% of all JOSS DOIs, 0.0% of all SoftwareX DOIs, and 49.2% of all Papers with Code (arXiv) DOIs.

- Repository metadata (name, description, programming languages, creation date)
- Contributor details (username, display name, email)
- Repository-level metrics (star count, fork count, issue count)

7.1.2. Developing a Predictive Model for Author-Developer Account Matching

7.1.2.1. Annotated Dataset Creation.

Creating an accurate author-developer account matching model required quality-labeled training data that reflects real-world identity matching challenges. Exact-matching on names or emails proved insufficient due to variations in formatting (e.g., “J. Doe” vs. “Jane Doe”), use of institutional versus personal email addresses, and incomplete information. However, author and developer account information often contains sufficient similarities for probabilistic matching, such as when author “Jane Doe” corresponds to username “jdoe” or “janedoe123”.

To efficiently build our training and evaluation dataset, we used JOSS articles as we believed they typically feature higher author-developer-account overlap, increasing positive match density. Our dataset creation process followed these steps:

1. We generated semantic embeddings for each developer account and author name using the multi-qa-MiniLM-L6-cos-v1 model from the Sentence Transformers Python library (Reimers and Gurevych, 2019).
2. We calculated cosine similarity between all potential author-developer-account pairs for each article-repository pair.
3. For annotation efficiency, we selected the three most similar authors for each developer account.

From these generated author-developer-account pairs, we randomly selected 3,000 for classification by two independent annotators as either matches or non-matches, resolving disagreements through discussion and verification. The resulting dataset contains 451 (15.0%) positive matches and 2548 (85.0%) negative matches, comprising 2027 unique authors and 2733 unique developer accounts.

Our collected data for annotation confirmed that exact matching would be insufficient—only 2191 (80.2%) of developer accounts had associated display names and just 839 (30.7%) had associated email addresses.

7.1.2.2. Training and Evaluation.

Our training and evaluation methodology began with careful dataset preparation to prevent data leakage between training and test sets. To ensure complete separation of authors and developers, we randomly selected 10% of unique authors and 10% of unique developers, designating any pairs containing these selected entities for the test set. This entity-based splitting strategy resulted in 2442 (81.4%) pairs for training and 557 (18.6%) pairs for testing.

For our predictive model, we evaluated three transformer-based architectures that have demonstrated strong performance in entity matching tasks:

- DeBERTa-v3-base (He et al., 2021a,b)
- mBERT (bert-base-multilingual-cased) (Devlin et al., 2018)
- DistilBERT (Sanh et al., 2019)

We systematically evaluated these base models across different combinations of developer-account features, ranging from using only the username to incorporating complete profile information (username, display name, and email address). We fine-tuned all models using the Adam optimizer with a linear learning rate of 1e-05 for training and a batch size of 8 for training and evaluation. Given the size of our dataset and the binary nature of our classification task, models were trained for a single epoch to prevent overfitting.

We evaluated model performance using standard binary classification metrics: precision, recall, and F1-score. This evaluation framework allowed us to directly compare model architectures and feature combinations while accounting for the balance between precision and recall in identifying correct matches.

Our comprehensive model evaluation revealed that fine-tuning DeBERTa-v3-base (He et al., 2021a) with developer username and display name as input features produces optimal performance for author-developer matching. This model configuration achieved a binary F1 score of 0.944, with an accuracy of 0.984, precision of 0.938, and recall of 0.95. Table 3 presents a complete comparison of model architectures and feature combinations.

