From Open Science to Open Source: CAROL[1], a Dataset Matching Corporate Publications to Code Repositories, 2008–2025

Nur Ahmed
Rafid Irtiza
Sam M. Walton College of Business, University of Arkansas

Abstract. To what extent do firms release the code behind their scientific research? Answering this question requires linking corporate publications to the repositories that contain related software. We introduce CAROL, a dataset that links scientific publications by U.S. publicly listed firms to repositories that release, implement, or cite them. We begin with 3,731 DISCERN 2.0 firms and 181.5 million works indexed by OpenAlex for 2008–2025. From these sources, we identify over 852000 papers by 2,480 firms and more than 5,250 GitHub organizations associated with 1,692 firms. We use exact matching, Sentence-BERT and Qwen3 embeddings, and curated links from Papers with Code (PwC) to construct 90,846 paper–repository pairs spanning 53,692 papers, 39,831 repositories, and 727 firms. We manually validate matches throughout the pipeline. Our data include confidence scores, repositories' licenses, and identifiers linking each pair to Compustat, DISCERN, and OpenAlex. Among firm papers accompanied by open-source code, 46 percent are outside computer science.

Introduction

Firms that publish their science also decide whether to release the code behind it. In July 2021, Alphabet's DeepMind published AlphaFold 2 in *Nature* and released its code when the paper appeared (Jumper et al., 2021). In May 2024, it published AlphaFold 3 in the same journal without code and released the code 187 days later (Abramson et al., 2024). Firms make this choice for thousands of papers each year, extending the strategic decision to publish (Arora et al., 2018; Hicks, 1995; Polidoro & Theeke, 2012) to the release of code.

We argue that matched scientific paper–repository pairs enable multiple lines of inquiry. First, to what extent do firms disclose strategically through open code? Firms reveal knowledge selectively (Alexy et al., 2013; Henkel, 2006). Pairs show *which* findings a firm releases as code, *when*, and *under what license*. Second, does releasing code lower a firm's research costs by attracting outside contributions and enabling reuse? Public code has substantial economic value (Hoffmann et al., 2024; Nagle, 2019). Firms may, however, release code for their most promising projects. This strategic disclosure confounds the effect of release with project quality. Our dataset helps address this problem because it records the papers a firm publishes without code as well as those it releases with code. Researchers can compare papers with and without code from the same firm, field, and year. They can also exploit changes in journal and conference code-availability policies, which shift release for reasons outside the firm's control.

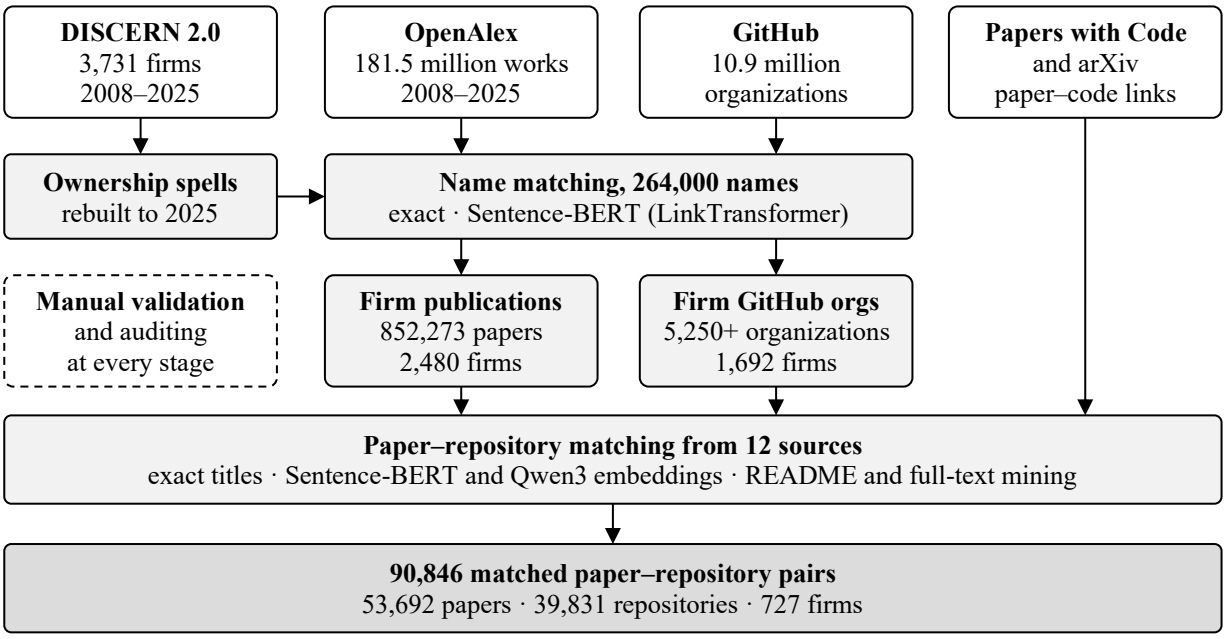
Progress on these questions has been impeded by the difficulty of matching a firm's papers to its code. Links to code appear inconsistently across papers and repositories, and GitHub does not record which firm owns an account. We introduce CAROL to fill this gap by adding code as a third channel of corporate disclosure, alongside the patents and publications that DISCERN 2.0 links to firms (Arora et al., 2024). The dataset is under development, and the current build already covers nearly all firms that lead corporate computer science research.

