

# From Funding to Findings (FIND): An Open Database of NSF Awards and Research Outputs

Kazimier Smith\*

Massachusetts Institute of Technology  
Cambridge, MA, USA  
kazim119@mit.edu

Yucheng Lu\*

New York University  
New York, NY, USA  
yuchenglu@nyu.edu

Qiaochu Fan

New York University  
New York, NY, USA  
qf2080@nyu.edu

## Abstract

Public funding plays a central role in driving scientific discovery. To better understand the link between research inputs and outputs, we introduce FIND (Funding–Impact NSF Database), an open-access dataset that systematically links NSF grant proposals to their downstream research outputs, including publication metadata and abstracts. The primary contribution of this project is the creation of a large-scale, structured dataset that enables transparency, impact evaluation, and metascience research on the returns to public funding. To illustrate the potential of FIND, we present two proof-of-concept NLP applications. First, we analyze whether the language of grant proposals can predict the subsequent citation impact of funded research. Second, we leverage large language models to extract scientific claims from both proposals and resulting publications, allowing us to measure the extent to which funded projects deliver on their stated goals. Together, these applications highlight the utility of FIND for advancing metascience, informing funding policy, and enabling novel AI-driven analyses of the scientific process.<sup>1</sup>

## 1 Introduction

Public funding is a cornerstone of the scientific enterprise, shaping the trajectory of research across disciplines and driving advances with broad social and economic impact. Understanding the relationship between funded proposals and the research outputs

they generate is important for scholars of science, policy makers, and funding agencies alike. To support such inquiry, we introduce the Funding–Impact NSF Database (FIND), a large-scale open resource that systematically links NSF grant proposals to their downstream publications. Unlike existing NSF award records, which often lack clean or consistent mappings to research outputs, FIND provides both structured metadata and textual content from proposals and publications, enabling researchers to examine funding outcomes with new precision. Beyond its value as a standalone dataset, FIND also opens the door to a variety of applications in metascience, economics, and natural language processing.

## 2 Preliminaries and Background

The National Science Foundation is a United States federal agency that funds scientific research in disciplines other than medicine [National Science Foundation, a]. The National Institutes of Health support medical research [National Institutes of Health]. The NIH has a robust dataset matching its funding to research outcomes via its RePORTER database, which directly links medical research grants to their subsequent publications and patents. The NSF lacks such a systematic matching. Although grants include a “project outcomes report” field, information is often missing or incomplete, and the data quality worsens for less recent grants. Moreover, the project outcomes report often does not include an explicit link (via DOI or otherwise) to publications resulting from the grant. Our dataset fills this gap.

---

\*Both authors contributed equally to this research.

<sup>1</sup>Our dataset will be publicly available soon.

The National Science Foundation funds research through grants [National Science Foundation, b]. Researchers seeking funding write a proposal in which they describe the research project, its goals, its methods, its potential outcomes, and the quantity and type of resources it requires. Proposals also describe how they will use the funds. Reviewers, often in groups and with expertise in the particular area of the proposal, examine proposals and choose which to accept (Hettich and Pazzani [2006]). Researchers submit their proposals to one of eight *directorates*, such as the Biological Sciences, which are broad research categories. Researchers may also list a more specific subfield for their proposal. Reviewers evaluate proposals and decide which to accept. Researchers can also submit proposals to renew funding for a specific project (if they need more time or more resources to complete it).

The NSF has evolved in structure since its inception in the 1950s. Here we describe some important changes relevant for our dataset (but certainly not all changes to the NSF in its history). First, Congress voted to enact the America COMPETES Act in 2007 and re-authorized it in 2010 [110th Congress, 2007]. This act substantially increased funding allocated to the National Science Foundation.

Second, in 2013, the Office of Science and Technology Policy required researchers funded by the NSF to upload data on their grants and resulting publications to a new database called the Public Access Repository. PAR aims to increase transparency of science funding in the wake of debates over what types of research should be funded, especially in the social sciences [National Science Foundation, 2023]. Since PAR is relatively new, it is quite incomplete. We analyze it as part of our dataset construction and show that it matches only a small number of grants to their resulting publications.

### 3 Related Work

Perhaps the closest work to ours is Rao et al. [2025]. They collect publicly available data on NSF grants as we do and use large language models to extract “scientific claims” and “investigation proposals” from materials science grants. They intend their dataset as training data to fine-tune a smaller model. We adapt their claim and proposal extraction for our purposes and apply it to all NSF grants after 2000. The extraction process is a vital part of our calculation of scientific success scores, but it could be costly. Rao et al. [2025] show, valuably, that smaller models can succeed at proposal extraction. This is an important

takeaway for future research that might try to calculate scientific success scores on a larger dataset. Moreover, we link proposals from NSF grants to scientific findings in publications resulting from the grants. The linkage is one of our key contributions that allows us to evaluate the scientific success of grants. While the NSF award dataset <sup>2</sup> does contain the field ”por” (Project Outcomes Report for the General Public), for most of the entries this field is missing. Even in cases where it is populated, it rarely provides a clean linkage to research output (i.e. to publication DOI).