Table 3: Comparison of Models for Author-Developer-Account Matching

Optional Feats.	Model	Accuracy	Precision	Recall	F1
name	deberta	0.984	0.938	0.950	0.944
name, email	bert-multilingual	0.984	0.938	0.950	0.944
name, email	deberta	0.982	0.907	0.975	0.940
name	bert-multilingual	0.982	0.938	0.938	0.938
name	distilbert	0.978	0.936	0.912	0.924
name, email	distilbert	0.978	0.936	0.912	0.924
email	deberta	0.957	0.859	0.838	0.848
email	bert-multilingual	0.950	0.894	0.738	0.808
n/a	deberta	0.946	0.847	0.762	0.803
n/a	bert-multilingual	0.941	0.862	0.700	0.772
n/a	distilbert	0.856	0.000	0.000	0.000
email	distilbert	0.856	0.000	0.000	0.000

Analysis of each model's performance revealed that including developer display names had the largest positive impact on model performance compared to username alone. We also observed that mBERT's performance was comparable to DeBERTa's while using the developer email address as an additional input feature. However, we selected the DeBERTa configuration as it had consistent strong performance across various feature combinations.

To facilitate the reuse of our work, we have made our trained model and supporting code publicly available. Complete fine-tuning, evaluation, and inference code is available as the Python package: sci-soft-models, and the fine-tuned model has been released on HuggingFace (evamxb/dev-author-em-clf).

7.1.2.3. Model Limitations.

While our model demonstrates strong performance, we acknowledge certain limitations in our approach:

1. **Short name sensitivity**: Shorter names (both usernames and display names) can affect the model's performance, as less textual information is available for matching.
2. **Organization accounts**: Research lab accounts used for project management present a potential challenge for accurate matching, as they don't correspond to individual authors. However, our filtering mechanisms applied before analysis help minimize their impact in modeling.

7.1.3. Dataset Characteristics and Repository Types

Our compiled dataset appears to contain a mix of repository types, varying from analysis script repositories to software tools and likely some "code dumps" (where code is copied to a new repository immediately before publication). This diversity is reflected in the commit duration patterns across different publication types. The median commit duration for repositories in our analysis is:

- 47 days for preprints
- 104 days for research articles
- 247 days for software articles

Complete statistics on commit durations, including count, mean, and quantile details, are available in Table 4.

Table 4: Commit duration (in days) distributions for different publication types. Only includes article-repository pairs with a most recent commit no later than 90 days after publication and excludes publications from research teams in the top 3% of total author sizes.

	count	mean	std	min	10%	25%	50%	75%	90%	max
article_type										
preprint	3080.0	104.860390	178.024291	-1520.0	0.0	5.0	47.0	130.00	270.0	2091.0
research article	19502.0	184.375551	247.760274	-931.0	0.0	13.0	104.0	258.75	474.0	3176.0
software article	222.0	372.509009	470.233626	-1.0	0.0	22.0	247.0	510.75	904.7	3007.0