[1]CAROL stands for Corporate Articles, Repositories, and Ownership Linkages. The name is tentative.

Building the Dataset

We begin with DISCERN 2.0 (Figure 1), which identifies research-active firms, links them to financial data, and provides a tested method for attributing publications to firms. Its 3,731 firms active between 2008 and 2025 each hold at least one patent and have Compustat records. DISCERN assigns papers to these firms through ownership spells that follow name changes, subsidiaries, and acquisitions. These features allow us to match papers to firms at scale and to connect code disclosure to patents, publications, and financial performance. We rebuilt the ownership spells through 2025 with DISCERN's method, drawing on CRSP, Compustat, SDC, and 10-K subsidiary lists. The rebuilt table agrees with DISCERN's on 94 percent of spells. We plan to expand the dataset beyond DISCERN after completing and refining the DISCERN-based build.

Figure 1. Construction of the Dataset



Note: Counts refer to the September 2026 build. Firms are DISCERN-adjusted CRSP PERMNOs, and firm counts are not nested.

In Step 1, we identified 852,273 papers by 2,480 firms among the 181.5 million works that OpenAlex indexes for 2008–2025 (Priem et al., 2022). We combined DISCERN's publication records with firm publications that we identified in OpenAlex for 2021–2025. To identify these publications, we matched about 264,000 standardized firm and subsidiary names against author affiliations. Exact name matching identified candidate papers. LinkTransformer (Arora & Dell, 2024), a Sentence-BERT record linkage tool, and manual review then screened these candidates.

In Step 2, we mapped 1,692 firms to more than 5,250 GitHub organizations. Two research assistants identified and verified firm accounts. Exact name matching against 10.9 million GitHub organizations and hand-verified, language-model-assisted searches added more. We attribute a repository to a firm when it was created under a GitHub organization during that organization's recorded ownership spell with the firm.

In Step 3, we matched papers to repositories using multiple sources. Two curated sources, Papers with Code (PwC; 29,775 pairs) and an arXiv paper–software dataset (50,190), complement our own matching. We read repository README files for paper titles (2,484 pairs) and paper identifiers (11,441). Sentence-BERT (Reimers & Gurevych, 2019) and Qwen3 (Zhang et al., 2025)

embeddings compare each paper's title and abstract with each repository's description and README (2,651). We also read repository URLs printed in abstracts, availability statements, and full texts. One in eight pairs (11,929) is found by two or more sources.