Another closely related work, Jones and Habermann [2025], compares the two main sources linking NSF awards to publications, Crossref [Crossref, 2025] and PAR [National Science Foundation, 2025] and show that Crossref is substantially more complete than PAR. We build on their work by (1) combining Crossref and PAR, (2) applying the linkage to all NSF awards from 2000 onwards, and (3) making our dataset publicly available. They approach the data from a metascience perspective, while we leverage the textual data to extract additional information.

Jones and Habermann [2025] note that “...the percentage of journal articles in Crossref with funder metadata ... averages less than 25%.” There are many articles that are not properly linked to their funding sources because they were issued prior to the NSF public access policy, because papers were published after the grant expired, or simply because principal investigators forgot to register their work. One avenue for future work is to use machine learning techniques similar to Brown et al. [2023] to create an extended linkage dataset.

More work exists on National Institutes of Health grants and related outcomes due to better data availability. While the NIH and the NSF fund different types of research, the ideas and conclusions in research on NIH funding are complementary to our work on the NSF. Azoulay et al. [2019] construct a novel dataset linking NIH grants to the biomedical patents they generate. They use the data to show that NIH funding boosts patent production; this conclusion demonstrates the value of linking funding sources to their outcomes. Other work has examined racial disparities in NIH funding Ginther et al. [2018]. Further studies have looked at the effect of funding on postdoc career outcomes Jacob and Lefgren [2011a], scientific productivity Jacob and Lefgren [2011b], and research novelty Packalen and Bhattacharya [2020]. Notably several of these studies use *rejected* NIH grants as a baseline to estimate the impact of an ac-

<sup>2</sup><https://www.nsf.gov/awardsearch/download.jsp>

cepted grant. Data on rejected NSF grants does not appear to be publicly available.

[Li and Yan \[2019\]](#) undertake a task very similar in spirit to our second NLP application. They examine keyword lists in NIH grants and resulting publications and use the lists to compute a measure of alignment between the grants and publications. We expand on this work two ways: one, we improve the alignment metric by using large language models excellent ability to process unstructured text. Two, we cover many more disciplines than the NIH grant-publication matching which is largely specialized to biomedical research. There are other sources of federal research funds, like the Department of Energy, and future work could attempt to match grants from these other agencies to resulting publications and patents. A full dataset linking all federally funded research to its outcomes would be extremely valuable.

Finally, firms like DataSeer (<https://dataseer.ai>) also contribute to standardizing and quantifying open science. Their tools can measure, for example, the degree to which publications in a given field generate open source code. They note that research funders receive “a better return on investment” when resulting publications generate open access data and code. Critically, DataSeer do not link funding sources to publications, so they cannot directly measure the amount of open access data and code resulting from a particular NSF grant. Our dataset fills the gap and we hope it can be used to study the impact of federal funding on open access science.

## 4 Methods

### 4.1 Initial data collection

We download our raw data from the NSF Award Search database. The NSF provides a web interface where users can search for awards based on award number, keywords, authors, etc. They also provide JSON files containing data on all NSF awards in each year from 1960-2025. The database includes historical records before 1960, but we focus on the data from 1960 onwards to facilitate linking with scientific papers. Matching older NSF grants to their scientific output is a valuable avenue for future research on the history of science in the United States.

We process the raw JSON files to obtain a dataset of 525,927 unique award IDs. The NSF assigns each award to a broad category called a “directorate”. A small number of awards are assigned to administrative directorates such as the Office of the Director. We largely ignore these awards in our analysis since they are likely part of the NSF administration rather

than true scientific production. After dropping these awards, Figure 1 plots the fraction of award IDs in each directorate over time for the universe of awards from 2000-2025. We focus on this period because all directorates exist after 2000 and because the match between NSF grants and publications is far better after 2000 (Figure 2). While the fraction of grants in each directorate remains relatively steady over time, the share of Technology, Innovation, and Partnerships grants more than doubles from about 3% of all grants in 2000 to more than 7% in 2024.