7.2. Modeling and Analysis Supporting Tables and Figures

7.2.1. Distributions of Author-Developer-Account Prediction Confidence

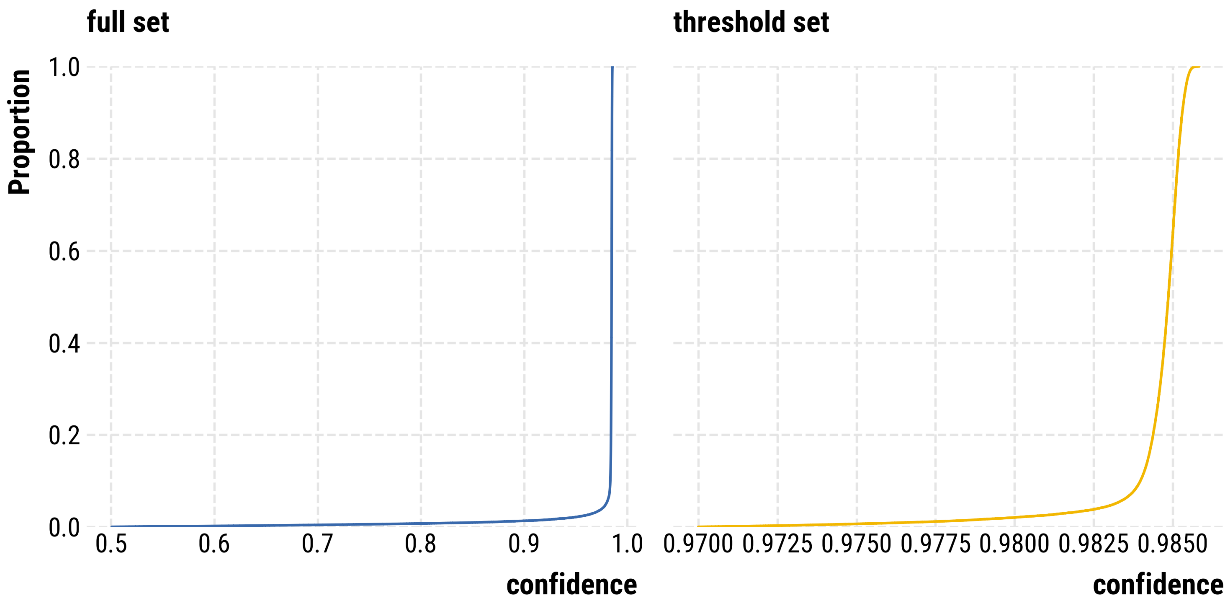

Figure 3: Distribution of author-developer-account prediction confidence scores. The left plot shows the distribution of all prediction confidence scores, while the right plot shows the distribution of prediction confidence scores for author-developer-account pairs with a confidence score greater than or equal to 0.97.

Thresholding the predictive model confidence at 0.97 resulted in a 3.2% reduction in the number of author-developer-account pairs. This threshold was chosen to ensure a high level of confidence in the matches while retaining a large number of pairs for analysis.

7.2.2. Linear Models for Software Development Dynamics Within Research Teams

Table 5: Article citations by code contributorship of research team. Generalized linear model fit with negative binomial distribution and log link function. Significant p-values are indicated with asterisks: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

Variable	**coef**	**P>|z|**	**[0.025**	**0.975]**
const ***	**0.98**	**0.00**	**0.93**	**1.03**
Total Authors ***	**0.07**	**0.00**	**0.06**	**0.08**
Code-Contrib. Authors ***	**0.05**	**0.00**	**0.03**	**0.07**
Code-Contrib. Non-Authors	-0.00	0.65	-0.01	0.01
Years Since Publication ***	**0.39**	**0.00**	**0.38**	**0.40**

Table 6: Article citations by code contributorship of research team controlled by open access status. Generalized linear model fit with negative binomial distribution and log link function. Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***)

Variable	**coef**	**P>|z|**	**[0.025**	**0.975]**
const ***	**0.61**	**0.00**	**0.49**	**0.73**
Total Authors ***	**0.07**	**0.00**	**0.06**	**0.08**
Code-Contrib. Authors	0.05	0.26	-0.04	0.14
Code-Contrib. Non-Authors	0.00	0.94	-0.03	0.03
Years Since Publication ***	**0.38**	**0.00**	**0.37**	**0.39**
Is Open Access ***	**0.41**	**0.00**	**0.29**	**0.53**
Code-Contrib. Authors × Is Open Access	-0.00	0.98	-0.09	0.09
Code-Contrib. Non-Authors × Is Open Access	-0.00	0.87	-0.03	0.03

Table 7: Article citations by code contributorship of research team controlled by domain. Generalized linear model fit with negative binomial distribution and log link function. Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***)

Variable	**coef**	**P>|z|**	**[0.025**	**0.975]**
const ***	**0.89**	**0.00**	**0.76**	**1.02**
Total Authors ***	**0.07**	**0.00**	**0.06**	**0.08**
Code-Contrib. Authors	0.02	0.66	-0.08	0.12
Code-Contrib. Non-Authors	0.01	0.62	-0.04	0.06
Years Since Publication ***	**0.40**	**0.00**	**0.39**	**0.41**
Domain Life Sciences **	**-0.20**	**0.01**	**-0.35**	**-0.06**
Domain Physical Sciences *	**0.13**	**0.04**	**0.00**	**0.25**
Domain Social Sciences *	**-0.19**	**0.02**	**-0.35**	**-0.03**
Code-Contrib. Authors × Domain Life Sciences	0.07	0.30	-0.06	0.19
Code-Contrib. Authors × Domain Physical Sciences	0.02	0.74	-0.09	0.12
Code-Contrib. Authors × Domain Social Sciences	0.11	0.09	-0.02	0.24
Code-Contrib. Non-Authors × Domain Life Sciences	-0.04	0.26	-0.11	0.03
Code-Contrib. Non-Authors × Domain Physical Sciences	-0.02	0.55	-0.07	0.04
Code-Contrib. Non-Authors × Domain Social Sciences	-0.03	0.34	-0.10	0.04