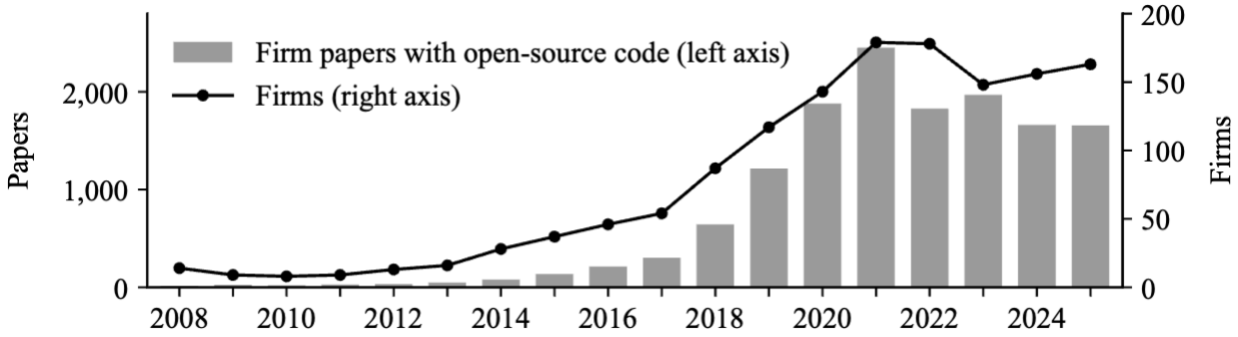
Three variables let researchers define the sample appropriate to their question. Match class records whether the repository is the paper's own open-source code release (companion), whether one side cites or uses the other (reference), or whether the sources leave the relation open (ambiguous). Ownership records whether the firm authored the paper, created the repository, or both. A confidence score from 1 to 6 grades each attribution and rises with corroboration across sources, exact title matches, and releases stated in the paper. As in Reliance on Science (Marx & Fuegi, 2020), we will release pairs at several score levels rather than at one cutoff, letting researchers trade precision against recall.

We conduct manual validation and auditing at every stage of the build. Research assistants verified GitHub organizations and affiliation matches by hand. We audit each build by tracing flagged cases, such as pairs claimed by several firms, forks, and namesakes, to their sources.

Current Status

Our data already cover most corporate computer science research and show that code release extends well beyond software, into pharmaceuticals and biotechnology. Firms in our dataset author 95 percent of the computer science papers and 96 percent of the AI papers in our firm publication universe. Nearly half (46 percent) of firm papers released with open-source code come from outside computer science, led by biochemistry, genetics, and molecular biology (13 percent) and medicine (7 percent). Companion pairs include research associated with pharmaceutical and biotechnology firms such as Amgen (320 pairs), Johnson & Johnson (280), and Pfizer (254).

Figure 2. Firm Papers Released with Open-Source Code, 2008–2025



Note: Bars report the number of firm-authored papers with a companion repository, that is, open-source code released with the paper, by publication year (left axis). The line reports the number of firms with at least one such paper (right axis).

Our data show that code release in corporate science has grown rapidly. The number of firm papers released with open-source code rose from 16 in 2008 to 2,451 in 2021 and has stayed above 1,650 per year since (Figure 2). Over the same period, the number of firms releasing such code rose from 14 to 179. About two in five companion repositories use MIT (20 percent) or Apache 2.0 (19 percent) licenses, 16 percent use custom or unrecognized terms, and 37 percent declare no license.

Next Steps

Our current dataset is a work in progress. Before public release, we will expand coverage, apply a consistent publication-matching procedure across years, refine firm attribution, and complete

independent validation. Planned additions include fuzzy matching of GitHub organization and repository names and mining of the arXiv versions of firm papers.

Validation will report precision and recall for each confidence score, following the design of Reliance on Science (Marx & Fuegi, 2020). Precision is the share of reported pairs that are correct, and recall is the share of existing pairs that we find. For precision, two research assistants will independently check 100 randomly drawn pairs at each confidence level. For recall, research assistants will search by hand for the code of about 1,000 randomly drawn firm papers to build a known-good test set. Validation will be conducted by research assistants who did not construct the matches. We will report both measures by confidence level and release only pairs at levels whose precision supports research use. Domain experts will also code how fully each repository in a subsample of companion pairs reproduces its paper (Bikard, 2018).