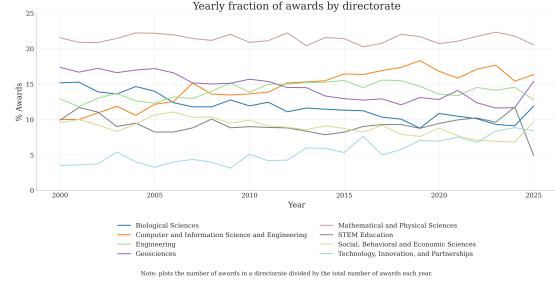


Figure 1: Grant allocation by directorate

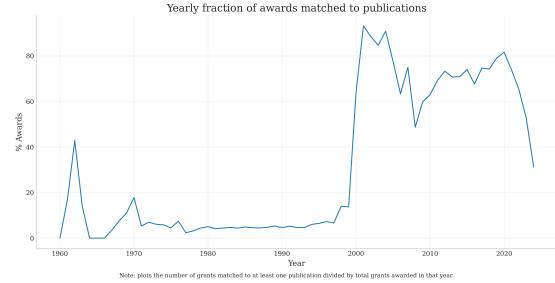


Figure 2: Grant-publication match rate

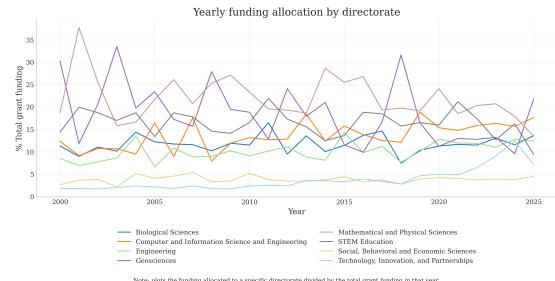


Figure 3: Fraction of grant funding in each directorate over time

Note that Figure 1 ignores the total monetary amount of the award. Figure 3 shows the fraction of total grant funding allocated to each directorate in each year. The social sciences receive substantially

Table 1: Comparison of Related Work

Study	Dataset	Method	Limitation
Rao et al. (2025)	NSF Grants	LLM	No link to outcomes
Jones and Haberman (2025)	NSF Grants	Database	No text analysis
<b>Our Work</b>	<b>NSF Grants</b>	<b>Database + LLM</b>	<b>Focus on post-2000 NSF grants</b>

less funding. Research costs in the social sciences are typically lower, so more work is needed to determine whether the social sciences are systematically underfunded.

The JSON files from the NSF award database contain 800 unique metadata fields for each award, although many of these fields are empty for most awards. For example, “pgm\_ele\_0.pgm\_ele.name” contains more granular information on the scientific field to which the grant contributes. The “por” field contains the award’s Project Outcomes Report, which is a required [110th Congress, 2007] report documenting the scientific contributions that resulted from the NSF award.

## 4.2 Matching awards to scientific publications

The next step in the dataset construction is finding the scientific publications associated with each NSF grant. We primarily use Crossref [Crossref, 2025] for this task. Crossref is a nonprofit organization that maintains a database of metadata about academic publications. Publishers enter information about their articles in the database. For this step, the most important piece of metadata in Crossref is funding sources for the publication. When an NSF grant supported a publication, the publisher can enter the NSF award ID into Crossref’s database. This allows us to link NSF grants to all the publications they supported. While participation in Crossref is voluntary, we find that 35% of NSF grants in our dataset (1960-2025) have at least one publication associated with them in Crossref. Although this is far from full coverage, it still provides a large dataset for downstream analysis and model finetuning.

We also use the NSF’s official Public Access Repository [National Science Foundation, 2025] to attempt to find publications that might not appear in Crossref. PAR began after a 2013 law required the NSF to disclose more information to the public about its grants and their outcomes. PAR is “the designated repository where NSF-funded investigators deposit peer-reviewed, published journal articles” [National Science Foundation, 2025]. Although submitting articles to PAR became mandatory in 2016, adherence

appears somewhat poor: we find that only 43.6% of NSF award IDs from 2017 onwards have a publication listed in PAR. From 2000-2016, this drops to 0.05%. Finally, 25% of publications found in PAR are also found in Crossref. We hope our publicly available dataset will help further the NSF’s goal of increasing public awareness of science funding.

Practically, we create a list of publications associated with an NSF award ID by taking the union of publications listed in Crossref and publications listed in PAR. Of our 525,927 total award IDs, 224,524 of them are matched to at least one publication (42.7%). Figure 2 shows this fraction over time. Coverage improves significantly after 2000 likely because the Digital Object Identifier system was introduced in that year doi Foundation. Improving the linkage between pre-2000 NSF grants and their publication outcomes is another important task for the study of the history of American science, and we hope our dataset can contribute to that effort.

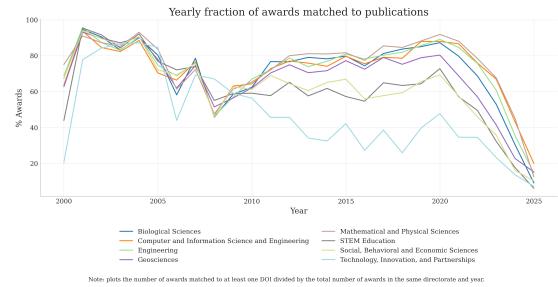


Figure 4: Grant-publication match rate by directorate

Figure 4 shows the data coverage from 2000 onwards broken down by directorate. The Social Sciences and STEM education have the lowest match rates, and it is worth noting that their match rates decrease after about 2008 while the Mathematical Sciences, Computer Sciences, and Engineering increase. We hope our dataset can help researchers explore the economic and political forces behind this divergence. Practically, using our dataset to finetune language models might lead to worse performance on social science related grants because we have fewer data points in that category.

### 4.3 Matching scientific publications to further metadata

The final step is collecting additional metadata about each publication that results from one of the NSF grants in our data. We primarily use Crossref [Crossref, 2025] and OpenAlex [Priem et al., 2022] for this. Crossref only uses publications registered with its database to calculate citation counts, so it likely provides an underestimate of the true citation count. To remedy this, we merge in citation information from OpenAlex. OpenAlex is the successor to Microsoft Academic Graph and aims to provide a completely open database of scholarly work. It includes data from Crossref and other sources (e.g. PubMed). OpenAlex contains an abstract for 512,644 publications in our dataset and a citation count for 936,703 publications. We restrict our first natural language processing task, predicting citation counts, to the subset of publications with a citation count available in OpenAlex. Our second task relies on publication abstracts, so we restrict it to that subsample of the OpenAlex data.