Table 8: Article citations by code contributorship of research team controlled by article type. Generalized linear model fit with negative binomial distribution and log link function. Significant p-values are indicated with asterisks: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

Variable	**coef**	**P>|z|**	**[0.025**	**0.975]**
const ***	**0.51**	**0.00**	**0.43**	**0.59**
Total Authors ***	**0.07**	**0.00**	**0.06**	**0.08**
Code-Contrib. Authors	-0.02	0.49	-0.07	0.03
Code-Contrib. Non-Authors **	**-0.03**	**0.00**	**-0.05**	**-0.01**
Years Since Publication ***	**0.40**	**0.00**	**0.39**	**0.41**
Article Type Research Article ***	**0.49**	**0.00**	**0.41**	**0.56**
Article Type Software Article ***	**-0.46**	**0.00**	**-0.72**	**-0.21**
Code-Contrib. Authors × Article Type Research Article **	**0.09**	**0.00**	**0.04**	**0.15**
Code-Contrib. Authors × Article Type Software Article	-0.07	0.30	-0.20	0.06
Code-Contrib. Non-Authors × Article Type Research Article ***	**0.04**	**0.00**	**0.02**	**0.06**
Code-Contrib. Non-Authors × Article Type Software Article	0.08	0.28	-0.07	0.22

7.2.3. Post-Hoc Tests for Coding vs Non-Coding Authors by Position

Table 9: Counts of Code-Contributing Authors ('Coding') as well as Total Authors by Position and Bonferroni Corrected p-values from Post-Hoc Binomial Tests. Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***).

Control	**Subset**	**Position**	**Coding**	**Total**	**p**
Domain	**Health Sciences**	First	1535	2404	0.000***
		Middle	510	11513	0.000***
		Last	228	2352	0.000***
	Life Sciences	First	2538	3890	0.000***
		Middle	815	11854	0.000***
		Last	469	3778	0.000***
	Physical Sciences	First	28426	41983	0.000***
		Middle	10022	111735	0.000***
		Last	3303	40407	0.000***
	Social Sciences	First	2766	4044	0.000***
		Middle	989	9251	0.000***
		Last	397	3861	0.000***
Article Type	**Preprint**	First	13146	19523	0.000***
		Middle	4872	52925	0.000***
		Last	1432	18598	0.000***
	Research Article	First	21595	32081	0.000***
		Middle	7034	89991	0.000***
		Last	2801	31188	0.000***
	Software Article	First	524	717	0.000***
		Middle	430	1437	0.000***
		Last	164	612	0.000***
Open Access Status	**Closed Access**	First	2587	3742	0.000***
		Middle	861	10965	0.000***
		Last	266	3642	0.000***
	Open Access	First	32678	48579	0.000***
		Middle	11475	133388	0.000***
		Last	4131	46756	0.000***
Overall	**Overall**	First	35265	52321	0.000***
		Middle	12336	144353	0.000***
		Last	4397	50398	0.000***

7.2.4. Post-Hoc Tests for Coding vs Non-Coding Authors by Corresponding Status

Table 10: Counts of Code-Contributing Authors ('Coding') as well as Total Authors by Corresponding Status and Bonferroni Corrected p-values from Post-Hoc Binomial Tests. Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***).