Data Description

The release will consist of four tab-separated files (Table 1). Each record in the main file will link a paper (OpenAlex work ID) and a repository (GitHub repository ID) to a firm (PERMNO).

Table 1. Files and Variables in the Planned Release

Variable	Provenance	Description
Matched pairs (CAROL_pairs.tsv)		
pair_id	this work	Identifier of the paper–repository pair, stable across builds
paper_id, doi	OpenAlex	OpenAlex work ID and DOI of the paper
repo_id	GitHub	Owner/name; other hosts use host:id (e.g., zenodo:7806192)
permno, gvkey	CRSP, Compustat	Firm identifiers, assigned through DISCERN ownership spells
match_class	this work	Companion, reference, or ambiguous
ownership	this work	Whether the firm authored the paper, created the repository, or both
confidence	this work	Score from 1 to 6 grading the attribution of the pair to the firm
match_source	this work, PwC, arXiv	Sources that found the pair
match_date	OpenAlex, GitHub	Earlier of publication date and repository creation date
Repositories (CAROL_repositories.tsv)		
repo_github_id, is_fork	GitHub	Numeric ID, stable when a repository is renamed, and fork status
created_at, pushed_at	GitHub	Creation date and date of last activity
stars, forks, pulls	GitHub	Stargazer, fork, and pull request counts at collection
language, license	GitHub	Primary language and SPDX license (e.g., MIT, Apache-2.0)
Firm publications and affiliations (CAROL_affiliations.tsv)		
paper_id, permno	OpenAlex, this work	Paper and the firm to which it is attributed
affiliation	OpenAlex	Raw author affiliation text that matched the firm
Firm–GitHub crosswalk (CAROL_firm_github.tsv)		
permno, github_org	this work	Firm, GitHub organization name, and period of ownership

Note: "This work" marks variables constructed for this dataset, following the data description format of Scharfmann et al. (2025). Variable names are provisional.

Uses and Dissemination

The dataset will link to existing research data and will support multiple lines of research streams. PERMNO, GVKEY, DOI, and OpenAlex identifiers connect each pair to Compustat, DISCERN patents and publications, and Reliance on Science (Marx & Fuegi, 2020). Researchers could examine how citations to a paper relate to forks and pull requests in its companion repository, documenting patterns of scientific attention and software activity around the same project. License records allow researchers to characterize licensing choices across firms and repositories. The data also show how much of industry's growing contribution to AI research (Ahmed et al., 2023) becomes reusable beyond the published paper.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500.

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635), 884–886.

Alexy, O., George, G., & Salter, A. J. (2013). Cui bono? The selective revealing of knowledge and its implications for innovative activity. *Academy of Management Review*, 38(2), 270–291.

Arora, A., & Dell, M. (2024). LinkTransformer: A unified package for record linkage with transformer language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 221–231).

Arora, A., Belenzon, S., Cioaca, L., Sheer, L., Shin, H. M., & Shvadron, D. (2024). *DISCERN 2.0: Duke Innovation & SCientific Enterprises Research Network* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13153196

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate R&D. *Strategic Management Journal*, 39(1), 3–32.

Bikard, M. (2018). Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science*, 29(5), 818–836.

Bikard, M. (2020). Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8), 1528–1543.

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, 35(7), 953–969.

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. *Industrial and Corporate Change*, 4(2), 401–424.

Hoffmann, M., Nagle, F., & Zhou, Y. (2024). *The value of open source software* (Working Paper No. 24-038). Harvard Business School.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.

Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9), 1572–1594.

Nagle, F. (2019). Open source software and firm productivity. *Management Science*, 65(3), 1191–1215.

Polidoro, F., Jr., & Theeke, M. (2012). Getting competition down to a science: The effects of technological competition on firms' scientific publications. *Organization Science*, 23(4), 1135–1153.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833.

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 and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992).

Scharfmann, E., Marx, M., & Fleming, L. (2025). Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science*, 390(6776), 891–893.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). Qwen3 Embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176.