Figure 5 compares the semantic embedding spaces of award abstracts and paper abstracts using the SPECTER2[Singh et al., 2022] model. Each point represents a document embedding projected into two dimensions via UMAP[McInnes et al., 2020] for visualization. Award-level embeddings (red) were computed from the concatenation of award titles and abstracts, while paper-level embeddings (blue) were obtained by mean-pooling the SPECTER2 embeddings of individual paper abstracts linked to the corresponding awards. The visualization reveals substantial overlap between the two distributions, indicating that award descriptions capture much of the same topical structure as the resulting papers, despite being broader and more heterogeneous in scope. A small region to the top right shows a clean separation of paper abstracts from award abstracts, suggesting topical or linguistic divergence for a subset of research outputs.

## 5 Experiments

### 5.1 Predicting citations

**Hurdle Model** To account for the highly skewed and zero-inflated distribution of citation counts across publications linked to NSF awards, we implemented a **two-stage hurdle modeling approach**. This framework separately models (1) the likelihood that an award leads to any cited research, and (2) the expected citation impact conditional on at least

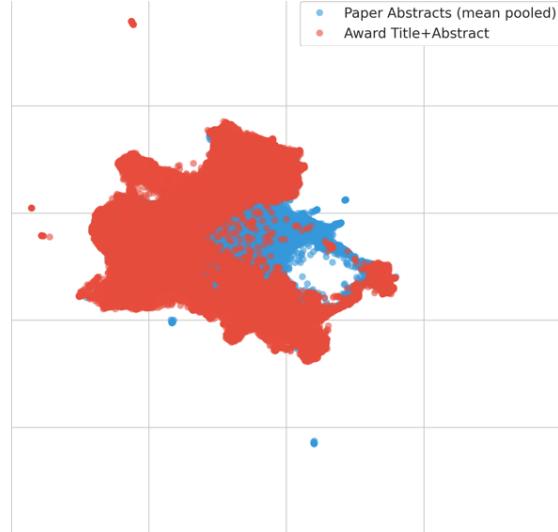


Figure 5: Award and Publication Embedding Comparison via UMAP dimension reduction

one cited publication. The hurdle model framework allows us to disentangle the drivers of “having any impact” from those that predict “how much impact” a grant might have. For example, broader or applied proposals may be more likely to produce at least one publication, while highly technical or ambitious proposals may produce fewer but highly cited outputs. This decomposition provides a more interpretable and flexible framework for assessing the predictive value of grant proposal content and metadata.

**Data and Features** We analyze NSF awards from 2000-2020, aggregating citation counts to the award level. Our approach enforces strict ex-ante prediction using only metadata available at funding time: award amount, temporal information (effective/expiration dates, duration), organizational directorate, and textual content from titles and abstracts. We generate dense embeddings from award text using SPECTER2 [Singh et al., 2022], a powerful science-domain adapted embedding model. Structured features undergo robust scaling and target encoding, with optional polynomial interaction terms to capture non-linearities. We first aggregated our dataset to the *award level*, computing the *average citation count* across all publications associated with each NSF grant. This transformation ensures each data point represents a distinct research award, capturing its total downstream citation impact rather than individual paper-level variation. We partition data into training (60%), validation (20%), and test (20%) sets with stratified sampling.

**Model Architecture** In Stage 1, we train binary classifiers ( LightGBM [Ke et al., 2017], MLPs, elastic-net logistic regression) to predict whether an award receives any citations, optimizing hyperparameters via Optuna with 5-fold cross-validation and ROC-AUC as the objective. In Stage 2, the target becomes the log-transformed citation counts for awards with citations. Beyond individual models, we construct voting, stacking ensemble, and calibrated weighted (only for classification) ensembles to leverage complementary predictive strengths.

**Results** Figure 6 presents the ROC and precision-recall curves for binary classification of citation occurrence and figure. In the first stage classification, The models achieve strong discriminative performance, with AUC scores ranging from 0.809 to 0.818 for gradient boosting and neural network approaches, and average precision scores exceeding 0.96. The precision-recall curves demonstrate that models maintain precision above 0.95 across a wide range of recall values, indicating robust identification of awards likely to generate cited publications. This high level of performance suggests that award metadata and textual content at the time of funding contain substantial signal about whether a project will produce cited research outputs. Incidentally, LightGBM (AUC = 0.818, AP = 0.964) alone seems to deliver most of the performance.