Control	Subset	Is Corresponding	Coding	Total	p
Domain	**Life Sciences**	Corresponding	1720	8014	0.000***
		Not Corresponding	2102	11508	0.000***
	Physical Sciences	Corresponding	5146	12483	0.000***
		Not Corresponding	36605	181642	0.000***
	Social Sciences	Corresponding	789	2455	0.000***
		Not Corresponding	3363	14701	0.000***
Article Type	**Preprint**	Corresponding	760	1036	0.000***
		Not Corresponding	18690	90010	0.000***
	Research Article	Corresponding	7545	27339	0.000***
		Not Corresponding	23885	125921	0.000***
	Software Article	Corresponding	206	438	0.464
		Not Corresponding	912	2328	0.000***
Open Access Status	**Closed Access**	Corresponding	250	468	0.304
		Not Corresponding	3464	17881	0.000***
	Open Access	Corresponding	8261	28345	0.000***
		Not Corresponding	40023	200378	0.000***
Overall	**Overall**	Corresponding	8511	28813	0.000***
		Not Corresponding	43487	218259	0.000***

7.2.5. Linear Models for Characterizing Code-Contributing Author H-Index

Table 11: Code-contributing authors h-index by coding status. Generalized linear model fit with Gaussian distribution and log link function. Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***).

Variable	coef	P>|z|	[0.025	0.975]
const ***	**3.19**	**0.00**	**3.18**	**3.20**
Works Count ***	**0.00**	**0.00**	**0.00**	**0.00**
Any Coding ***	**-0.32**	**0.00**	**-0.33**	**-0.30**
Majority Coding ***	**-0.77**	**0.00**	**-0.79**	**-0.74**
Always Coding ***	**-0.97**	**0.00**	**-1.02**	**-0.92**

Table 12: Code-contributing authors h-index by coding status controlled by most freq. author position. Generalized linear model fit with Gaussian distribution and log link function.Significant p-values are indicated with asterisks: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

Variable	coef	P>|z|	[0.025	0.975]
const ***	**2.38**	**0.00**	**2.31**	**2.44**
Works Count ***	**0.00**	**0.00**	**0.00**	**0.00**
Any Coding **	**0.14**	**0.00**	**0.05**	**0.22**
Majority Coding	-0.07	0.08	-0.14	0.01
Always Coding ***	**-0.24**	**0.00**	**-0.33**	**-0.16**
Common Author Position Last ***	**1.05**	**0.00**	**0.98**	**1.11**
Common Author Position Middle ***	**0.74**	**0.00**	**0.68**	**0.81**
Any Coding × Common Author Position Last ***	**-0.28**	**0.00**	**-0.37**	**-0.19**
Any Coding × Common Author Position Middle ***	**-0.45**	**0.00**	**-0.53**	**-0.36**
Majority Coding × Common Author Position Last ***	**-0.35**	**0.00**	**-0.45**	**-0.26**
Majority Coding × Common Author Position Middle ***	**-0.61**	**0.00**	**-0.69**	**-0.52**
Always Coding × Common Author Position Last ***	**-0.37**	**0.00**	**-0.51**	**-0.22**
Always Coding × Common Author Position Middle ***	**-0.51**	**0.00**	**-0.64**	**-0.38**

Table 13: Code-contributing authors h-index by coding status controlled by most freq. domain. Generalized linear model fit with Gaussian distribution and log link function.Significant p-values are indicated with asterisks: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