Figure 7 shows the performance of the second stage regression. While the predictive performance is not particularly strong, it is nonetheless nontrivial, especially considering that our model uses award-level information rather than features derived from the actual papers themselves. Predicting citation counts is inherently difficult, as citations are influenced by numerous unobserved social, disciplinary, and temporal factors. Interestingly, even studies that leverage features from research papers themselves typically report values in the range of 0.4-0.5 (e.g., [Petersen et al., 2011, Pobiedina and Ichise, 2014]), suggesting that our approach captures a meaningful portion of the variation despite relying on higher-level award metadata. Predicting exact citation count is

## 5.2 Grant success scores

NSF officials and scientists are likely interested in how “successful” grants are. Success has several possible definitions in this context; we focus on publication outcomes given the linking we create. NSF grants aim to advance scientific discovery, and their results are disseminated in peer-reviewed publications. Although many of the grants in our dataset

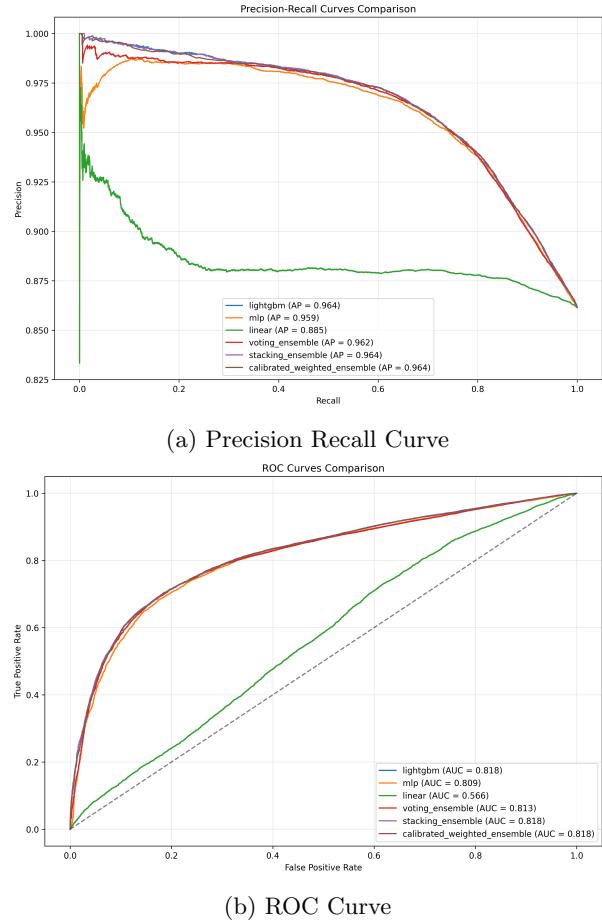


Figure 6: First Stage Performance

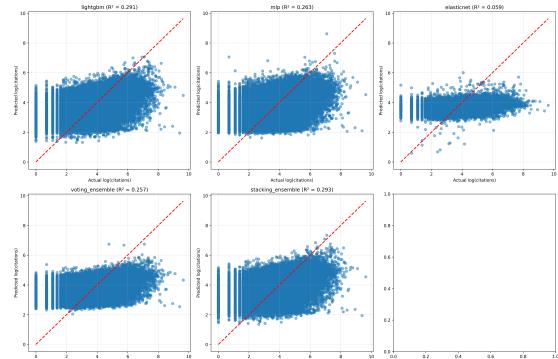


Figure 7: Second Stage Performance

result in at least one publication, it is not obvious that these publications achieve the grant’s original goals. To measure this, we calculate a “scientific success score” for each grant (we sometimes refer to it as the “alignment score”). This score measures how well each publication achieves the research proposals in the grant. A grant might produce many publications, but if few of them align with the grant’s original aims, the grant has not been entirely successful. On the other hand, a grant that produces a small number of publications which directly achieve its goals could be considered more successful.

To construct this score, we take advantage of large language models’ ability to quickly extract information from unstructured text. First, we use gpt-4.1.nano to process NSF grant abstracts. We ask the model to extract “verifiable claims” and “investigation proposals” (see Appendix A for the full prompt) from each grant abstract. A verifiable claim is an existing fact which the grant states. A grant researching plant growth might note that “plants require sunlight to produce energy” before describing its research goals. An investigation proposal is one such goal: some scientific project the grant applicant will carry out using the funding. The grant on plant research might state “we will measure the impact of partial shade on photosynthesis”. Grants can have multiple proposals. From 525,927 grants we extract 2,032,600 unique investigation proposals; the median grant has five proposals. We fail to extract proposals from 40,877 grants. The pragmatic reason to extract verifiable claims is to reduce the number of instances in which the model interprets settled science as proposed research [Rao et al. \[2025\]](#). While we do not use verifiable claims in our NLP applications here, we include them in the dataset for future research. For example, the degree to which a grant relies on existing facts might proxy for its novelty; this is a useful measure for studying changing appetites for risk at the NSF and among scientists.

Next, we process the abstracts of the 512,644 publications matched to our grants data again using gpt-4.1-nano. In particular we extract “reported methods” and “scientific findings” (see Appendix A for the full prompt). Reported methods are scientific techniques, for example, electrochemical impedance spectroscopy, that the authors used to conduct their research. Scientific findings describe the results of the publication. For example, the publication using electrochemical impedance spectroscopy might find that “A new lithium-sulfur cathode design improves cycling stability by 40% over conventional designs”. We focus on scientific findings here but include reported methods in the dataset for future work. Again, ask-

ing the model to extract reported methods helps delineate true research results from techniques and methodologies. We extract at least one scientific finding from 472,861 of 512,643 publications; among these, the median publication has three scientific findings.

Each grant has one or more investigation proposals, and each publication has one or more scientific findings. Each grant is matched to one or more publications. Within each grant, we generate all pairwise combinations of a proposal and a finding. This generates 12,424,084 (proposal, finding) pairs; each of these is associated with a unique NSF grant. Our goal is to assign a score to each of these pairs which captures how well the scientific finding achieves the goals set out in the grant proposal. Because proposals and findings are free-form, it is not obvious how to assign a score with a traditional NLP algorithm. Instead, we use gpt-4.1-nano to generate the scores: we pass each pair to the model and ask it to assign a score from zero to 100 capturing how well the finding achieves the goals in the proposal (see Appendix A for the full prompt). We are able to assign a score to 12,424,016 of the pairs, and the median pair has a score of 20. Since we generate all pairwise combinations of proposals and findings, it is unsurprising that the median score is low. Some pairs will be mismatched since a given publication will not necessarily attempt to answer all the questions in the original grant (indeed, since the median grant generates two publications, it seems plausible that grant awardees publish separate papers for distinct proposals in the grant). The mean pair has a score of 30.81 and the standard deviation of scores is 25.44. Aggregating to the grant level, the median grant has five (proposal, finding) pairs with a score above 50.

Finally, we investigate the pattern of these scientific success scores. Within each grant award year (the year in which the grant began, not the expiration year) we calculate the mean success score across all (proposal, finding) pairs. Figure 8 plots these means

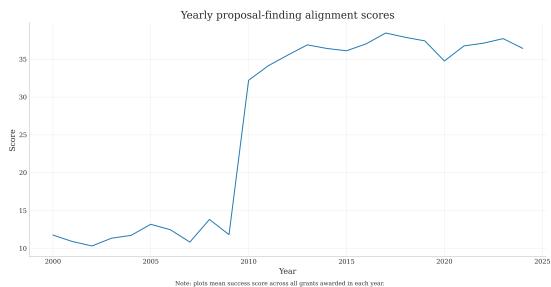


Figure 8: Scores over time

from 2000 to 2024 and shows a sharp increase in the average score starting in 2010. Figure 9 shows that this did not result purely from a decrease in the number of scored pairs, and Figure 10 shows that the pattern persists across NSF directorates. Our primary suggestion for future work is to investigate this drastic change. A change in the way the NSF awards or evaluates grants could generate the sudden jump; since there is no clear upward trend a policy change is a plausible explanation. Two notable policy changes around 2010 were:

1. The federal government increased NSF grant funding as part of the America COMPETES Act [110th Congress \[2007\]](#) which was reauthorized in 2010. If this funding improved scientists' ability to carry out the research proposed in their grants, it might have led to more publications that directly achieve grant proposals.
2. The NSF began requiring quarterly reports on grant progress. This might have simply increased pressure on scientists to deliver on their proposed research, thereby increase success scores. Alternatively, pressure to deliver positive reports might have led scientists to alter the language they use in grant proposals and in resulting publications. Since our methodology relies heavily on the actual language researchers use rather than the true underlying facts, a change in language could cause a sharp change in calculated success scores.

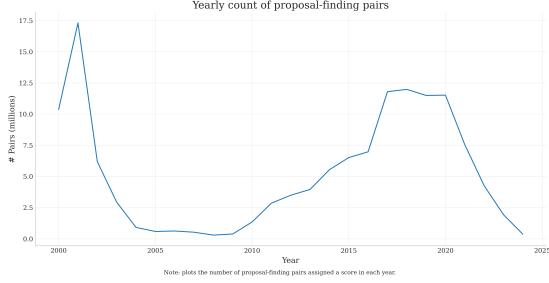


Figure 9: Number of proposal-finding pairs with a score over time

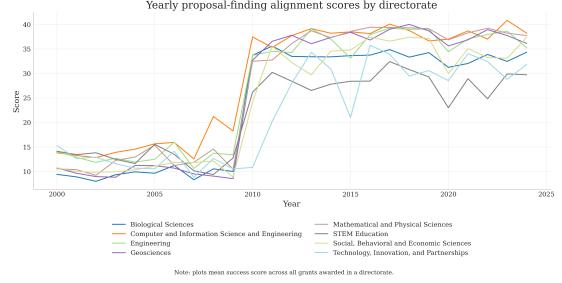


Figure 10: Scores over time by directorate

through two NLP applications: predicting citation counts and calculating grant “success scores”.

We suggest two directions for future work. First, filling in the matching between pre-2000 NSF grants and their resulting publications would create a valuable dataset for metascience research. This is a difficult task that would likely involve multiple methods and data sources. A first step might be to research the methods Crossref uses to link publications to grants and to try to extend those methods to unmatched publications. We hope our dataset can serve as a baseline to validate this matching. In particular, our (proposal, finding) pair alignment scores offer an estimate of how well publication and grant text should match. Significant divergence between pre-2000 pair scores and post-2000 pair scores could indicate a matching issue.