Variable	coef	P>|z|	[0.025	0.975]
const ***	**3.32**	**0.00**	**3.29**	**3.36**
Works Count ***	**0.00**	**0.00**	**0.00**	**0.00**
Any Coding ***	**-0.39**	**0.00**	**-0.47**	**-0.30**
Majority Coding ***	**-1.44**	**0.00**	**-1.64**	**-1.24**
Always Coding ***	**-1.22**	**0.00**	**-1.59**	**-0.86**
Common Domain Life Sciences ***	**0.10**	**0.00**	**0.05**	**0.15**
Common Domain Physical Sciences ***	**-0.14**	**0.00**	**-0.18**	**-0.11**
Common Domain Social Sciences ***	**-0.16**	**0.00**	**-0.22**	**-0.10**
Any Coding × Common Domain Life Sciences	0.10	0.09	-0.02	0.21
Any Coding × Common Domain Physical Sciences	0.07	0.12	-0.02	0.16
Any Coding × Common Domain Social Sciences	-0.00	1.00	-0.14	0.14
Majority Coding × Common Domain Life Sciences ***	**0.80**	**0.00**	**0.57**	**1.03**
Majority Coding × Common Domain Physical Sciences ***	**0.69**	**0.00**	**0.49**	**0.89**
Majority Coding × Common Domain Social Sciences ***	**0.74**	**0.00**	**0.48**	**1.00**
Always Coding × Common Domain Life Sciences	0.31	0.14	-0.10	0.73
Always Coding × Common Domain Physical Sciences	0.25	0.17	-0.11	0.62
Always Coding × Common Domain Social Sciences	0.32	0.14	-0.11	0.75

Table 14: Code-contributing authors h-index by coding status controlled by most freq. article type. Generalized linear model fit with Gaussian distribution and log link function.Significant p-values are indicated with asterisks: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***).

Variable	coef	P>|z|	[0.025	0.975]
const ***	**3.10**	**0.00**	**3.09**	**3.11**
Works Count ***	**0.00**	**0.00**	**0.00**	**0.00**
Any Coding ***	**-0.30**	**0.00**	**-0.32**	**-0.27**
Majority Coding ***	**-0.77**	**0.00**	**-0.81**	**-0.73**
Always Coding ***	**-0.99**	**0.00**	**-1.07**	**-0.92**
Common Article Type Research Article ***	**0.18**	**0.00**	**0.17**	**0.20**
Common Article Type Software Article ***	**0.22**	**0.00**	**0.12**	**0.33**
Any Coding × Common Article Type Research Article	-0.02	0.15	-0.05	0.01
Any Coding × Common Article Type Software Article	0.19	0.06	-0.01	0.39
Majority Coding × Common Article Type Research Article	0.01	0.80	-0.05	0.06
Majority Coding × Common Article Type Software Article ***	**0.36**	**0.00**	**0.19**	**0.54**
Always Coding × Common Article Type Research Article	0.00	0.93	-0.10	0.10
Always Coding × Common Article Type Software Article **	**0.37**	**0.00**	**0.15**	**0.58**

7.3. Analysis of Project Duration and Percentage Code-Contributors Who Are Authors

In our pre-registered analysis plan (https://osf.io/fc74m), we originally hypothesized about the relationship between project duration and authorship recognition. Specifically, we posited that sustained technical engagement and scientific recognition might be meaningfully related, with longer project durations potentially leading to higher rates of code-contributor authorship. We saw repository histories as providing a unique opportunity to examine this relationship, leading us to hypothesize that projects with longer commit durations would be associated with higher percentages of developers receiving authorship recognition (pre-registered as H2).

However, our analysis found no evidence to support this hypothesis. When examining the relationship between a repository's commit duration and the percentage of developers who receive authorship recognition, we found no significant correlation (r = -0.00, p = n.s.). This suggests that the length of time a project has been in development has no meaningful relationship with the proportion of developers who are recognized as authors.

We ultimately decided to move this analysis to the appendix for two key methodological reasons. First, our approach of using repository-level commit duration as a proxy for individual contribution patterns proved too coarse-grained. A more precise analysis would need to examine individual-level contribution durations and patterns rather than overall project length. Second, our method did not account for the varying levels of contribution that different developers make to a repository. Simply correlating overall project duration with authorship rates fails to capture the nuanced ways that sustained, meaningful technical contributions might influence authorship decisions.

These limitations suggest potential directions for future work that could more rigorously examine the relationship between long-term technical engagement and scientific recognition. Such work might benefit from more granular analysis of individual contribution patterns, perhaps incorporating measures of contribution significance and sustainability rather than just temporal duration.