Second, the substantial increase in alignment scores after 2009 demands further inquiry. Both changes in NSF funding and changes in reporting requirements may have contributed, and future work could disentangle and quantify these two forces. [Packalen and Bhattacharya \[2020\]](#) also note that “...NIH funding has become more conservative despite initiatives to increase funding for innovative projects.” If innovative research is less predictable from the grants that fund it, then innovative publications will have lower alignment scores with their funding grants. An increase in risk-aversion at the NSF after 2009 could therefore explain the sudden increase in alignment scores.

## 6 Conclusion

We present the FIND dataset, an open-access collection of National Science Foundation grants across all disciplines linked to the scientific publications they ultimately generate. We demonstrate its utility

## References

- 110th Congress. America COMPETES Act. Public Law 110-69, August 2007. URL <https://www.govinfo.gov/app/details/PLAW-110publ69>. 121 Stat. 572.
- Pierre Azoulay, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. Public r&d investments and private-sector patenting: evidence from nih funding rules. *The Review of economic studies*, 86(1):117–152, 2019.
- Eva Maxfield Brown, Lindsey Schwartz, Richard Lewei Huang, and Nicholas Weber. Soft-search: Two datasets to study the identification and production of research software, 2023. URL <https://arxiv.org/abs/2302.14177>.
- Crossref. Crossref. <https://www.crossref.org/>, 2025.
- doi Foundation. What is a doi? <https://www.doi.org/the-identifier/what-is-a-doi/>. Accessed: 2025-10-06.
- Donna K Ginther, Jodi Basner, Unni Jensen, Joshua Schnell, Raynard Kington, and Walter T Schaffer. Publications as predictors of racial and ethnic differences in nih research awards. *PloS one*, 13(11):e0205929, 2018.
- Seth Hettich and Michael J Pazzani. Mining for proposal reviewers: lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 862–871, 2006.
- Brian A Jacob and Lars Lefgren. The impact of nih postdoctoral training grants on scientific productivity. *Research policy*, 40(6):864–874, 2011a.
- Brian A Jacob and Lars Lefgren. The impact of research grant funding on scientific productivity. *Journal of public economics*, 95(9-10):1168–1177, 2011b.
- Jamaica Jones and Ted Habermann. Leveraging the global research infrastructure to characterize the impact of national science foundation research, 2025. URL <https://arxiv.org/abs/2501.06843>.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Kai Li and Erjia Yan. Are nih-funded publications fulfilling the proposed research? an examination of concept-matchedness between nih research grants and their supported publications. *Journal of Informetrics*, 13(1):226–237, 2019.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL <https://arxiv.org/abs/1802.03426>.
- National Institutes of Health. Frequently asked questions (faqs). <https://report.nih.gov/faqs>. Accessed: 2025-10-06.
- National Science Foundation. About NSF. <https://www.nsf.gov/about>, a. Accessed: 2025-10-06.
- National Science Foundation. Overview of the nsf proposal and award process. <https://www.nsf.gov/funding/overview>, b. Accessed: 2025-10-06.
- National Science Foundation. Response to media reports on NSF-supported glaciers and glaciology award. Technical report, National Science Foundation, August 2023. URL [https://nsf-gov-resources.nsf.gov/2023-08/NSF\\_Response\\_supported\\_Glaciers\\_and\\_Glaciology\\_Award.pdf](https://nsf-gov-resources.nsf.gov/2023-08/NSF_Response_supported_Glaciers_and_Glaciology_Award.pdf).
- National Science Foundation. NSF public access repository (NSF-PAR). <https://par.nsf.gov/>, 2025.
- Mikko Packalen and Jay Bhattacharya. Nih funding and the pursuit of edge science. *Proceedings of the National Academy of Sciences*, 117(22):12011–12016, 2020.
- Alexander M. Petersen, H. Eugene Stanley, and Sauro Succi. Statistical regularities in the rank-citation profile of scientists. *Scientific Reports*, 1:181, 2011. doi: 10.1038/srep00181. URL <https://doi.org/10.1038/srep00181>.
- Natalia Pobiedina and Ryutaro Ichise. Predicting citation counts for academic literature using graph pattern mining. In Moonis Ali, Jeng-Shyang Pan, Shyi-Ming Chen, and Mong-Fong Horng, editors, *Modern Advances in Applied Intelligence*, pages 109–119, Cham, 2014. Springer International Publishing.
- Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works,

authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Delip Rao, Weiqiu You, Eric Wong, and Chris Callison-Burch. Nsf-scify: Mining the nsf awards database for scientific claims, 2025. URL <https://arxiv.org/abs/2503.08600>.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scireperval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL <https://api.semanticscholar.org/CorpusID:254018137>.

## Appendix A: prompts

Here, we describe the prompts we passed to large language models to accomplish various tasks in the paper. To extract verifiable claims and investigation proposals from NSF grant abstracts we use the following system prompt:

```
You are an expert at structured data extraction. You will be given unstructured text from a scientific abstract and should convert it into the given structure.
```

The full prompt to extract information from grant abstract is

```
You are a scientific text analyzer specializing in NSF grant abstracts . Extract verifiable claims and investigation proposals from the given abstract.
```

### DEFINITIONS:

- Verifiable claims: Statements presented as facts, established knowledge, or explicit/implicit assumptions that could be verified through existing evidence or prior research. Examples include:
  - \* Established scientific facts or findings
  - \* Current state of knowledge in the field
  - \* Existing problems or challenges
  - \* Prior research results referenced
  - \* Stated assumptions or premises
- Investigation proposals: Future research activities, hypotheses to test, questions to answer, or methods to develop that this award will pursue. Examples include:
  - \* Research questions to be addressed
  - \* Hypotheses to be tested
  - \* Methods to be developed or applied
  - \* Systems to be built or studied
  - \* Experiments to be conducted
  - \* Analyses to be performed

### INSTRUCTIONS:

1. Read the entire abstract carefully
2. Identify ALL verifiable claims - there may be many throughout the abstract
3. Identify ALL investigation proposals - capture the complete research agenda
4. Keep each item concise but complete (typically 1-2 sentences)
5. Preserve technical terminology and specificity exactly as written
6. Extract only what is stated or directly implied - do not add interpretation
7. Maintain the distinction between current knowledge (claims) and

future work (proposals)

### EXAMPLE:

Input: Metal-organic frameworks (MOFs) have shown promise for CO<sub>2</sub> capture due to their high surface areas and tunable pore structures. However, water stability remains a critical challenge for practical deployment.

This project will develop new water-stable MOF architectures using hydrophobic ligands and investigate their CO<sub>2</sub> adsorption performance under humid conditions.

The team will also use machine learning to predict stability and guide synthesis.

### Output:

```
{  
  "verifiable_claims": [  
    "Metal-organic frameworks (MOFs)  
      have shown promise for CO2  
      capture due to their high  
      surface areas and tunable pore  
      structures",  
    "Water stability remains a critical  
      challenge for practical  
      deployment of MOFs"  
  ],  
  "investigation_proposals": [  
    "Develop new water-stable MOF  
      architectures using hydrophobic  
      ligands",  
    "Investigate CO2 adsorption  
      performance of new MOFs under  
      humid conditions",  
    "Use machine learning to predict  
      MOF stability",  
    "Use machine learning to guide MOF  
      synthesis"  
  ]  
}
```

### OUTPUT REQUIREMENTS:

- Include ALL claims and proposals found (there is no maximum limit)
- Use empty arrays [] if no claims or proposals are found

### INPUT:

The full prompt to extract information from publications is

```
You are a scientific text analyzer specializing in extracting key results from research publication abstracts. Extract verifiable scientific findings and reported methods from the given abstract.
```

### DEFINITIONS:

- Scientific findings: Statements reporting discoveries, measurements, relationships, or conclusions based on the study. These are the

results presented in the paper. Examples include:

- \* Quantitative or qualitative outcomes
- \* Newly identified patterns, mechanisms, or effects
- \* Comparative or benchmarking results
- \* Conclusions supported by evidence

- Reported methods: Specific techniques, procedures, models, data, or tools used to produce the findings. Examples include:

- \* Experimental protocols
- \* Modeling or simulation approaches
- \* Data collection or analysis methods
- \* Instruments or datasets used
- \* Computational tools or algorithms applied

#### INSTRUCTIONS:

1. Read the entire abstract carefully
2. Identify ALL scientific findings - what the paper reports as results
3. Identify ALL reported methods - how those results were obtained
4. Keep each item concise but complete (typically 1-2 sentences)
5. Preserve technical terminology and specificity exactly as written
6. Extract only what is explicitly stated or directly implied - do not speculate
7. Clearly separate findings from methods

#### EXAMPLE:

##### Input:

We demonstrate a new lithium-sulfur cathode design that improves cycling stability by 40% over conventional designs. Electrochemical impedance spectroscopy and scanning electron microscopy were used to characterize failure mechanisms. Our model shows that capacity fade correlates with polysulfide accumulation at the separator.

##### Output:

```
{
  "scientific_findings": [
    "A new lithium-sulfur cathode design improves cycling stability by 40% over conventional designs",
    "Capacity fade correlates with polysulfide accumulation at the separator"
  ],
  "reported_methods": [
    "Electrochemical impedance spectroscopy was used to characterize failure mechanisms",
  ]
}
```

```
"Scanning electron microscopy was used to characterize failure mechanisms",
"A model was used to correlate capacity fade with polysulfide accumulation"
]
```

**OUTPUT REQUIREMENTS:**

- Include ALL findings and methods found (there is no maximum limit)
- Use empty arrays [] if no findings or methods are found

#### INPUT:

Finally, to calculate the “success score” between a specific proposal and a specific finding, we use the following system prompt:

You are an expert scientific evaluator. Your task is to score how well scientific findings achieve the research goals outlined in corresponding investigation proposals.

The main prompt is

Main prompt: Rate 0-100 how well this scientific finding achieves the research goal in this proposal: