Scooped! Estimating Rewards for Priority in Science

Ryan Hill

Northwestern University

Carolyn Stein

University of California Berkeley

The scientific community assigns credit or "priority" to individuals who publish an important discovery first. We examine the impact of losing a priority race (colloquially known as getting "scooped") on publication and career outcomes. To do so, we analyze data from structural biology where the nature of the scientific process together with the Protein Data Bank enables us to identify priority races and their outcomes. We find that scooped teams are less likely to publish in top journals and receive 21 percent fewer citations. We further study the implications of priority racing on research strategy, academic inequality, and scientist beliefs.

I. Introduction

> In short, property rights in science become whittled down to just this one: the recognition by others of the scientist's distinctive part in having brought the result into being. (Robert K. Merton 1957)

We thank the editor and four anonymous referees for valuable feedback on this paper. We are very grateful to our advisors Heidi Williams, Amy Finkelstein, Pierre Azoulay, and

Electronically published January 7, 2025

Journal of Political Economy, volume 133, number 3, March 2025.

https://doi.org/10.1086/733398

Basic science is a critical input to innovation, but it may be underprovided in competitive markets because discoveries are not directly marketable and property rights are difficult to enforce. Unlike applied research, basic (or "pure") scientific research advances our fundamental understanding of the world, but typically does not yield immediate opportunities for commercialization (Nelson 1959; Arrow 1962). As a result, credit for ideas, rather than direct profits, is an important potential motivator of innovative activity (Dasgupta and David 1994). Within academia, there is a widespread notion that the first person to publish a new discovery receives the bulk of the credit. Scientists therefore compete fiercely for priority (Merton 1957). Famous examples of priority disputes include Isaac Newton versus Gottfried Leibniz over the invention of calculus, Charles Darwin versus Alfred Wallace over the discovery of natural selection and evolution, and, more recently, Grigori Perelman versus Shing-Tung Yau, Xi-Peng Zhu, and Haui-Dong Cao over the proof of the Poincaré conjecture. This competition for recognition shapes the culture and professional structure of many disciplines, and scientists regularly worry about their work being "scooped" or preempted by a competitor (Hagstrom 1974). However, there is little empirical evidence documenting how credit is allocated in science or how rewards are shared between the "winners" and "losers" of these races. The additional credit given to the winner—what we call the *priority premium*—is an important parameter because it dictates the intensity of the competition to publish first. A relatively even credit split could lead to less competition than a winner-take-all scenario, which could meaningfully affect the pace, direction, and quality of research.

Therefore, the contribution of this paper is to empirically measure this priority premium. We analyze the impact of getting scooped on the losing project (in terms of probability of publication, journal placement, and citations), as well as on a scooped scientist's subsequent career. We also

Josh Angrist for their invaluable mentoring and support. This paper has also benefited from feedback and suggestions from David Autor, Sydnee Caldwell, Jane Choi, Brigham Frandsen, Colin Gray, Benjamin Jones, Madeline McKelway, Tamar Oostrom, Christina Patterson, Jim Poterba, Otis Reid, Jon Roth, Adrienne Sabety, Cory Smith, Ariella Kahn-Lang Spitzer, Scott Stern, Liyang Sun, Quitzé Valenzuela-Stookey, Sean Wang, and many participants in the MIT Labor and Public Finance Seminar. We thank Paula Stephan and Matt Marx for helpful discussions at the National Bureau of Economic Research Summer Institute and the European Virtual Innovation Seminar. We especially thank Scott Strobel, Stephen Burley, and Steve Cohen for detailed advice about structural biology and the Protein Data Bank. Thomas Barden, Alexia Witthaus Viñé, and Haiyi Zhang provided excellent research assistance. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant no. 1122374 (Hill and Stein) and the National Institute of Aging under Grant no. T32-AG000186 (Stein). We apologize to any authors that were inadvertently scooped by this paper; we hope that they also receive their due share of recognition. Both authors contributed equally. This paper was edited by Chad Syverson.

investigate whether competition for academic attention contributes to inequality within scientific disciplines.

Conceptually, our goal is to measure the cost of getting scooped by constructing comparisons in which multiple teams of scientists are working independently and concurrently on identical or very similar projects, which we call *races*. In practice, these races are challenging to identify for three reasons. First, many academic fields use a variety of methods and seek to answer fairly open-ended questions, and so finding near-identical projects is difficult. Second, even if the questions are well-defined, it is difficult—especially without expertise in a given scientific field—to quantify the intellectual distance between two papers in topic space. Third, scooped projects are often abandoned, making them impossible to track in publication data. We tackle these challenges by analyzing project-level data from the field of structural biology. Specifically, we examine projects in the Protein Data Bank (PDB), a repository for structural coordinates of biological macromolecules. The PDB is a centralized, curated, and searchable database of biological details contributed by the worldwide research community and contains over 150,000 macromolecule structures.[1] Several features of the PDB allow us to make headway on the key empirical challenges described above. First, structural biology papers have a well-defined objective, which is to describe the 3-dimensional shape of a known protein. Once the first paper about a protein structure is published, any follow-up publications serve mostly to confirm the result of the first. Second, projects are grouped by the PDB according to molecular similarity, which allows us to identify papers written by separate teams that solve identical or very similar molecular structures. Lastly, the PDB uniquely allows us to observe projects that are scooped shortly after completion but before publication. Scientists are required by journals to upload structures to the PDB prior to publication, so we can see projects that were completed but never appeared in print. Moreover, the rich metadata in the PDB allows us to reconstruct the timelines of projects, and find instances where teams were—unbeknownst to each other—working on the same molecule at the same time. Structural biology is a secretive field: in our data, most teams that lose priority races are scooped unexpectedly near the end of their projects.[2]

[1] The vast majority of these macromolecules are proteins, and therefore we will often refer to the entire collection as such.

[2] Historians of the field suggest that crystallography is unusually secretive due to a combination of (a) high project costs and (b) ease of imitation by competitors after those high costs have been sunk. The field has worked actively to encourage data sharing through the PDB, though the competitive nature of the field was an impediment. The compromise struck by the PDB was that scientists must only share their data at the time of publication, not before (Strasser 2019). In a survey of structural biologists we conducted, 80 percent of the respondents say they rarely if ever circulate their findings in a working paper or preprint prior to journal publication. Klebel et al. (2020) find that 40 percent of journals have unclear policies about the admissibility of preprint submissions, which may exacerbate the reluctance to share early work.

We construct races using two key dates that are recorded for all PDB projects. First, the *deposit date* marks when the scientist first uploaded her findings to the PDB. Scientists typically deposit their findings shortly after a manuscript has been submitted for publication. The second is the *release date*, which closely corresponds to the date of publication and is usually 2 to 6 months after deposit. Critically for our design, the data is hidden from the public (and from competing scientists) between deposit and release. To construct races, we find instances where two or more teams have independently deposited a structure discovery for identical macromolecules prior to their competitor's release date. The order of release then defines the outcome of the race. The first team to release is the winner, and the second team is scooped. We identify 1,611 races in our data. These races consist of 3,279 separate projects out of 67,297 total projects in our sample period from 1999 to 2017, suggesting that 5 percent of all structural biology projects are involved in a late-stage race to publication. These races are composed of a diverse set of scientific teams from different countries, institutional prestige, and experience. In the main analysis of this paper, our definition of *scooped projects* focuses only on late-stage races where both teams are on the cusp of publication. Focusing primarily on these late-stage scoops is advantageous for the economic interpretation of our results. Since both projects had been completed independently prior to publication, we can infer that the second-place team would have published the priority paper in the counterfactual where they had not been scooped. The estimated difference in observed outcomes therefore isolates the premium for novelty awarded by editors and readers. The downside of focusing on these narrow postdeposit scoops is that the scientists are passive at this point. The research has been largely completed and the timing of release is in many ways out of their hands, so these races offer little insight into the strategic interactions between racing teams, a central topic in the economics of R&D racing. Therefore, as an extension in section V, we study a sample of teams that were scooped after they had begun their experiments but before they had deposited their final project, in order to learn more about these strategic interactions.

While getting scooped is not randomly assigned, we use multiple methods to assess the validity of the causal identification assumptions. We estimate the effect of winning a race using the naturally occurring variation in the priority ordering of races. Therefore, omitted-variables bias is a threat to the causal interpretation of the estimates. If the winners are positively selected on experience, research ability, or university prestige, then our estimates of the scoop penalty will be biased upwards (in terms of magnitudes). However, we find that the outcome of races—even if not perfectly random—is highly unpredictable. We observe cases of high-ranked teams scooping low-ranked teams and low-ranked teams scooping

high-ranked teams. Throughout the analysis, we carefully document potential sources of bias and assess treatment balance using the observable team and author characteristics. To further mitigate concerns of omitted-variables bias, we use the post–double-selection Lasso method for control-variable selection (PDS Lasso; Belloni, Chernozhukov, and Hansen 2014; for Lasso, an acronym for "least absolute shrinkage and selection operator," see Tibshirani 1996).

We find that getting scooped has a moderate impact on the success of the scooped project. Scooped projects are 2.6 percent less likely to be published. Scooped papers appear in journals with impact factors that are lower by 0.19 standard deviations; they are nearly 20 percent less likely to appear in a top-10 journal. Scooped papers receive 21 percent fewer citations, and are 24 percent less likely to be a *hit paper*, defined as reaching the top 10 percent in citations for their publication year. While these effect sizes are meaningful, they are far from a winner-take-all division of credit. Focusing on citations as an outcome, our estimates imply that the losing paper receives 44 percent of the total citations accrued by both papers, a much higher share than the 0 percent assumed by a winner-take-all model. Much of the citation effect is driven by journal placement, with only a 4 percent difference in citations once we control for journal fixed effects. We provide suggestive evidence that editors and reviewers have a strong taste for novelty. Papers that are scooped prior to submission to a top journal are rarely, if ever, accepted for publication. Some scooped papers do appear in top journals, but only if they were far along in the review process on the date they are scooped.

Does getting scooped have a detrimental impact on the careers of individual authors? We compare the future publications, citations, and academic longevity of scientists on the winning and losing teams. We find that scientists who are scooped are about 6 percent less likely to be actively depositing in the PDB 5 years after this setback, and 2 percent less likely to be publishing in life and medical sciences as a whole. We do not find statistically significant effects on intensive margin publication rates. However, scooped scientists receive 20 percent fewer citations to their future work, an effect that is stronger for novice scientists (34 percent) than their veteran counterparts (16 percent).

The main analysis focuses only on scoops where the losing team had already deposited and was therefore limited in its opportunity to change its research. When considering cases of *predeposit scoops* (i.e., scoops that occur before the losing team has deposited their work), we find that scientists are able to strategically respond to being scooped by adjusting the scope and direction of their project and also by integrating insights from the winning publication. We identify this subsample of races using the "collection date" feature of the PDB, which allows us to find teams that had done their initial experiments but had not yet deposited their findings

in the PDB. Teams scooped in this intermediate stage take 1.4 years longer from collection to deposit than teams that are scooped after depositing. In that time, they tend to include additional structure deposits in their paper and shift the focus of their writing away from narrowly describing the structure itself by incorporating more analysis of protein function. They are also more likely than our main sample of scoops to build on the priority findings using a technology called molecular replacement. Although some of these strategic responses to getting scooped slow the scientists down, they also help to offset the growing scoop penalty.

We analyze and discuss how the priority-reward system relates to inequality in science. Our sample of races provides unique insight into how reputation affects academic attention, because we see teams of varying reputation and affiliation competing to publish the same discovery first. We find that when a high-reputation lab scoops a relatively unknown lab, they receive 65 percent of the total citations, but when a low-reputation lab scoops a high-reputation lab, they only receive 46 percent of the total citations. We rationalize this asymmetry in priority rewards with a model of academic attention based on the statistical discrimination literature (Phelps 1972; Aigner and Cain 1977). This relationship between priority credit and reputation suggests that compensation in science is not formulaic but may be influenced by the attention constraints and biases of editors and readers.

Finally, we benchmark the size of the scoop penalty by comparing it to the perceptions of active structural biologists. We survey 822 corresponding authors of papers linked to the PDB and pose a hypothetical scenario about getting scooped. The respondents estimate a 27 percent probability of getting scooped between submission and publication, much larger than the 3 percent chance we document in the PDB data. We then ask them to predict the probability of publication and expected citations if they are scooped by a competitor's paper. They predict that they only have a 67 percent chance of publishing the paper—again, much lower than the 85 percent of scooped projects that we observe being published in the PDB data. Finally, they estimate a 59 percent penalty in citations compared to the hypothetical winner, much higher than the 21 percent penalty we estimate in the PDB data.[3] These comparisons suggest that scientists may be overly concerned about the probability and cost of getting scooped and that perhaps better information about the true outcome of races might alleviate concerns about risk and competition in academia.

We choose to focus on structural biology because the unique features of the PDB allow us to estimate an internally valid priority effect in a way

[3] We also estimate these numbers in a subsample of the PDB data that is most similar to the hypothetical posed in the survey and still find evidence of pessimism. See table 8 for details.

that—to the best of our knowledge—would not be possible in other fields of science. However, a narrow focus on a single field naturally raises questions of external validity. Varying norms, institutions, and technology across different academic fields might lead to different distributions of priority and mechanisms for assigning credit. The scoop penalty may be higher in structural biology than in, for example, economics, because structure discoveries are "one right answer" solutions and therefore similar papers are potentially more substitutable. On the other hand, because structural biology is an experimental field, there could be inherent value in replication, which might increase the attention granted to scooped papers as compared to more theoretical fields like pure mathematics. We argue that structural biology is an important area of research per se and is therefore worthy of our attention. However, the research questions and methods structural biologists use are similar to other important fields in the basic life sciences, and so we suspect that our qualitative conclusions may apply to these fields as well.

The size of the priority premium directly relates to the level of competition in science. In a scenario where priority rewards are evenly split between the first- and second-place teams, there is no reason to compete to publish first. At the opposite extreme, if priority rewards are winner-take-all, the competition will be intense. This competition, in turn, has important implications for how science functions. On one hand, sharp priority rewards can encourage intense effort on solving frontier problems. A priority system also has the public benefit of encouraging disclosure, which is critical for fostering follow-on innovation (Williams 2013). On the other hand, some have theorized that R&D racing might induce overinvestment and duplication of effort on particular projects (Loury, 1979; Hopenhayn and Squintani 2021). In a companion paper (Hill and Stein 2024a), we study how high levels of competition generated by unequal priority rewards also impact the quality of scientific work. Our results suggest that the competition to publish first induces scientists to rush and ultimately results in lower-quality research. Some journals—seemingly in response to these rushing concerns—have begun to explicitly offer a grace period in which they will consider scooped papers for publication (Marder 2017; PLoS Biology Staff Editors 2018). This appears to be an effort to directly reduce the priority premium by ensuring more credit for the second-place team. Moreover, competition may affect science along other dimensions. For example, high levels of competition may reduce collaboration and the free sharing of information, ultimately slowing scientific progress. Therefore, measuring the priority premium—which maps directly to the intensity of scientific competition—is a critical first step in this agenda.

The remainder of the paper proceeds as follows. The following paragraphs offer a brief literature review. Section II provides some scientific

background and a description of our data. Section III describes the empirical design and identification. Section IV presents results for the short-run impact of scoops on publication, journal placement, and citations, as well as the long-run career results. We also discuss the role of editors and the timing of races for the distribution of priority rewards. Section V studies the strategic response to being scooped in races where the scooped team had not yet completed the project. Section VI describes a model of academic attention and reports results for heterogeneity of the scoop penalty by preexisting reputation. Section VII benchmarks the size of our estimates against the beliefs of surveyed structural biologists about the probability and cost of getting scooped. Section VIII concludes.

Related literature.—This paper contributes to several distinct but connected literatures, both in economics and disciplines interested in the "science of science." First, and most broadly, it contributes to our understanding of how incentives for basic research are structured. Second, it adds to a more narrow empirical literature about the causes and consequences of innovation races. Finally, it contributes to a literature about career dynamics in scientific labor markets and the role of academic reputation.

Priority races in science are often compared to patent races in industry. However, incentives for basic scientific advances are in many ways distinct from patents. Inventors in a patent race are competing for profits, while researchers in a priority race are competing for journal placement, citations, and recognition from their peers. However, both systems compensate researchers for the production of public goods, incentivize timely disclosure of knowledge, and hasten the pace of discovery. Both systems are usually conceptualized as tournaments for a discrete innovation reward or prize, with the first innovator getting the outsized share of rewards.

Theoretical models of patent races have considered how racing affects the amount of R&D investment (Loury 1979; Lee and Wilde 1980) as well as the pace of research and the amount of risk-taking induced by the structure of races (Dasgupta and Stiglitz 1980). Many of these models presuppose a winner-take-all reward that has implications for the outcome of innovation tournaments and the strategic behavior of the participants. The conventional wisdom in the sciences—and the assumption underlying much of the theoretical economics work on the topic—is that the process of scientific discovery is also a winner-take-all tournament, even if the prize is priority recognition rather than a patent (Merton 1957; Stephan 1996). Dasgupta and David (1994) explain that a discontinuous priority reward might arise in science because of a fundamental verification problem. Because of the public goods nature of new knowledge, a team that tries to publish the second paper cannot credibly prove to the community that they would have successfully completed the project

absent the help of the priority paper. Even if it would be socially optimal to share more credit with teams who were working in parallel, these information frictions might make credit-sharing difficult. The discontinuous priority-reward structure has implications for the pace of research and the strategic interaction of teams (Bobtcheff, Bolte, and Mariotti 2017). The literature on innovation systems has yielded influential models but as yet very little empirical evidence about the actual distribution of rewards in these races. Therefore we believe our estimates provide important context for theoretical and policy discussions about the incentives for scientific innovation.

This paper joins a small literature that aims to study innovation races empirically (Lerner 1997). Most related to our work, Thompson and Kuhn (2020) document that winners of patent races do more innovation in the future, and that this innovation is more likely to be related to the original patent. The authors identify patent races by looking for patents that were rejected for lack of novelty. Bikard (2020) studies the phenomenon of simultaneous discovery in science, and documents many cases of papers that are similar in content, are published around the same time, and are frequently cited together. However, our method of using biological details to link competing papers allows us to find simultaneous discoveries where one paper goes unpublished or is cited infrequently in the future.

Our heterogeneity by reputation estimates contribute to work in sociology and economics about path-dependent advantage in academic prestige, commonly called the *Matthew effect* (Merton 1968). Our results build on recent empirical work that has documented evidence of the Matthew effect in life sciences (Azoulay, Stuart, and Wang 2013), astronomy (Hill 2019), and grant funding (Jacob and Lefgren 2011; Bol, de Vaan, and van de Rijt 2018; Wang, Jones, and Wang 2019).

II. Background and Data Construction

A. *Scientific Primer: Structural Biology and the Role of Proteins*

In this section we provide a primer on the field of structural biology, a setting particularly conducive to studying scientific races. Structural biology is the study of the 3-dimensional structure of biological macromolecules. These macromolecules include DNA, RNA, and, most commonly, proteins. Proteins contribute to almost every process inside the body: hemoglobin transports oxygen in blood, actin and myosin trigger muscle contractions, and insulin regulates blood sugar. In many ways, the form or structure of a protein determine its function. For example, antibodies are Y-shaped immune system proteins that bind to foreign molecules (like viruses

or bacteria) with two of their arms, while recruiting other immune system proteins with the remaining arm. It is exactly this Y shape that allows the antibody to function (NIGMS 2017). Protein folding and structure has important applications, particularly in medicine, and 15 Nobel Prizes have been awarded for advances in structural biology (Wlodawer et al. 2008; Martz et al. 2019).

Proteins are composed of chains of amino acids, which range in length from a few dozen to several thousand amino acids long. These chains fold, giving the protein its three-dimensional shape. Scientists have long known how to determine a protein's amino acid sequence, but it is much more difficult to understand how they are folded. Most protein structures are solved using a technique called x-ray crystallography, and each structure-determination project may take many months or years. Scientists grow proteins into crystals, subject them to x-ray beams at large synchrotron facilities, and use the resulting diffraction data to determine a model of the protein's structure (Goodsell 2019). Although knowledge about protein structures is useful for applied technologies, the discovery of the structure itself is not patentable.[4] New structures are usually solved by academic researchers at universities or research centers, although 15 percent of the scientists in our sample work at nonprofit research laboratories or private companies.

B. The PDB

We focus on structural biology because the PDB contains detailed, organized, and comprehensive project-level data that is publicly available. The PDB is a worldwide repository of biological macromolecule structures, 95 percent of which are proteins.[5] The PDB was established in 1971 at Brookhaven National Laboratories with just seven structures. Today, the PDB contains over 150,000 macromolecule structures, and is growing at a rate of about 10 percent annually (Berman et al. 2000; Burley et al. 2019).

The PDB spent many decades trying to actively encourage contribution and overcome norms of secrecy that had been pervasive in the field of crystallography. Researchers are encouraged (and in many cases required) by the PDB to disclose experimental details, methodology, atomic coordinates describing the structural model, and raw experimental data if

[4] A 2013 Supreme Court ruling (Assoc. for Molecular Pathology v. Myriad Genetics Inc., 569 U.S. 576 [2013]) precludes patents on naturally occurring products such as proteins, genes, and bacteria in the United States. However, even prior to this ruling, patents on the 3-dimensional structure of proteins were rare and difficult to obtain (Seide and Russo 2002; Shimbo et al. 2004).

[5] The remaining types of molecules in the PDB are DNA, RNA, or a complex of protein, DNA, and/or RNA.

possible. Crystallographers are particularly tight-lipped about their research progress because each project represents a huge investment of time and resources. Once results are produced, they are easy to imitate and highly useful to competing scientists working on similar or related projects (Strasser 2019). There was an obvious public benefit for systematic contribution of discoveries in the PDB, particularly for comparative modeling and survey research, but there were very low private incentives for participation (Hill, Stein, and Williams 2020). In early days, the small community of crystallographers was able to maintain an honor system that discouraged encroaching on projects known to be in progress, but this norm broke down as the field grew in size and competitiveness (Ramakrishnan 2018). For many years, the PDB used a variety of schemes to try to encourage community participation and data sharing, including direct solicitation, public cajoling, and even prize drawings (Strasser 2019). However, since the early 1990s, the majority of scientific journals have required that any published structures be deposited in the PDB (Barinaga 1989; Berman et al. 2000, 2016). Furthermore, in 1998 top journals including *Science*, *Nature*, and *PNAS* formalized a policy to ensure simultaneous release of academic papers and PDB details (Campbell 1998; Sussman 1998), as encouraged by the PDB and the International Union of Crystallography.

Because of these strict public disclosure policies, we believe the PDB represents a near-complete census of macromolecule structure discoveries. Whenever a structural biologist completes a project, she uploads the structure, experiment, and discovery details to the PDB. This typically happens shortly before or after she submits an academic paper describing her findings for publication. An important feature of this process is that the uploaded data is confidential. No other user of the PDB can access the data or see that the deposit has been created. Even the editor and reviewers only receive a receipt of deposit from the PDB and author, and they do not see the underlying structure data until the date of publication. Only at the point of publication is the data released to the public. If any project goes unpublished, the data is released by default after 1 year (wwPDB 2019).

The primary unit of analysis in the PDB is a structure deposit, which is a unique report about the determination of a single protein by one research lab. Each structure deposit is assigned a unique identification number. For example, PDB ID 4HHB, deposited in 1984 by Giuilio Fermi and coauthors, reports the structure of human deoxyhemoglobin, the form of hemoglobin without oxygen that is the predominant protein in red blood cells (Fermi et al. 1984).

The PDB provides three key pieces of information that we will use in our analysis. The first is a measure of similarity between proteins. This is calculated by comparing how similar a protein's amino acid chain is to other proteins in the PDB. For a given protein, the PDB uses an algorithm

to construct a list of other proteins that are 100 percent similar, 90 percent similar, and so on, all the way down to 30 percent similar. These groupings, or "clusters," allow us to determine whether two structure deposits from different teams correspond to the same or a very similar protein. The second key piece of information that the PDB provides is a list of dates for the structure deposit, including when the data was deposited and when it was released. This allows us to construct a timeline for the projects and identify cases in which two or more teams were working simultaneously on the same protein. Finally, each PDB structure is linked to the academic paper that the structure was published in (if any). This link includes the PubMed ID, which we link to PubMed bibliographic data and Web of Science citation data.

C. Identifying Priority Races: Challenges and Solutions

Identifying priority races in scientific data is difficult for three reasons. First, to facilitate identification, research questions should be well defined and share a common approach to solving the problem. To underscore the importance of this requirement, consider economics, a field where this is not the case. There are many papers on the same topic or question (e.g., what is the effect of raising the minimum wage on employment?) that are often published in close succession (e.g., Cengiz et al. 2019 and Jardim et al. 2022). And yet, because there are a variety of methods, settings, and approaches, these papers may be quite distinct. Therefore, the first paper to be published does not necessarily scoop subsequent papers that aim to answer the same question. For our purposes, we need a field where the questions are tightly defined with a common approach, a feature that seems more common in the hard sciences than the social sciences. The second challenge is identifying papers that answer the same question. Manually comparing papers to decide whether they address the same question is infeasible at scale. Ideally, we would have some objective measure of scientific proximity, which can tell us whether two teams are working on an identical problem. Finally, the third challenge is that scooped papers are often abandoned without publication. If authors abandon their projects when they see that a similar paper has been published, many scooped papers will never show up in bibliographic data.

The PDB enables us to make significant progress on these three obstacles. First, the questions in structural biology are well-defined because scientists are typically trying to solve the structure of a known protein. Moreover, the methods are consistent: 91 percent of proteins are solved using x-ray crystallography. This means that if we observe two papers that study the structure of the same protein, these two papers are likely to be very similar in terms of their questions, methods, and conclusions. Second,

as mentioned in section II.B, the PDB measures how biologically similar different proteins are to one another. This allows us to link research projects based on objective measures of their scientific proximity, rather than relying on text similarity or citation patterns. Finally, scientists are required to deposit their structures in the PDB prior to publication. This gives us the ability to observe some projects that never reach publication. Given that scientists might abandon projects that get scooped, this record of unpublished projects is a key feature of our data. We will discuss the timeline in more detail in the next section. To the best of our knowledge, we are the first to measure scientific races in a data-driven manner.[6]

D. Defining Priority Races

Broadly speaking, we define a *priority race* as an instance where two or more teams are working on the same protein independently and concurrently and are likely uncertain about the identity or progress of their competitors. Following Brown and Ramaswamy (2007), we define the *same protein* as meaning two proteins within the same 50 percent or higher sequence-similarity group (called a *cluster* in the PDB). This is a conservative cutoff, as 30 percent has been suggested as sufficient similarity for building homology models (Moult 2005; Dessailly et al. 2009). In other words, the first publication within these 50 percent similarity clusters is often highly cited because it provides a novel structure model that other crystallographers can build on to solve very similar proteins.[7] The PDB assigns identification numbers to clusters of similar proteins, and we say that the first structure released in that cluster is the *priority* structure deposit. There are often many subsequent deposits that report similar structure coordinates as the priority deposit, only some of which we define as being scooped. These follow-on deposits appear for a variety of reasons. Some are concurrent projects by authors that were racing to be first but were scooped or are replication projects of the same protein by future teams, while others are new projects that solve the structure for closely

[6] Thompson and Kuhn (2020) are able to identify patent applications that were engaged in a patent race by finding patents that were rejected for lack of novelty. Bikard (2020) identifies paper "twins" using papers that are frequently cited together, but this approach precludes cases where one team captured the outsized share of citations by construction, or cases where a project is abandoned.

[7] Figures A1 and A2 (figs. A1–A4 are available online) provide evidence that, at each level of similarity above 50 percent, paper pairs (i.e., one scooped and one winning paper) in our sample have very similar titles and have similar rates of citation. For robustness, we can restrict to scoops by proteins within the same 100 percent cluster and find similar results, which we report in table A5 (tables A1–A11 are available online). If a protein is scooped by more than one other protein, we give preference to the protein that is biologically closer (i.e., in the "higher" cluster). See app. B (apps. A–E are available online) for details on the data construction.

related proteins that either derive from different organisms or else are bonded with different macromolecules in a novel way.[8]

We use project timelines reported in the PDB to determine whether a follow-on deposit qualifies as scooped by the priority deposit. The PDB provides two key dates at the structure level that help us determine whether two teams were working concurrently: the deposit date and release date.[9] The deposit date corresponds to the date that the scientist uploaded her solved structure to the PDB. Importantly, the structure is not yet visible to the public. Nearly all scientific journals require that authors upload their structures to the PDB prior to publication, so deposit typically occurs slightly before or after the date that the scientist first submits her paper. The release date is the date that the PDB deposit is made public. This typically corresponds to the publication date. In cases where the structure is never published, the PDB releases the deposit by default 1 year after the deposit date. Figure 1 provides a visual timeline of these dates, as well as some summary statistics. Throughout this analysis we will always use the release date as the relevant marker of priority. An alternative approach would be to use paper publication dates to determine priority ordering. But these dates are often unavailable, especially for older publications, or are ambiguous in recent data because online publication may come before print edition publication. Further, we treat publication as an outcome variable; we would risk potential bias if we conditioned on publication as a requirement for treatment assignment. Lastly, PDB releases are publicly salient events that the community pays attention to, so the release dates are therefore good markers of priority order. Appendix A4 discusses implications and presents evidence about the concordance between release dates and publication dates in greater detail.

Figure 2 illustrates how we define a scoop event. Consider two projects, A and B, authored by two distinct teams working on the same protein. Suppose project A is a priority project in one of the similarity clusters. We say that project A scoops project B if (i) A is released before B is released but (ii) after B has deposited to the PDB. Condition i guarantees that A finishes first, while condition ii guarantees that B did not know about A until after the structure was deposited in the PDB. Since B had already deposited a completed structure, they likely would have been the priority deposit had they not been scooped by A. Requiring that B has deposited before A is released ensures that we observe abandoned projects, since all deposited structures appear in our data even if they

[8] For example, there are 30,153 clusters of proteins in the PDB that are 50 percent similar, and each cluster has an average of six deposits, only some of which are eligible to be considered racing according to our definition.

[9] The scientists also report a collection date, which is the date the scientist took her crystals to the synchrotron and collected her experimental data. Typically deposit occurs about one to two years after collection.

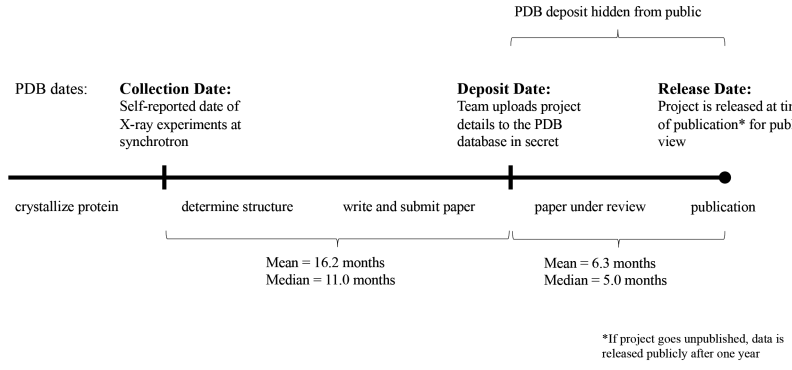

FIG. 1.—Project timeline and key dates. This figure shows the timeline of a typical PDB project in our regression sample. Dates in bold above the line are observed in our data. Events listed below the timeline are the approximate timing of other project events, including the submission and review process. Deposit and structure data is hidden from public until the structure is released.

are scooped and fail to publish. We allow the priority project to scoop more than one team, and 5.6 percent of the races we identify have three or more competitors. Appendix B provides a more detailed description of the data work necessary to construct these races in practice. In our

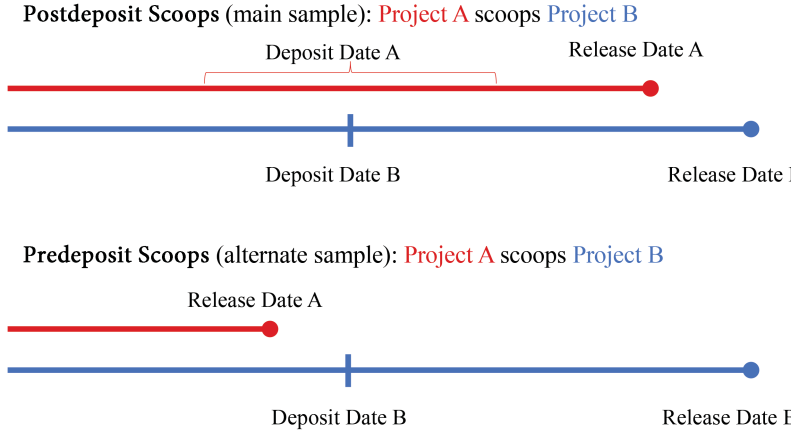

FIG. 2.—Defining priority races. This figure shows visually the timing rule we use to define scoops. In the first example, project A scoops project B because both projects were deposited prior to project A's release. These postdeposit scoops make up our main analysis sample of races. In the second scenario, project A releases before project B, but project B had not yet deposited at the time of project A's release. Therefore this example would be excluded from our main regression sample but is used in our analysis of predeposit scoops in sec. V.

main analysis, we exclusively focus on these clean, but narrowly defined scoops that occur after B has already deposited. However, in section V we expand our analysis to include earlier-stage scoops, that occur before B deposits.

1. An Example

To help understand our procedure, consider the example outlined in table 1. The table shows two structures: 4JWS and 3W9C. Both are structures of the Cytochrome P450cam protein complexed with its redox partner, putidaredoxin (Pdx-P450cam complex). This enzyme is involved in metabolism and clearing toxins, such as in the human liver. Figure 3 shows the nearly identical biological assembly models that each team deposited independently and confidentially to the PDB. The scientists at Leiden University (3W9C) collected their data a few months before the scientists at the University of California, Irvine (4JWS) (February 3, 2012 versus September 14, 2012). However, by the time of deposit, the UC Irvine team had pulled ahead, depositing one week before the Leiden team (March 27, 2013 versus April 3, 2013). Ultimately, UC Irvine won the priority race, with their structure being released two months before Leiden (June 19, 2013 versus August 21, 2013). Importantly, when Leiden deposited their structure on April 3, 2013, UC Irvine had not yet released their structure. This means that Leiden was likely unaware of their competitor's

TABLE 1
EXAMPLE PRIORITY RACE: PDX-P450CAM COMPLEX

	Winning project	Scooped project
PDB structure ID	4JWS	3W9C
Protein name	Pdx-P450cam complex	Pdx-P450cam complex
Paper title	"Structural Basis for Effector Control and Redox Partner Recognition in Cytochrome P450"	"The Structure of the Cytochrome P450cam-Putidaredoxin Complex Determined by Paramagnetic NMR Spectroscopy and Crystallography."
Key dates:		
Collection	September 14, 2012	February 3, 2012
Deposit	March 27, 2013	April 3, 2013
Release	June 19, 2013	August 21, 2013
First author affiliation	University of California, Irvine	Leiden University
Journal	*Science*	*Journal of Molecular Biology*
JIF	31.5	4
5-year citations:	52	39

NOTE.—This table presents an example of a racing pair identified in the PDB using the scoop rules outlined in section II.D See fig. 3 for the image of the structure models deposited by each team.

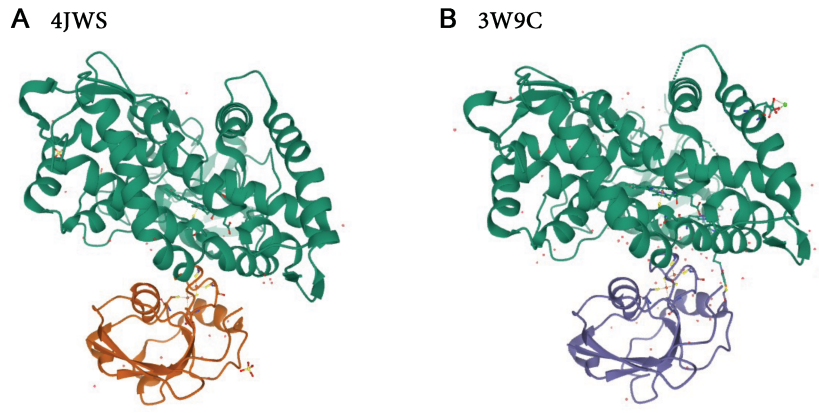

FIG. 3.—Example priority race: Pdx-P450cam Complex. This figure presents a side-by-side comparison of the biological assembly models of the Pdx-P450cam complex protein deposited by two independent racing teams. According to the scoop definition in sec. II.D, structure deposit 4JWS scooped structure deposit 3W9C. See table 1 for more details.

progress or results when they were preparing their publication and depositing the structure. Comparing the outcomes of the winner (4JWS) and the loser (3W9C), we observe that the winning paper was more successful. It was published in a better journal (*Science*, with an impact factor of 31.5, vs. *Journal of Molecular Biology*, with an impact factor of 4.0) and received about 30 percent more citations over the next 5 years (Hiruma et al. 2013; Tripathi, Li, and Poulos 2013). In this case, the Leiden authors became aware that they were scooped during the manuscript review. In the conclusion of their paper, they write, "While this manuscript was under review, Tripathi et al. published the crystal structure of the Pdx–P450cam complex that was obtained via cross-linking of the two proteins. It is interesting to compare our complex with those reported in that study. Tripathi et al. found a position and orientation of Pdx relative to P450cam that is essentially identical with ours" (Hiruma et al. 2013).[10]

2. Additional Sample Restrictions

We make three further restrictions to minimize cases of ambiguity in the race-construction procedure. First, we drop some proteins that are

[10] Overall, 33 percent of the scooped papers in our sample directly cite the winning paper. The probability that this citation occurs increases with a larger gap in time between publication. For scooped projects that are released less than 1 month after the winner, fewer than 14 percent cite the winning paper. That probability increases to 64 percent for races with more than an 8-month gap between release dates. See fig. A3.

exceedingly complex. Some very large proteins are composed of many entities that are sometimes solved piece by piece over many years instead of all at once. This introduces the possibility that a scientist could be scooped on only a fraction of their project.[11] Second, we drop projects that are published in a paper that is linked to 15 or more other structures. Among the set of papers included in our final analysis sample, 46 percent are linked to more than one structure, and the average number of structures per paper is 1.9. Multistructure papers are at risk of being scooped on a fraction of the full project. This restriction allows for some fractional scoops to enter our data, but ignores papers where each protein becomes a very small fraction of the full contribution of the paper. Finally, we drop races that end in a near or exact tie. Occasionally, two racing papers will be submitted to the same journal and the editor will publish them as companion pieces in the same issue, and we drop these cases. We also drop races where the two papers were released closer than 2 weeks apart from each other. We make this restriction to help ensure that the first project has a clear claim of priority and that the order of release is more likely to correspond to the order of publication.[12]

E. Additional Data Sources

This section describes the additional data sources that we use to define outcome variables, control variables, and provide further details about our setting. Additional details on data sources can be found in appendix A.

Journal Citation Reports.—Journal Citation Reports is an annual report published by Clarivate Analytics that evaluates journal influence using a metric called *journal impact factor* (JIF). Let $\text{Cites}^{j}_{t,t-k}$ be the number of citations that journal j received in year t for articles written in year $t - k$. Let $\text{Articles}^{j}_{t-k}$ be the number of articles published by journal j in year $t - k$. Then journal j's impact factor in year t is given by

$$\text{JIF}^{j}_{t} = \frac{\text{Cites}^{j}_{t,t-1} + \text{Cites}^{j}_{t,t-2}}{\text{Articles}^{j}_{t-1} + \text{Articles}^{j}_{t-2}}. \tag{1}$$

In words, JIF attempts to capture a journal's rolling average citations per article. We standardize the impact factors within a year t to account

[11] Proteins are often composed of subunits called entities. The clustering algorithm in the PDB groups similar molecules at the entity level, not the structure level. Therefore we define clear rules for dealing with proteins that are scooped on more than one of their constituent entities. We also drop projects with 15 or more entities because of exceeding complexity. Appendix B describes in more detail how we deal with multientity structures in the data.

[12] The PDB only releases structures once per week, which can also make very close scoops ambiguous in terms of which truly came first. Our 2-week restriction helps eliminate these cases but has a minimal impact on our results. See app. A4 for more details on the correspondence between the PDB release date and publication date.

for the fact that impact factors have been rising over time as the rate of publishing within the life sciences has increased. We also use JIF to create a list of top-10 journals. In order to focus on journals that are both high-impact and also relevant to structural biology, we restrict to a potential list of the 30 journals with the most PDB linkages in each half decade. That set is then restricted to the 10 highest-impact journals in each 5-year span. The list contains top-ranked general interest journals as well as top-ranked life science journals.[13]

PubMed, Author-ity, and Web of Science.— The Web of Science is a database of over 73 million scientific publications written since 1900 which are linked to their respective citations. The data are owned and maintained by Clarivate Analytics. We link the PDB to the Web of Science using PubMed identifiers, which are unique IDs assigned to research papers in the medical and life sciences by the National Library of Medicine. We use these data to compute citation counts for PDB-linked papers. Our primary outcome is citations in the 5 years following publication, excluding self-citations. We also construct a measure of whether a structure was published in a hit paper by ranking PDB articles by 5-year citation counts and marking the top 10 percent with the highest citation counts within years. The version of the Web of Science that we use ends in 2018, therefore we restrict the regression samples for these outcomes to 1999–2013 to allow for time for publications to accrue citations we can observe.

We construct career histories of variables before and after the priority date of each race to serve as control variables and long-run outcomes. Reconstructing publication records for individual authors is difficult because names are not disambiguated in the PubMed or PDB. We use a dataset called Author-ity, which groups PubMed IDs into distinct author identifiers using coauthor and topic patterns (Torvik et al. 2005; Torvik and Smalheiser 2009). However, because not all PDB deposits are published, it is hard to link unpublished deposits to the correct name identity in Author-ity. Therefore, in the long-run results section, we restrict to a subset of authors that have uncommon names and uniquely match to an individual in Author-ity. We also use simple name-matching techniques within the PDB to construct control variables of team productivity prior to treatment, which we can do for all deposits including those that are not published. We describe the name disambiguation procedures in detail in appendix A6.

For long-run outcomes, we count PubMed publications, PDB-linked publications, top-10 publications, citation-weighted publications, and hit publications for the years following the treatment date. Besides analyzing

[13] Top-10 journals in 2017: *Nature, Science, Cell, Journal of the American Chemical Society, Nature Chemical Biology, Nature Structural and Molecular Biology, Nature Communications, Angewandte Chemie, Nucleic Acids Research,* and *Proceedings of the National Academy of Sciences.*

the effects of race outcomes on the intensive margin of publication, we also consider the extensive margin of exit from publishing PubMed papers and PDB-linked papers altogether.

QS World University Rankings.—We use information about the affiliation ranking of the PDB scientists as control variables and to predict their academic reputation. The QS World University Rankings is an annual publication that globally ranks universities both overall and within subjects. We use the 2018 life sciences and medicine rankings, as this field is the most relevant to our setting. The ranking methodology combines four sources: a global survey of academics (academic reputation), a global survey of employers (employer reputation), citations per paper, and faculty *h*-index values. These four sources are aggregated to create a total score which is used to rank the 500 best universities.

Editorial Dates.—In section IV.C, we analyze how the scoop penalty is affected by the timing of the scoop event relative to the journal review and publication timeline. We supplement our data with the received, accepted, and publication dates for papers published in journals owned by a handful of large publishers. While we were not able to obtain these dates for all articles, we chose to focus on journals based on their prevalence in the PDB and the availability of the data for download. The journals included in the subsample are flagship or field journals from the following journal groups: Science, Nature Journals, Cell Press, and Public Library of Science (PLoS). This subsample covers 21 percent of our primary regression sample.

Scientist Survey.— In order to benchmark the magnitudes of our findings, we surveyed structural biologists about their perceptions of the probability and costs of getting scooped. Email surveys were conducted in September of 2019. We collected email addresses from the Web of Science, which provides a contact email for many of the corresponding authors on academic publications. The recruitment sample was defined as any corresponding author on a PDB-linked publication from 2014–2019 that had an email address available in the Web of Science files. We sent recruitment emails to 8,984 unique email addresses, and encouraged respondents to participate on a volunteer basis. We received 822 responses, for a total response rate of 9.1 percent. Each potential recruit received one initial solicitation and two follow-up reminders to complete the survey. Relevant text of the questionnaire is provided in appendix D.

F. Summary Statistics

By identifying priority races, we effectively split the PDB into two mutually exclusive groups: structures involved in a priority race (the *racing* sample) and structures not involved in a priority race (the *nonracing* sample).

TABLE 2
SUMMARY STATISTICS FOR STRUCTURE-LEVEL DATA

Variable	Racing (1)	Nonracing (2)	Difference (racing − nonracing) (3)	Difference (SE) (4)
A. Team Characteristics				
Authors (no.)	7.120	7.454	−.333	(.079)***
Affiliation:				
North America	.291	.351	−.060	(.008)***
Europe	.152	.158	−.006	(.006)
Asia	.191	.134	.056	(.007)***
University (rank 1-50)	.250	.240	.010	(.008)
University (rank 51–200)	.238	.261	−.023	(.008)***
Other	.512	.499	.013	(.009)
Industry or nonprofit	.152	.171	−.018	(.006)***
First author experience (years)	5.444	5.983	−.538	(.109)***
Last author experience (years)	7.418	7.806	−.387	(.120)***
B. Project Outcomes				
Published	.866	.752	.114	(.006)***
Standardized impact factor	.113	−.045	.158	(.021)***
Top-10 journal	.356	.283	.073	(.010)***
5-year citations (no.)	26.178	17.245	8.933	(.736)***
Hit paper	.148	.083	.065	(.007)***
Observations	3,279	64,018		

NOTE.—This table presents summary statistics for the racing and nonracing samples. Observations are at the structure level. Column 1 shows the means of the racing sample, and col. 2 shows the means of the nonracing sample. Column 3 shows the difference between the racing and nonracing projects, and col. 4 shows the heteroskedasticity-robust standard error of the difference. Hit papers are those in the top 10 percent for 5-year citations among articles in their publication year.

*** $p < 0.01$.

Table 2 shows summary statistics at the structure level for both of these samples. Just under 5 percent of the structures in our sample are involved in a priority race. We look at both team characteristics and deposit outcomes. Teams involved in priority races tend to be smaller, younger, and more likely to come from a top university. The racing scientists were also more likely to work in Asia, and less likely in North America. The deposit outcomes suggest that proteins involved in priority races are scientifically more important. Proteins in the racing sample are more likely to be published, appear in higher-ranked journals, and receive more citations.

III. Empirical Design

The analysis is designed to identify the causal effect of getting scooped on the short-term success of the project (publication, journal placement,

and citations), as well as on subsequent academic success of the scooped authors. We estimate the difference in outcomes between the winners and losers of the priority races in the PDB. In an ideal setting for causal inference, the winners and losers would be randomly assigned. In reality, the outcome of these late-stage races is not exactly random but is highly unpredictable. We present evidence that although some characteristics of the teams are correlated with winning a race, these observables can only explain very small differences in outcomes. In this section, we present the main estimating equations of our analysis, describe and test for potential sources of bias, and explain the control selection strategy we use to deal with potential selection bias.

A. *Baseline Specification*

Equation (2) presents the basic specification for the project-level regressions. For deposit i studying protein p, we estimate

$$Y_{ip} = \alpha + \beta \text{Scooped}_{ip} + \mathbf{X}'_{ip}\delta + \gamma_p + \epsilon_{ip}, \tag{2}$$

where Y_{ip} is an outcome, such as publication, JIF, or citations. Scooped_{ip} is an indicator for losing a priority race, $\mathbf{X}_{ip}$ is a vector of covariates,[14] and γ_p is a protein (i.e. race) fixed effect.[15] The main coefficient of interest is β, which identifies the scoop penalty. All standard errors are clustered at the protein level. Our identifying assumption is that Scooped_{ip} is uncorrelated with the error term once we condition on observable covariates and the protein involved in the priority race.

In section IV.B, we consider the long-run effect of getting scooped on academic career outcomes. The regression specification is similar to equation (2), but the unit of observation is a scientist, rather than a project. For scientist s who coauthored deposit i that was in a priority race over protein p, we estimate

$$Y_{isp} = \alpha + \beta \text{Scooped}_{isp} + \mathbf{X}'_{isp}\delta + \gamma_p + \epsilon_{isp}, \tag{3}$$

[14] Covariates include all variables listed in table 2, excluding resolution and *R*-free. Variables in panel A are included for both first and last author. We also control for variables in panels B and C calculated over the full career (in addition to the counts calculated over 5 years). Lastly, we control for indicators that tag first and last authors that have common last names as defined in app. A6.

[15] The main econometric justification to include protein fixed effects is that we have a small number of races with more than one scooped team (i.e., some races involve three teams: one winner and two losers). To the extent that these races differ from the standard two-team races in some unobserved way, there will be a mechanical correlation between losing the race and that unobserved factor, because in races with more than two teams, there are multiple losers but only a single winner. Including race fixed effects is an efficient way to non-parametrically control for this potential omitted-variables bias.

where Scooped_{isp} is a dummy equal to one if scientist s was scooped on project i. $\mathbf{X}_{isp}$ is a vector of scientist-project covariates, such as the number of publications accumulated by scientist s in the 5 years before the priority date associated with project i. We also include cubic controls for career age, which is defined as the number of years since the author's first publication in the PDB, as well as the university rank of the first author affiliation and the continent where the first author is located. Again, γ_p is a protein fixed effect. The long-run outcomes are calculated as the sum of each outcome in the 5 years following the priority date. Importantly, we exclude the publication that is linked to the structure ID of the PDB projects that were involved in the race. These outcomes therefore represent productivity in other projects not including the winning or losing paper in each race. Although each scientist may win or lose races multiple times, we include each appearance as a separate treatment event, and consider the subsequent outcomes for all scoop events.

B. Identification and Balance

Comparing outcomes of winners and losers of the PDB races identifies the causal effect of getting scooped if the race ordering is as good as randomly assigned. There are many reasons a team might win or lose a priority race, and it is plausible that the order of completion is somewhat idiosyncratic. The randomness of the scientific process, day-to-day operation of scientific labs, and the vagaries of the journal review process leave ample opportunity for random chance to dictate the timing of these races. Anecdotal accounts of ill-timed personnel issues, lab accidents, or unlucky experiment failures suggest that the timing of project completion is oftentimes out of the hands of even the most diligent and skilled scientist (Ramakrishnan 2018; Yong 2018). Furthermore, after the deposit date and submission of a manuscript, the team has very little discretion over the timing of the review process, which may be delayed by editor preference, reviewer inattention, or publisher congestion. Moreover, scientists typically have little information about the identities or progress of their competitors.

On the other hand, skill, experience, or resources could provide an advantage to certain teams that would allow them to systematically start earlier or work faster and therefore win priority races. This is a threat to identification because these characteristics may simultaneously increase the probability of winning and improve project outcomes. For example, suppose a technological breakthrough marks the starting point of a race that many diverse teams enter. If one team from Harvard has exceptional resources to adopt the technology and complete the project first, we will observe them win the race and receive many citations. But since Harvard is a high-reputation university and has a track record of success, they

would likely have received many citations even in the counterfactual where their competitor won the race. Therefore, we rely on the assumption that well-resourced or otherwise high-reputation teams are not able to systematically win priority races, and we test this using observable characteristics of each team.

If winning a priority race is random, then winning and losing teams should look balanced based on observables. We assess this observed balance between winners and losers in table 3. Using the information disclosed by the teams in the PDB, we inspect a variety of observable characteristics that might reasonably be correlated with the probability of treatment or with outcomes. These include the number of authors, the location of the lab, the rank of the university affiliation, and the experience in years of the first and last authors. We also calculate measures of the authors' productivity in PDB-related publications in the 5 years prior to the racing deposits. These include the number of PDB deposits, publications, and publications in top-ranked journals.[16]

Table 3 shows the mean values of each covariate for the winning and losing teams, as well as for the teams in the non-racing sample, for reference. We report test statistics for the difference in means between the winning and losing teams, as well as an *F*-statistic for a test of joint significance of all covariates. We find that many of the covariates are balanced between the winning and losing teams. But winning and losing teams are statistically different in a few notable dimensions. North American and European teams are more likely to win than lose, while Asian teams are more likely to lose than win. Scientists from top-50 ranked universities are more likely to win, as well as first and last authors with slightly less experience. The prior productivity of these labs is more balanced, with most measures of productivity being statistically insignificant for both first and last authors (though winning first authors appear to have deposited more). We also test whether the scientific results that are being deposited by both teams are similar. Refinement resolution and *R*-free are two variables reported by the PDB that describe the objective quality of the experimental data and model in each deposit. Resolution describes the degree of precision in the diffraction data produced during crystallography experiments, and *R*-free measures the goodness-of-fit between the experimental data and the proposed structure model. For both of these measures, smaller values imply better quality. These two measures are very close to balanced between winners and losers, suggesting that the quality of the science or the skill of the scientists is likely not driving our results. Taking the table as a whole, we reject the null hypothesis of balance on the full battery of covariates based on an *F*-statistic of 4.02.

[16] We do not use citations accrued to the racing papers because many of those citations would be assigned after the treatment date of the priority races and could therefore be endogenous to the outcome of the race.

TABLE 3
COVARIATE BALANCE BETWEEN WINNING AND LOSING TEAMS

Variable	Nonracing (1)	Losing Projects (LP) (2)	Winning Projects (WP) (3)	Difference (LP − WP) (4)	Difference (SE) (5)
	A. Team Characteristics				
Number of authors	7.454	7.183	7.056	.127	(.205)
Affiliation:					
North America	.351	.262	.320	−.057	(.022)***
Europe	.158	.134	.170	−.036	(.018)**
Asia	.134	.224	.156	.067	(.018)***
University (rank 1–50)	.240	.223	.278	−.055	(.021)***
University (rank 51–200)	.261	.248	.228	.020	(.020)
Other	.499	.529	.494	.035	(.023)
Industry or nonprofit	.171	.153	.152	.001	(.018)
First author experience (yrs.)	5.983	5.744	5.134	.611	(.279)**
Last author experience (yrs.)	7.806	7.521	7.311	.210	(.313)
	B. First Author Productivity (prior 5 years)				
Deposits	12.361	3.791	5.504	−1.714	(.687)**
Publications:					
Total	2.893	2.591	3.139	−.548	(.464)
In top-10 journals	.656	.709	.670	.038	(.065)
In top-5 journals	.222	.262	.239	.023	(.032)
	C. Last Author Productivity (prior 5 years)				
Deposits	44.269	30.937	28.991	1.946	(4.327)
Publications:					
Total	9.905	12.513	13.398	−.884	(2.241)
In top-10 journals	4.027	4.669	4.622	.047	(.511)
In top-5 journals	1.421	1.653	1.799	−.146	(.190)
	D. Project Quality Metrics (lower is better)				
Resolution (Å)	2.244	2.328	2.315	.013	(.062)
R-free goodness-of-fit	.236	.245	.243	.002	(.002)
Observations	64,018	1,668	1,611	*F*-statistic: 4.019***	

NOTE.—This table compares characteristics of winning and losing projects in order to check for treatment balance. Observations are at the structure level. Column 1 shows the means of the nonracing sample, col. 2 shows the means of the losing projects in the racing sample, and col. 3 shows the means of the winning projects in the racing sample. Column 4 shows the difference between the losing and winning projects, and col. 5 shows the heteroskedasticity-robust standard error of the difference. The *F*-statistic and associated *p*-value is calculated in a regression in which all of the variable values are stacked into a single left-hand side outcome variable and the treatment indicator is interacted with variable fixed effects on the right-hand side.

** $p < 0.05$.

*** $p < 0.01$.

Unbalanced covariates lead to biased estimates only if they are systematically correlated with the outcome variable. Therefore, to further assess potential selection bias, we visually inspect the difference in expected citations between winners and losers. We estimate a paper-level model

of citations using a Lasso regression of 3-year citation counts on the battery of team covariates. This model is estimated only in the sample of nonracing deposits. We then take the selected variables and estimated coefficients to predict citations in the racing sample in a post-Lasso ordinary least squares (OLS) procedure. The covariates we include are counts of publications, citations, and journal placements in the 5 years prior to the deposit for the first and last author, as well as the squares of these variables. We also use the career age of the first and last authors, the rank of the first author's institution in 10-school bins, and the country and university of the first author. The Lasso model selects many of the variables one would expect to be important, including dummies for being in the US and dummies for university rank. The full Lasso results are reported in table A1.

Figure 4 plots a histogram of the difference in predicted citations between each pair of winning and losing teams (races with three or more

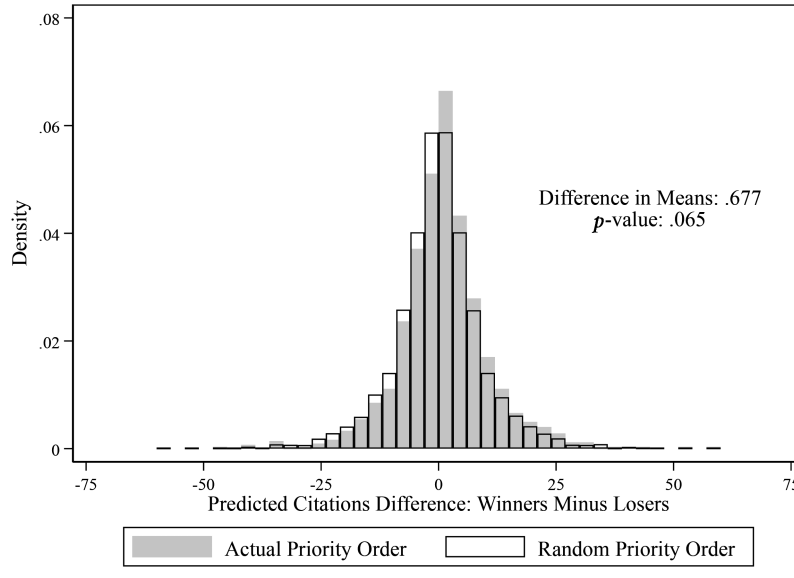

FIG. 4.—Histogram of team reputation difference. An observation in this figure is a racing pair. The gray shaded distribution shows the actual difference in predicted citations. Bars to the right of zero represent instances when the winning team had higher predicted citations than the losing team, and bars to the left of zero represent instances when the winning team had lower predicted citations than the losing team. The distribution outlined in black shows the difference in predicted citations if the winning and losing teams were randomly chosen. This random selection of winners was simulated 100 times to create the histogram and is therefore close to symmetric and centered around zero.

teams are omitted here). A perfectly balanced sample would be centered around zero and symmetric. If winners were systematically better resourced, had a better reputation, or had more experienced, then the histogram would be skewed to the right. As a benchmark for perfect balance, we compare this distribution to a simulated distribution where we randomly assign one of the paired teams as the winner. We simulate this coin flip 100 times per pair. The true distribution is shifted slightly to the right of the randomly simulated distribution, suggesting that winners are slightly more likely to be high-reputation than would be predicted by chance. But the differences in the distribution are small. The difference in means between the two distributions is 0.68 predicted citations with a p-value of 0.065 (for reference, the sample average is about 12 citations, so this represents a 6 percent difference). This slight lack of balance motivates our control strategy discussed in the next section.

C. *Control Selection Using PDS Lasso*

In light of potential treatment imbalance, we rely on an identification assumption that treatment is exogenous conditional on observable control variables. There are many potential control variables in our data, so we use a method called PDS Lasso (Belloni, Chernozhukov, and Hansen 2014) to optimally select control variables. Consider a partially linear model similar to equation (2),

$$Y_{ip} = \alpha + \beta \text{Scooped}_{ip} + \mathbf{g}(\mathbf{Z}_{ip}) + \gamma_p + \epsilon_{ip}, \tag{4}$$

where $\mathbf{Z}_{ip}$ is a large set of control variables. Assume that ϵ_{ip} satisfies an exogeneity assumption such that the treatment is mean independent of ϵ_{ip} conditional on controls. Then β will be consistently estimated if we can control for a sufficiently good approximation of $\mathbf{g}(\mathbf{Z}_{ip})$. Rather than relying on an ad hoc procedure to choose controls, PDS Lasso offers a robust approach to estimation and inference for β.

The PDS-Lasso method uses two steps. First, it estimates a Lasso regression of Scooped_{ip} on $\mathbf{Z}_{ip}$ to select a set of regressors that are predictive of treatment. Then it uses a second Lasso regression of Y_{ip} on $\mathbf{Z}_{ip}$ to select regressors that are predictive of the dependent variable. The selected control variables are highly informative of treatment assignment and outcomes and therefore reduce bias in estimation. The superset of selected regressors from those two regressions are used as the control variables in a post-OLS regression of Y_{ip} on Scooped_{ip}. The potential set of regressors we use are the variables listed in note 14, as well as squares of those variables and university rank binned into 10-school dummies. The protein fixed effects γ_p are included as unpenalized regressors in all steps of the method.

IV. Results

A. *Short-Run Effect on Projects*

Table 4 reports the regression results for the project-level effect of getting scooped. We focus on five primary outcomes: (1) an indicator for whether the project was published; (2) the JIF (standardized within year); (3) an indicator for publishing in a top-10 journal as measured by impact factor; (4) total citations accrued in 5 years, transformed with the inverse hyperbolic sine function; and (5) an indicator for becoming one of the top 10 percent of publications measured by 5-year citation counts.[17] Not all projects are published, and if they are, they may not be published in a ranked journal. We count unpublished papers as having zero citations. If the project is not published in a ranked journal, we impute the impact factor of their publications as being equivalent to the minimum journal ranking in the regression sample. The sample is restricted in columns 4 and 5 to projects released before 2014 to allow a full 5 years of data coverage to count citations in that window before our citation data ends in 2018. We present regression results from three different specifications. Panel A shows the results from a simplified version of equation (2) with no control variables. Panel B adds all controls listed in table 3, and panel C uses controls selected from the PDS-Lasso procedure described in section III.C. The results across all five outcomes suggest that covariates have very little impact on the coefficients between panel A and panel C, assuaging concerns about omitted-variables bias. We will use panel C as the preferred specification to report our estimates throughout the paper. To further test for selection bias on unobservables, we implement a robustness check following Oster (2019) in table A2.[18]

Scooped projects are 2.6 percentage points less likely to be published, off of a baseline publication rate for winning projects of 88 percent. This represents a 3 percent decrease in probability of publishing, or, framed

[17] The inverse hyperbolic sine transform is a standard way of dealing with a right-skewed distribution that contains zeroes and/or negative numbers (Burbidge, Magee, and Robb 1988; Bellemare and Wichman 2019). The transformation is given by

$$\operatorname{asinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right).$$

The coefficients on variables transformed by the hyperbolic sine function can be interpreted similarly to logs (i.e., proportionally).

[18] Adding controls and protein fixed effects increases the R^2 from less than 0.01 to over 0.60 in all regressions, suggesting that most of the variance in the outcome is explained by treatment and observable controls. Implementing the suggested bias adjustment, we conservatively assume a maximum $R^2 = 1$ and $\delta = 1$ (unobservables are equally important for treatment selection as observables), and find that the adjusted coefficients are almost identical to our baseline findings. Further, the δ needed to reduce the estimate to zero ranges from 8 to 60 across all specifications, meaning there would need to be an unrealistic degree of selection on unobservables to threaten the robustness of the results.

TABLE 4
EFFECT OF GETTING SCOOPED ON PROJECT OUTCOMES

Dependent Variable	Published (1)	JIF (2)	Top-10 Journal (3)	5-Year Citations (4)	Hit Paper (5)
	A. No Controls				
Scooped	−.025	−.192***	−.065***	−.245***	−.037***
	(.015)	(.044)	(.020)	(.071)	(.014)
	B. Base Controls				
Scooped	−.026**	−.182***	−.063***	−.216***	−.028**
	(.013)	(.045)	(.021)	(.063)	(.014)
	C. PDS-Lasso–Selected Controls				
Scooped	−.026***	−.186***	−.062***	−.208***	−.036***
	(.010)	(.032)	(.015)	(.045)	(.010)
Winner Y mean	.879	−.027	.320	28.830	.149
Observations	3,279	3,279	3,279	2,514	2,514

NOTE.—This table presents regression estimates of the scoop penalty, following equation (2). Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls listed in table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses and are clustered at the race level. The JIF in col. 2 is standardized by year. The regression in col. 4 uses asinh(5-year citations) as the dependent variable, but winner Y mean is reported in levels for ease of interpretation. Hit papers are those in the top 10 percent for 5-year citations among articles in their publication year.

** $p < 0.05$.

*** $p < 0.01$.

differently, a 20 percent increase in the probability of abandoning the project. This modest discouragement rate is likely driven by the low cost of publishing once the project has already been deposited in the PDB (recall that in our sample, all scooped projects have already been deposited in the PDB when they learn that they have been scooped). In many cases, the scooped teams may be well into their submission and revision process at the time of being scooped, and therefore will persist to publication. Even if they are rejected from a journal, there are many lower-ranked outlets that may be more willing to accept scooped papers, a mechanism we explore in section IV.C.

In column 2, we estimate a statistically significant penalty in JIF. Scooped papers are published in journals with impact factors 0.19 standard deviations below winning papers. In column 3, this translates to a 6 percentage point (20 percent) decrease in the probability of publishing in a top-10 journal. Column 4 shows that scooped papers face a significant citation penalty as well. The winning projects receive 29 citations on average in the first 5 years. The scooped projects receive 21 percent fewer citations in the same time span. Column 5 suggests that this means scooped projects are 3.6 percentage points (24 percent) less likely to be

one of the top 10 percent of papers in that publication year, ranked by 5-year citations. These results are robust to a variety of cutoffs, including a shorter or longer citation window and different percentiles for the high-citation mark (see table A3). Table A4 shows results are robust to the exclusion of protein (i.e., race) fixed effects. As a further robustness check, we reproduce the regressions using a subsample of races that have projects with 100 percent similar sequence structure according to the algorithm used by the PDB. Table A5 shows that the magnitudes are very similar for all outcomes, even if statistical precision is lower due to the smaller sample size.

Scooped projects may be penalized not only in terms of journal placement and citations but also by less formal means of recognition, such as reader downloads, coverage in the scientific press, and mentions on social media. Scientists value these interactions as they build standing and reputation in both the academic community and general public. Table A6 shows results of project-level regressions using outcomes sourced from Altmetric. We find that getting scooped has statistically significant negative effects on downloads, news mentions, Wikipedia citations, patent citations, and Twitter mentions.

Taken together, these results suggest that there is a significant penalty for being scooped, both in the likelihood of publication, the journal rank of publication, and the number of citations accrued in the early life cycle. However, these results also indicate that the rewards for priority are not winner-take-all. Losing teams receive a smaller, but still substantial share of the credit as measured by publication and citations. Translating the citation penalty to shares of total citations, losing projects receive approximately 44 percent of the total citations accrued to both papers, a much larger share of credit than the zero percent typically assumed by classic models of innovation races.[19]

B. Long-Run Effect on Authors

In this section we analyze the long-run consequences of being scooped on the careers of the various authors of scooped papers following equation (3). Table 5 reports the results of the long-run outcomes regression. Panel A contains results for regressions in the full sample of authors. Panel B restricts to *novices* only, who are defined as authors for whom 7 years or less had elapsed between their first publication and the time

[19] The estimated share of 44 percent is calculated by dividing the mean citations of the losing teams, $28.8\times(1-0.208)$, by the implied total citations, $28.8+28.8\times(1-0.208)$, based on the estimate of the citation penalty from col. 4 of table A6, panel C.

TABLE 5
Effect of Getting Scooped on 5-Year Productivity

Dependent Variable	Any PubMed within 5 years (1)	Any PDB within 5 years (2)	Publication Count within 5 Years: PubMed (3)	PDB (4)	Top-10 Journals (5)	Citation-Weighted (6)	Hit Papers (7)
	A. All Authors						
Scooped	−.018***	−.042***	−1.129	−.072	−.139	−.200***	−.589***
	(.006)	(.010)	(1.046)	(.221)	(.101)	(.045)	(.202)
Winner Y mean	.841	.702	45.869	7.154	3.610	497.203	7.741
Observations	8,624	8,624	8,624	8,624	8,624	6,484	6,484
	B. Novices						
Scooped	−.058***	−.040**	−.019	.003	.106	−.335***	−.181
	(.018)	(.019)	(.276)	(.168)	(.068)	(.102)	(.118)
Winner Y mean	.469	.356	4.243	1.890	.616	75.691	1.165
Observations	2,033	2,033	2,033	2,033	2,033	1,529	1,529
	C. Veterans						
Scooped	−.006*	−.040***	−1.179	−.163	−.235	−.160***	−.821***
	(.003)	(.012)	(1.553)	(.309)	(.145)	(.046)	(.286)
Winner Y mean	.990	.839	61.681	9.261	4.787	667.421	10.388
Observations	5,821	5,821	5,821	5,821	5,821	4,378	4,378

Note.—This table presents regression estimates of the long-run scoop penalty, following equation (3). Observations are at the author level. Each coefficient is from a separate regression. The dependent variable in col. 6 is the total citations accrued in 3 years to all papers published in the 5 years after the race, transformed with the inverse hyperbolic sine function (winner Y means reported in level citations). The dependent variable in col. 7 is the total number of publications that reach the top 10 percent of 3-year citations for their publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as authors with 7 years or less of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all nonnovices). All regressions include author-level covariates selected by PDS Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

of the scoop event.[20] Panel C restricts to *veterans*, who are all authors not defined as novices.[21]

Getting scooped has a statistically significant negative effect on the probability of publishing any subsequent articles in the PDB and PubMed in the 5 years after the race (not including the paper linked to the focal PDB deposits). Column 1 shows that novice scientists who get scooped are 12 percent less likely to have any subsequent PubMed publications and 11 percent less likely to publish any PDB-linked paper in the next 5 years. Although there is not an economically significant negative effect on the extensive margin for veterans in the PubMed data broadly (the estimated effect is less than 1 percent), veterans are 5 percent less likely to publish PDB-linked articles after being scooped. Although veteran careers appear more resilient to being scooped than novice careers, it is possible that getting scooped might encourage some scientists to steer away from the PDB in the future.

Despite a significant extensive-margin effect, we find no significant changes to publication counts on the intensive margin for novices or veterans. Losing teams have no statistically significant differences in publications or PDB-linked publications in the following years as shown in columns 3 and 4, and they are not more or less likely to publish in top-10 journals. This difference in intensive- and extensive-margin effects might mirror a similar dynamic documented by Wang, Jones, and Wang (2019), where scientists that persevere through setbacks (in their case, the denial of a grant) do not experience negative productivity effects in the long run, perhaps due to grit or psychological persistence. However, we do estimate significant penalties in citations for all categories of authors. In the full author sample, the scooped individuals receive 20 percent fewer citations (measured by inverse hyperbolic sine citation-weighted publications) in the next 5 years, where citations are counted up to 3 years after each paper's publication. This effect falls particularly hard on novices, who receive 34 percent fewer citations, while veterans receive only 16 percent fewer citations. The effect on hit papers is reported in column 7 and also suggests that getting scooped decreases attention to future work. The full sample of scientists publish 0.59 fewer hit papers in the five years following a scoop event. The negative effect is lower for novices in levels (0.18 papers versus 0.82 papers for veterans), and not statistically significant for novices. However, if we scale the effect size by the average number of hit papers, the effect is larger for novices (a 16 percent decline versus an 8 percent decline). We also consider outcomes in the following

[20] Seven years is the 30th percentile of the distribution of years since first publication.

[21] The sum of the sample sizes in panels B and C is smaller than the sample size in panel A because the race fixed effects specification in practice restricts identification to races that have at least one novice (or veteran) in the winning and losing team of each race.

three years in table A7 and ten years in table A8. The results are similar in the 3-year window, but are smaller and imprecise after 10 years, in part because we restrict to a smaller balanced sample of races that ended before the last 10 years of our sample window. Finally, in table A9 we restrict to first, middle, and last authors separately because first and last authors are considered to have a larger reputational stake in life science papers, but we find broadly similar effects for all types of authors.

C. Mechanisms: Role of Scoop Timing in the Publication Process

Scooped projects receive about 21 percent fewer citations than their winning counterparts, suggesting that academic researchers pay less attention to the projects that are scooped. In this section, we investigate how the editorial process affects the scoop penalty, and we argue that journal placement is a primary driver of the citation penalty. Further, the size of the penalty is highly correlated with the timing of races. Teams that are scooped early (very shortly after they deposit their findings) receive a much larger penalty than teams that are scooped late (shortly before publication). We provide evidence that top journal editors are unlikely to accept scooped papers; therefore, scooped papers consistently fall to lower-ranked journals, excepting those already deep into the review process at the time they were scooped. These results suggest that editors and reviewers are key policymakers in determining the distribution of academic credit for novel research.

1. Decomposing the Citation Effect by Journal

First we show that the citation penalty is largely driven by journal placement. We decompose the citation effect into an editor/reviewer effect and a reader effect by controlling for journal placement. Column 1 of table 6 replicates the citation penalty effect from column 4 of table 4, but uses a subsample of races in which both papers were published in ranked journals. When both papers are published, the citation penalty is 16 percent for scooped papers. In columns 2 and 3, we add controls for JIF, first as a linear term and then as a cubic polynomial. The citation effect falls to 10 percent, but remains statistically significant. Finally, in column 4 we include journal fixed effects to control completely for any direct effect of the publication outlet on citations. The effect falls to 4 percent. These results suggest that nearly three-fourths of the citation penalty comes through the channel of the publishing journal. Any remaining effect on citation attention comes through readers differentially citing winning and losing papers in similar journals.

TABLE 6
DECOMPOSING CITATION AND JOURNAL EFFECT

	5-YEAR CITATIONS			
DEPENDENT VARIABLE	(1)	(2)	(3)	(4)
Scooped	−.155***	−.111***	−.102***	−.044*
	(.032)	(.029)	(.028)	(.027)
Journal controls	None	Linear JIF	Cubic JIF	Journal FE
Winner Y mean	34.7	34.7	34.7	34.7
Observations	1,891	1,891	1,891	1,891

NOTE.—This table reports the scooped coefficients in regressions with 5-year citations as the outcome where we control for JIF. The citation counts are transformed with the inverse hyperbolic sine function in the regression, but the winner Y mean is reported in levels for ease of interpretation. The regression sample is restricted to races where both papers were published in a ranked publication. Column 1 reestimates the table 4, col. 4 regression in this subsample. Columns 2 and 3 add linear and then cubic controls for JIF. Column 4 includes fixed effects for journal. All regressions also include PDS-Lasso–selected controls and protein (i.e., race) fixed effects.

* $p < 0.1$.

*** $p < 0.01$.

2. Editors' Role in Priority Credit

We further explore the role of editors in adjudicating priority credit by focusing on the submission, review, and publication timelines of scooped projects submitted to leading science journals. Academic journals compete fiercely to publish the highest quality and most novel scientific articles. Many of these journals have explicit policies for accepting only highly original and novel research. For example, *Science* provides the following guidelines to peer reviewers: "Recommend in your review whether the paper should be published in *Science* and provide a more detailed critique based on the following. . . . Novelty: Indicate in your review if the conclusions are novel or are too similar to work already published" (AAAS 2019). Editors and reviewers therefore likely drive much of the scoop penalty if they choose to reject scooped papers when they come across their desk. In this section we look at how the scoop penalty is affected by the timing of journal submissions. Many of the papers in our sample had already been submitted to a journal when they were scooped, and a few papers had already been accepted. Even if an editor would prefer to reject a scooped paper, they may be unable to do so if the paper had already been accepted or was far along in the review process. We use the supplementary data collected from journal websites to examine how the scoop penalty is affected by the timing of the review process. Ideally, we would compare the scoop date to rejection dates at leading journals, but data on rejected papers is not publicly available. Therefore, we use instead the timing of submission and acceptance to present suggestive evidence that editors at top journals are reticent to publish scooped papers.

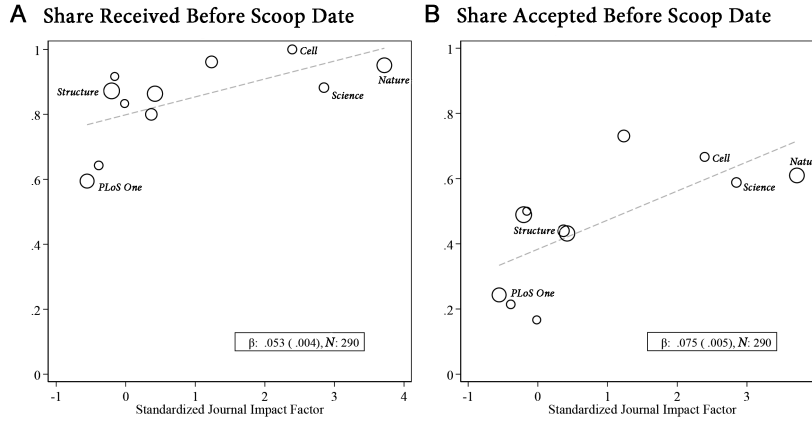

FIG. 5.—Journal placement and timing of scoops. The figure reports the share of scooped papers that were received and accepted before the scoop date at different journals. Each circle represents one of the 11 largest journals for which we collected supplemental data on the editorial timeline. Journals are arranged along the *x*-axis by their standardized JIF. The size of the circles is proportional to the number of scooped papers published in each one.

In our data, scooped papers occasionally appear in top journals like *Science*, *Nature*, and *Cell*, but 90 percent of those papers were already under review on the date that they were scooped. Furthermore, about 60 percent of those papers were scooped after they had already been accepted. Figure 5 further shows that this pattern varies greatly by the impact factor of the journal that eventually publishes the scooped paper. For lower-ranked journals, such as *PLoS One*, only 60 percent of scooped papers had been received by the journal on the date they were scooped, and just over 20 percent had been accepted. Among the 11 large journals for which we have information about received and accepted dates, there is a positive and statistically significant relationship between the share accepted before the scoop date and the impact factor: a scooped paper published in a journal whose rank is 1 standard deviation higher is 8 percentage points more likely to have already been accepted on the scoop date. Although we cannot directly observe scooped papers being rejected from these journals, we can infer from this pattern that top journals are less willing to accept papers that were scooped before submission or early in the review process. Many of these scooped papers fall to lower-ranked general interest journals or highly specialized structural biology journals.[22] Some of these lower-ranked journals, such as *PLoS Biology*, have

[22] One possible strategy a team might consider to win a race is to submit to a lower-ranked journal that has a faster average review time. Indeed we find that top-ranked journals take about 120 days on average from submission to acceptance while lower-ranked

explicit policies of accepting scooped papers. *PLoS Biology* editors write, "Just as summiting Everest second is still an incredible achievement, so too, we believe, is the scientific research resulting from a group who have (perhaps inadvertently) replicated the important findings of another group. To recognize this, we are formalizing a policy whereby manuscripts that confirm or extend a recently published study ('scooped' manuscripts, also referred to as complementary) are eligible for consideration at *PLoS Biology*" (PLoS Biology Staff Editors 2018). But even some lower-ranked journals are concerned about the fierce competition for novel research. When we approached one publisher about sharing their data on received and accepted dates, they only offered to provide the data anonymously, stating their concern about presenting public evidence that they publish scooped papers.

3. Time Lag and the Scoop Penalty

The severity of the scoop penalty is correlated with the time lag between the release of the winning and losing projects. In figure 6, we plot the difference in outcomes separately for 3 terciles of races divided by the time between the release dates of the winning and losing projects. The points are placed on the *x*-axis at the average delay time within the subset of races. The first panel shows the JIF penalty and the second panel shows the citation penalty. Both plots have a strong decreasing trend in the penalty—in other words, the longer the lag between the priority paper and the scooped paper, the less credit the scooped paper receives. The JIF penalty is 0.1 standard deviations in the first 3–4 months, then drops to 0.3 standard deviations by 8 months. Similarly, projects released within 1 month of each other have no difference in citations. The scoop penalty grows to 50 percent for scooped projects with an 8 month delay. In fact, much of the negative effect that we present in table 4 is driven by the tercile of races with the longest delays. An important caveat to these results is that the delay to release after being scooped is potentially endogenous. While much of release lag may be due to idiosyncrasies of the publication process that are out of the researchers' hands, teams may also make strategic decisions about whether to rush to publish, revise and delay, or give up publication altogether, so the delay times should be viewed as potentially selected on team or project characteristics. We explore some of these forces in more detail in the next section. These results suggest that the delay time between projects is relevant for editors and readers, perhaps

journals take about 90 days on average. However, as the results in table 6 show, the bulk of the scoop penalty is due to journal placement, suggesting that the citation-maximizing strategy is to submit to the best possible journal first, despite the potential for a slightly longer review.

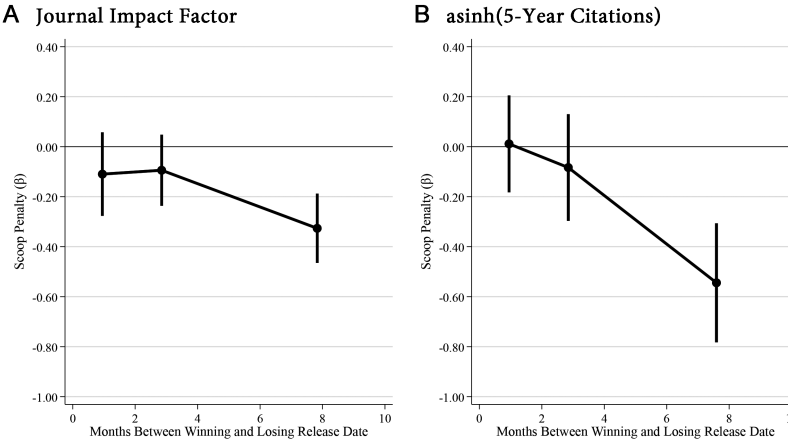

FIG. 6.—Journal impact factor (JIF) and citation penalty by scooped project release delay. The sample of races is divided into 3 terciles along the distribution of time between winning and losing release date. Races are positioned along the *x*-axis at the average scoop release delay within each group. Projects released in close proximity are to the left, and those with a long delay are to the right. The *y*-axis shows the difference in JIF (*A*) and citations (*B*) between the winner and loser.

because the community can more clearly attribute priority credit with more time separating similar projects.

V. Strategic Responses to Getting Scooped before Project Completion

Thus far, we have focused exclusively on races where two teams had completed the project before the knowledge of the scoop is revealed. We chose this restriction because it minimizes the scope for researchers to endogenously respond to the scoop event. In cases where scientists are scooped after depositing, they are usually preparing a manuscript or have submitted to a journal already. The PDB also mandates that the project is released to the public 1 year after deposition at the latest, and this forced disclosure likely puts pressure on the team to publish quickly if they have already deposited. Therefore, they have less flexibility to respond to the scoop event by repositioning their research, changing direction, using insights from the winning paper, or abandoning the project altogether. This allows us to estimate the impact of being scooped, all else equal. However, the endogenous response itself is interesting. How do scientists use the knowledge that they have been scooped to reoptimize? In this section, we compare projects that were scooped before and after deposit to show how scientists respond when they learn that they have been scooped before completing the project.

The classic patent-race literature has focused on the strategic decisions of a follower in a race for a discontinuous reward, typically the profits from a patent (Loury 1979; Dasgupta and Stiglitz 1980; Lee and Wilde 1980; Gilbert and Newbery 1982; Reinganum 1983). Depending on the modeling assumptions, these models predict a range of outcomes: for example, the follower will persist at a steady R&D pace, the follower will increase effort in an attempt to leapfrog the leader, or the follower will choose to drop out of the race altogether. The optimal strategy is dependent on the R&D technology, the information structure of the game, and other features such as whether the race has a single or multiple stages (Fudenberg et al. 1983). Our setting differs from those models for important reasons, but insights from this literature are relevant for interpreting scientist behavior in our setting, especially for those scooped before they had deposited.

Like these classic innovation-race models, researchers in our setting can choose to accelerate a research project or abandon it altogether. However, there are other important choice margins in our setting. First, unlike the models described above, the game does not automatically end when the first team releases their structure. Instead, the second-place team still has an opportunity to adjust the pace, direction, and scope of their project. This is more akin to recent patent-race models where races are multistaged or endless (Judd 1985; Aoki 1991; Doraszelski 2003; Horner 2004). Second, early models rarely grappled with the public goods nature of innovation, where a loser can benefit from the winner's discovery through imitation or improvement of the winner's disclosed discovery (Arrow 1962; Dasgupta and David 1994).

In the remainder of this section, we study the strategic decisions of a scientist who is scooped early enough in the project's life that she still has an opportunity to reoptimize the path of the project. The key idea is that once a scientist learns that she has been scooped, she faces a trade-off. She knows that on one hand, she will get more credit if she publishes quickly because the scoop penalty grows with time (as shown in fig. 6). On the other hand, she can expand the project in new directions (e.g., by adding additional structures or experiments). This will take time—leading to a larger penalty—but will also make the project more valuable overall. Moreover, because she can now take advantage of informational spillovers from the first paper, it might be easier to expand the project than before that paper was released. We formalize this trade-off in appendix C. Broadly speaking, there are three possible cases. In one case, the scientist speeds up when she learns that she has been scooped to minimize the penalty. In a second case, she slows down and improves or broadens her project to maximize its value. Finally, it is possible that the cost of completing a project is no longer offset by the reward, leading to a third case where she abandons the project upon learning it has been scooped.

In our data, we can observe the behavior of some scientists that were scooped before they had a chance to complete their projects, and thus have the flexibility to reoptimize. Although not required by the PDB, many deposits (81 percent) report a collection date, which is the date that the scientist collected the x-ray diffraction data at a synchrotron. Using these dates, we can identify races where scientists had successfully crystallized their protein and collected diffraction data but then learned they were scooped by another team prior to depositing their completed structure model (see fig. 2).

Overall, the empirical evidence is consistent with the second case: researchers spend longer to expand the scope of their projects when they know they have been scooped. Figure 7 compares the timeframe of projects between postdeposit scoops (our original sample) and predeposit scoops. On the left, we show the number of years that pass between the original collection of the data and the time of being scooped. Not surprisingly, predeposit scoops tend to be slightly earlier in the life of the project (mean of 1.6 years for predeposit scoops and 1.7 for postdeposit scoops). There are very few projects that have a short (less than 4-month) lag between collection and scoop in the postdeposit sample of races because the scientists would not have had time to analyze the experimental data and deposit their structure. However, there is considerable overlap of the distributions, suggesting that these two types of scoops occur in similar timeframes on average.

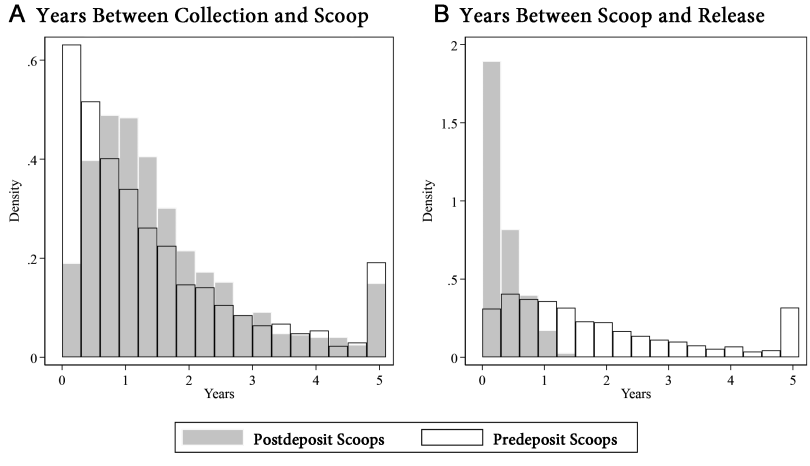

FIG. 7.—Gaps between collection, scoop, and release for pre- and postdeposit scoops. The figure shows the amount of time that passes between the collection date and the scoop (*A*) and between the scoop date and release date (*B*) for predeposit and postdeposit scoops. Histogram is top-coded at 5 years.

TABLE 7
STRATEGIC RESPONSES TO PRE- AND POSTDEPOSIT SCOOPS

		Proteins in Paper		Paper Title Keywords		
Dependent Variable	Maturation	Count	Multiple	"Structure"	"Function," "Mechanism," or "Analysis"	Molecular Replacement
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.173***	1.395***	.437***	.876***	.156***	.614***
	(.068)	(.088)	(.013)	(.016)	(.010)	(.010)
Scooped	.025	−.041	−.005	−.017	.001	.022
	(.062)	(.072)	(.021)	(.016)	(.017)	(.017)
Predeposit × Scooped	1.362***	.201**	.061**	−.062***	.042**	.080***
	(.095)	(.090)	(.027)	(.021)	(.021)	(.021)
Observations	5,398	5,398	5,398	5,398	5,398	5,398

NOTE.—This table presents regression estimates of strategic response outcomes on a scooped indicator and an interaction between a predeposit indicator and the scooped indicator. Predeposit scoops are those where the scooped team had collected data but not yet deposited at the time of the first paper release. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. All regressions include controls selected by the PDS-Lasso method. Standard errors are in parentheses and are clustered at the race level.

** $p < 0.05$.

*** $p < 0.01$.

However, the right panel shows the number of years between the scoop and final release of the scooped paper. This release gap is much longer on average for predeposit scoops (mean of 0.36 years for postdeposit scoops, 2.13 for predeposit scoops), suggesting that, for scientists who know they have been scooped but decide to continue, the preferred strategy is to invest more time into the project rather than abandon it. One important point of context is that postdeposit scooped projects are mandated to release the findings after 1 year, so even if postdeposit scooped teams wanted to change their research, add experiments, or rewrite their paper, they have much less flexibility after they have already deposited.

This delay in release appears to be consistent with scientists electing to add additional experiments and differentiate their project from the race winner. Table 7 presents regression results using the full sample of races associated with 1,778 predeposit scoops and 979 postdeposit scoops combined for which we have available data.[23] We regress a series of project characteristics that relate to the margins of adjustment discussed above on a scooped indicator and an interaction between a predeposit indicator and the scooped indicator.[24] In column 1, we can see that there is a very

[23] There are some cases where there is one predeposit scoop and one postdeposit scoop in the same cluster, i.e., scooped by the same priority deposit. For clarity in the regression specifications, we drop the predeposit scoops from these clusters.

[24] The predeposit main effect is absorbed by the protein fixed effect.

large increase in maturation time (time between collection and release) for the predeposit scooped teams relative to the postdeposit scooped teams, with the predeposit teams spending 1.4 more years on average. Next, we consider how trailing scientists may adjust the scale or scope of their research to offset the scoop penalty. We find in columns 2 and 3 that predeposit scooped teams are much more likely to include multiple protein structures in their paper relative to the postdeposit scooped teams, suggesting that predeposit scooped teams expand the scope of their papers. Next, we look at how scooped teams may have adjusted the content of their paper by analyzing keywords from the paper titles. Column 4 suggests that predeposit scooped teams are much less likely to use the word "structure" or "structural" than postdeposit scooped teams. However, as seen in column 5, they are more likely to use words like "function," "mechanism," or "analysis" in the title. It appears that if teams have the flexibility to adjust the direction of their research after being scooped, they choose to shift the focus away from the structure determination itself and toward describing the function or biological mechanisms that the structure implies.

Finally, we use another unique feature of the data to test whether trailing teams benefited from seeing the priority deposit if they were scooped before completing their own work. The PDB contains a flag for a technology called molecular replacement, which is a crystallography technique that improves model prediction. Importantly, it relies on using another similar structure model as a pattern to refine the new model from diffraction data (see Kim 2023 for a detailed explanation of the technology). In other words, trailing teams can use molecular replacement—which makes completing their projects easier—if they can observe the winning structure before they finish their own structure. Column 6 suggests that scooped teams are more likely to use this technology, but *only if they were scooped before they deposited.* If the winning structure is released after the trailing team deposits, as is the case in our postdeposit sample, the trailing team is unable to take advantage of insights from the winning structure model in their own process. However, if the winning structure is released before the trailing team deposits, as is the case in our predeposit sample, they are able to benefit from this information. This suggests there are meaningful knowledge spillovers that benefit the losing team. In addition to being consistent with our model, we think this provides strong empirical evidence that the release of the project represents a meaningful information shock. Overall, it appears that scientists who are scooped before they have a chance to deposit their findings are more likely to delay the release of their structure, increase the scope or change the direction of their research, and integrate the knowledge from the first discovery into their project.

Finally, we compare the cost of being scooped before and after deposit. We interpret the results of this exercise cautiously because of the

endogenous selection into the predeposit sample and the additional flexibility that predeposit scooped teams have to strategically respond to the scoop. Table A10 reproduces table 4 in the predeposit sample. We find that the difference in most outcomes between the winners and losers is about 15 to 40 percent larger in predeposit scoops compared to our primary postdeposit sample. The citation gap is 28 percent in the predeposit scoops compared to 21 percent in the postdeposit (main sample) scoops. The relative reduction in the probability of publication is comparable between the two groups. Scientists who persist despite being (knowingly) scooped are likely a selected set who are determined to publish.

VI. Reputation and the Scoop Penalty

Scientific races provide a unique setting to study how academic recognition is affected not only by priority, but also by the preexisting reputations of winners and losers. We find that when a high-reputation team scoops a low-reputation team, they receive 65 percent of the total citations, but when a low-reputation team scoops a high-reputation team in a comparable race, they only receive 46 percent of the total citations. This asymmetry in attention suggests that the distribution of priority rewards is not formulaic and may be affected by the institutions, norms, or biases of the academic community. In appendix C, we present a model of academic attention based on a standard statistical discrimination model (Aigner and Cain 1977). Here we present empirical results that support the predictions of the model.

Priority rewards are allocated by a decentralized set of actors, including journal editors and readers, in a market for academic attention. Because scientists have limited time for reading and reviewing new papers, it may be difficult to determine the quality of new research. Therefore, editors and readers may rely on signals of ability based on the reputation of the researchers or their institution to supplement their judgement of a paper's quality. The model considers cases where two types of teams, high- and low-reputation, publish identical papers. Readers decide who to cite based on priority and reputation. In cases where teams are of the same type, the priority effect is isolated, and the first team to publish receives more than 50 percent of the total citations. However, in cases where teams are of different types, the priority and reputation effects will either work in the same or opposite direction, depending on which team finishes first. If the high-reputation team wins the race, the two effects reinforce each other, meaning the high-ranked team will have an equal or greater share of citations compared to the low-ranked team than they would competing against another high-reputation team. If the low-reputation team scoops the high-reputation team, the net effect is ambiguous. If the reputation effect is stronger than the priority effect,

the low-reputation team may receive less than 50 percent of the total citations, despite publishing first.

To test our model, we measure the share of total citations received by winning and losing labs, and compare these shares in races where the reputation varies between the two racing teams. More specifically, if lab A and lab B race to write a paper about the same protein, we compute $\text{CitationShare}_A = \text{Citations}_A/(\text{Citations}_A + \text{Citations}_B)$. This citation share maps to the probability of citation outlined in the model above.[25]

We proxy for the preexisting reputation of each lab using the Lasso-estimated predicted citations from the nonracing data sample as described in section III.B. Labs with above-median predicted citations are categorized as *high-reputation* labs, while teams below median are called *low-reputation* labs. In figure 8 we plot the predicted citations of the losers on the *x*-axis and the predicted citations of the corresponding winners on the *y*-axis. Each point on this scatter plot represents the observed match between two racing labs. If all labs were equally matched in preexisting reputation, all points would lie on the dashed 45-degree line. Of course, labs are rarely perfectly matched in the data, providing variation in the difference of reputation between the winners and losers.

The median lines in figure 8 conveniently partition the sample into four subsamples that line up with the four types of "matchups" we discuss in our model. The top right and bottom left corners represent subsamples of closely matched races where both labs were either high-reputation or both low-reputation. The top-left and bottom-right subsamples represent mismatched races where an above-median team scooped a below-median team and vice versa.

In mismatched races, we interpret the difference between citations as being caused by an additive effect of priority and reputation. One potential confounder in that interpretation is that high- and low-reputation teams might produce different quality of scientific outputs for the same structure discovery. If high-reputation teams produce higher-quality or more convincing results, then the additional citations they receive may not be caused by their high-profile reputation alone. Although it is difficult to quantify all aspects of paper quality, we examine two important measures of quality reported by the PDB: resolution and *R*-free (goodness-of-fit), described in more detail in section III.C. Table A11 compares the average resolution and *R*-free of the winning and losing structures in each of the four subsets of races. We find very little evidence of statistical difference in quality metrics between high- and low-reputation teams

[25] The model does not include the possibility of cocitations, where both papers are cited together, but the empirical results are proportional to an analysis where cocitations are excluded.

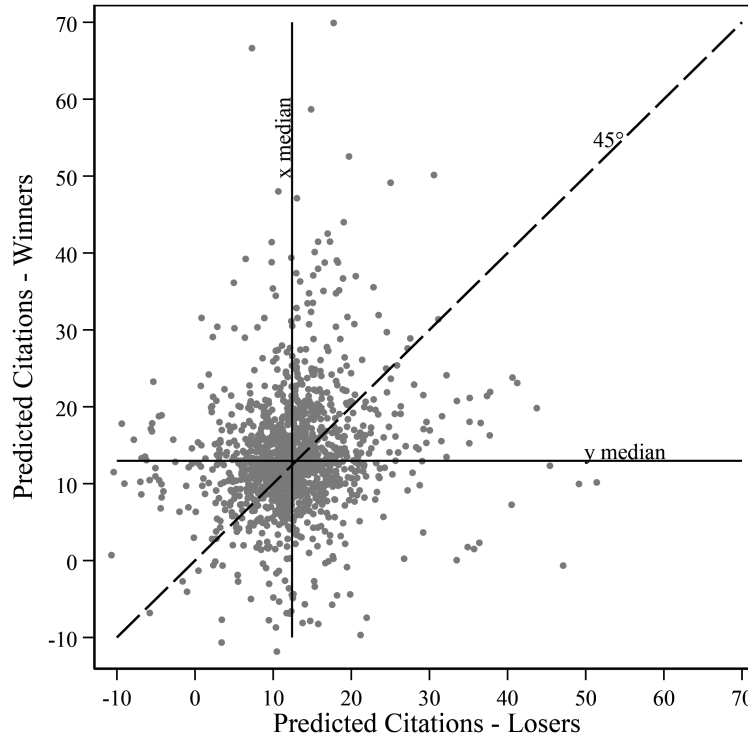

FIG. 8.—Scatter plot of team-reputation difference. Each observation in this figure is a racing pair. The *y*-axis shows the predicted citations for the winning team, and the *x*-axis shows the predicted citations for the losing team. If the winning team has higher predicted citations than the losing team, the dot will lie above the 45-degree line. If the winning team has lower predicted citations than the losing team, the dot will lie below the 45-degree line. Perfectly matched teams would lie on the 45-degree line.

engaged in a race. This suggests that any difference in citations is not driven by the quality of science that each team is producing.

Figure 9 shows the average citation counts by matchup type, as well as the citation shares. Panel A shows the evenly matched races, which isolates the priority effect. As predicted by the model, the winning labs receive more citations. Moreover, if we look at the share received by the winning team, we see that it is identical in the high-versus-high–reputation matchups and the low-versus-low–reputation matchups (in each case, the winning team receives about 55 percent of the total citations), which is consistent with the model.[26]

[26] The restriction to evenly matched teams in panel A is also a convenient check on the identification assumptions for a causal interpretation of the estimated scoop effect. Even when competitors are well-matched on observables, there exists a statistically significant priority premium that is unlikely to be driven by positive selection of winners.

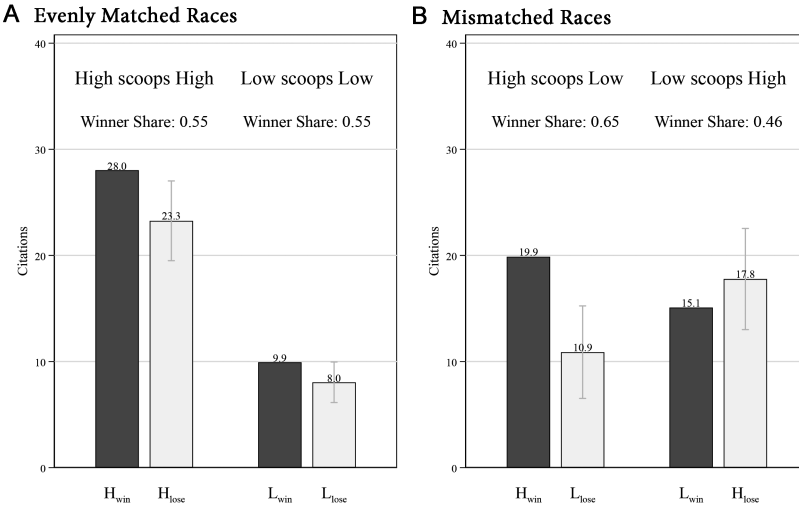

FIG. 9.—Priority effect by reputation matchup. We divide the sample of races from fig. 8 into four quadrants, depending on whether the winning and losing teams are above median (high reputation) or below median (low reputation) in expected 3-year citations defined by the Lasso estimation. Dark gray bars represent the actual citations of the winning team and light gray bars of the losing team. *A*, The comparison between evenly matched races, in which a high-reputation team scoops a high-reputation team or a low-reputation team scoops a low-reputation team. *B*, Comparison between mismatched races, in which a high-reputation team scoops a low-reputation team or a low-reputation team scoops a high-reputation team. The winner's share of total citations are reported above each set of bars.

Panel B shows the unevenly matched races. When a high-reputation lab scoops a low-reputation lab, the priority effect and the reputation effect work in the same direction. Here we see that, consistent with proposition 2, the winning team receives an even larger share of the total citations (65 percent). Conversely, when a low-reputation lab scoops a high-reputation lab, the priority effect and the reputation effect move in opposing directions. In this case, it appears that the reputation effect is the stronger of the two, with the winning team receiving less than half (46 percent) of the total citations. Again, this matches the prediction outlined by proposition 2 of the model.

Collectively, we interpret this as evidence that statistical discrimination based on prior lab reputation can rationalize our heterogeneity results. The lack of symmetry exhibited in panel B suggests that being first is not the sole determinant of credit in science. In science, there is no central arbiter that gives legally binding credit or property rights to the first-place team. Here the teams vie for attention, and although the low-reputation teams may benefit by winning a race, there appears to be

built-in inequality in attention that prevents them from capturing as much of the credit as their high-reputation competitors.

VII. Benchmarking Magnitudes: Survey Results

We estimate that getting scooped causes a decrease in the probability of publication, leads to publication in lower-impact journals, and reduces citations. However, priority races are not winner-take-all. Our citation estimate suggests that winners get 56 percent of the total citations, a far cry from the 100 percent often assumed in the theoretical literature. But how does this estimated share of credit compare to scientists' beliefs? In an email survey of structural biologists, we pose a hypothetical situation about a late-stage race to publication. The full text of the questions can be found in appendix D. First we ask, "Suppose you have just completed a very promising research project . . . what do you think is the probability that your project will be scooped between now and when it is published?" We next state that their hypothetical project has indeed been scooped by a paper in the journal *Science*. In this scenario, we ask them the following questions: "Would you choose to abandon your manuscript? Assuming you submit, what is the probability the article will eventually be published? What is the best journal that would accept your paper? If your competitor receives 100 citations, how many citations do you expect your publication to receive?"

Table 8 reports the average responses of the biologists in columns 3–6 and compares them to the magnitudes estimated in the PDB data in columns 1 and 2. The hypothetical scenario in the survey was designed to match the instances of racing that we have in our data. However, because we tried to pose the survey questions as concretely as possible for clarity, the racing situation does not exactly match the average situation in the PDB. In particular, in the survey the losing team is scooped early in the submission process, and the project is very high quality, with an expected journal placement in *Science*. Therefore we report estimates in column 2 from a subset of the PDB data where (1) the losing team is scooped soon after they deposit their data,[27] and (2) one of the teams published in one of the three highest-impact journals (*Science*, *Nature*, or *Cell*). These restrictions make some of the PDB estimates smaller or larger, but we still consistently find evidence of pessimism among respondents. Surveyed scientists report a 27 percent chance of being scooped between submission and publication, more than three times the 8 percent scoop probability

[27] Specifically, we sort races by the time elapsed between the loser deposit date and the winner release date and keep the one-quarter of race losers that were scooped earliest in the process.

TABLE 8
SURVEY BENCHMARK OF SCOOP PENALTY

	PDB ESTIMATE		SURVEY ESTIMATE			
	Full Sample (1)	Comparable Subsample (2)	All Respondents (3)	Low-Reputation (4)	High-Reputation (5)	Difference (Col. 4 − Col. 5) (6)
Prob(Scoop)	.029	.081	.266	.267	.266	.001
						(.016)
Prob(Publication)	.854	.976	.665	.636	.694	−.059***
						(.022)
JIF penalty	−.186	−1.208	−2.918	−2.937	−2.900	−.036
						(.084)
Citation penalty	−.208	−.135	−.594	−.614	−.575	−.040*
						(.024)
Observations	67,933	3,152	822[1]			

NOTE.—This table reports results from the PDB (cols. 1 and 2) and from a survey of structural biologists (cols. 3–6). In col. 1, we report the mean values of a scoop indicator in the full sample. The remaining estimates are estimated identically to table 4, panel C. In col. 2, we repeat this procedure in a subsample of the PDB that is comparable to the survey text. Specifically, we restrict to PDB races where one racer published in *Science*, *Nature*, or *Cell* and the losing team was scooped early in the process (quarter of sample with the shortest time between loser deposit and winner release). In col. 3, we report the results of a survey of structural biologists. The survey asked respondents to estimate the probability and consequences of getting scooped on a hypothetical project. See app. D for full survey text. In cols. 4 and 5, respondents were divided into two groups, high- and low-reputation, using the predicted citations measure used for heterogeneity in section VI of the text. Column 6 reports the difference in response means between cols. 4 and 5 and reports the heteroskedastic-robust standard error in parentheses.

* $p < 0.1$.

*** $p < 0.01$.

1 Note that not all respondents answered all questions, so the sample size varies across rows: $N = 768$, 672, 597, and 675, respectively).

in the comparable PDB sample.[28] Six percent of respondents report that they would abandon the project, but only 68 percent think they would succeed at publishing conditional on submitting, implying a 67 percent unconditional probability of publishing as shown in column 3. This is much lower than the 85 percent of scooped papers that are actually published in the PDB data, and the 98 percent that are published in the comparable subsample. Scientists are very pessimistic about the potential journal placement of scooped papers, expecting that the best journal they could publish in would be almost 3 standard deviations below *Science*, which has a standardized impact factor of about three in most years. Finally, we ask about expected citation effects. When asked to guess the number of citations they would receive compared to the hypothetical winner's 100 citations, the average guess was only 41 citations, which translates to a 59 percent penalty or a share of 29 percent of the total citations. The corresponding estimate in the PDB is no more than a 21 percent penalty or a 44 percent share. Ultimately, PDB scientists expect much worse consequences from being scooped than can be found in the data.

Table 8 also reports survey responses separately for high- and low-reputation scientists. We split the survey sample using the same Lasso-predicted citation measures used in section VI. Column 4 reports the average responses for below-median–reputation scientists, column 5 reports the average responses for above-median–reputation scientists, and the difference with standard errors is reported in column 6. High- and low-reputation respondents predict equal probabilities of being scooped. Low-reputation respondents are more pessimistic however about the probability of publishing conditional on being scooped, with 7 percentage points lower probability that they will be able to publish their scooped paper. Perhaps surprisingly, both types of respondents had similar expectations for the types of journals that they would publish in, all expecting that the scooped papers would fall to field journals or middling general interest journals with average impact factor. But they again depart on their expected citations, with high-reputation scientists expecting a citation penalty that is about 4 percentage points smaller than low-reputation scientists (57.5 percent penalty vs. 61.4 percent penalty). This difference in expectations is consistent with our results about the role of reputation in determining priority rewards. Since both types of authors suggest they would submit to similar journals, it may be that the difference in citations

[28] One caveat to this comparison is that we identify scooping papers in the PDB that have a very specific and perhaps narrow type of overlap, a structure determination for a protein that is similar enough in amino acid sequence to register in our cluster definitions. It may be that a scientist could see other types of papers related to their protein that have conceptual overlap that is different than the dimension we are measuring, which might explain why they report a higher probability of being scooped in expectation than we observe in the PDB.

is driven by statistical discrimination of editors, reviewers, and readers as explained in the model in section VI. It appears that although all scientists are pessimistic about the cost of getting scooped, less-prominent authors are particularly concerned. Our estimates of significant inequality in citation patterns suggest that these beliefs may be justified.

VIII. Conclusion

Priority races are a common feature of academic science, and credit for priority is considered an important motivator for the generation of new knowledge. Yet, we have little empirical evidence on how these priority rewards are structured. Racing is hard to analyze empirically because proximate research projects are difficult to link in data and many scooped projects are abandoned before entering the scientific record. This paper makes progress on these empirical challenges by focusing on project-level data in a setting that captures the near universe of completed projects in structural biology. By taking advantage of the unique data collected by the PDB, we are able to construct credible estimates of the priority premium in the field of structural biology. We find that rewards are far from winner-take-all; rather, our preferred estimates suggest a 56-44 split in citations between the winning and losing paper.

This paper contributes to our understanding of the role of priority and the structure of incentives in basic research. Academic science is an atypical marketplace of productive activity. New ideas are valuable for the world but are not immediately marketable, and are therefore unlikely to be produced by private firms or individuals seeking profits. A patent system is therefore a less-effective instrument for encouraging investment, risk-taking, effort, or disclosure of scientific studies. Instead, a system of priority rewards has developed to encourage research investment, which is reinforced through norms in the scientific community. Individuals who produce new knowledge are given credit by the community that can accumulate into a reputation that likely has both intrinsic and monetary value to the scientist. Although R&D races have been posed as winner-take-all tournaments in past literature, we find that priority rewards are not winner-take-all, but are potentially still an important motivator of both effort and novelty in science. Even if the result of one race has a small impact on careers, the accumulation of credit may still be important.

In this paper, we establish that priority is a relevant incentive in science, but we do not analyze the overall welfare implications of the priority system and size of the priority premium, nor do we consider alternative systems or policies. How would a larger or smaller priority premium affect the efficiency of science? There are many margins to consider, including how changes would affect effort, project selection, collaboration, and even participation in science. A particularly interesting concern raised

in popular and academic writing is that priority may be pursued at the expense of quality. Racing to complete projects may stimulate effort and hasten the pace of discovery, but it may lead scientists to cut corners on the quality of the results that they disclose. If the incentives for replication are low and the costs of replication are high, science as a whole may suffer as quick and sloppy research becomes the norm. In Hill and Stein (2024a), we analyze objective measures of the quality of crystal diffraction data and corresponding structure models to study how racing in science affects quality outcomes. We find that proteins with high ex ante potential have more competitors racing to complete the structure, are deposited faster, and are completed with lower quality. This evidence suggests that racing in science does indeed hasten disclosure but has negative effects on quality. Concerns about the cutthroat nature of racing have led to suggestions of policies that might dampen the strong incentives for novelty. These include allowing a grace period for journal acceptance a few months after being scooped, providing opportunities to establish priority for early-stage work through preprints, or directly incentivizing replication efforts through directed grant funding.

Finally, the results of our survey suggest that scientists are very pessimistic about the cost and probability of being scooped. If the perceived threat of being scooped has a negative influence on the pace, direction, quality, and openness of science, we believe that this paper should help assuage concerns about competition for priority and foster a more productive research environment.

Data Availability

Data and code for replicating the tables and figures in this article can be found in Hill and Stein (2024b) in the Harvard Dataverse, https://doi.org/10.7910/DVN/TJ5VCW.

References

AAAS (American Association for the Advancement of Science). 2019. "Instructions for Reviewers of Research Articles." American Assoc. Advancement Sci., New York.

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Indus. and Labor Relations Rev.* 30 (2): 175–87.

Aoki, Reiko. 1991. "R&D Competition for Product Innovation: An Endless Race." *A.E.R.* 81 (2): 252–6.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors.* Princeton, NJ: Princeton Univ. Press.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2013. "Matthew: Effect or Fable?" *Management Sci.* 60 (1): 92–109.

Barinaga, Marcia. 1989. "The Missing Crystallography Data." *Science* 245 (4923): 1179.

Bellemare, Marc F., and Casey J. Wichman. 2019. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bull. Econ. and Statis.* 82 (1): 50–61.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *Rev. Econ. Studies* 81 (2): 608–50.

Berman, Helen M., Stephen K. Burley, Gerald J. Kleywegt, John L. Markley, Haruki Nakamura, and Sameer Velankar. 2016. "The Archiving and Dissemination of Biological Structure Data." *Current Opinion on Structural Biology* 40:17–22.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Res.* 28 (1): 235–42.

Bikard, Michaël. 2020. "Idea Twins: Simultaneous Discoveries as a Research Tool." *Strategic Management J.* 41 (8): 1528–43.

Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti. 2017. "Researcher's Dilemma." *Rev. Econ. Studies* 84 (3): 969–1014.

Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt. 2018. "The Matthew Effect in Science Funding." *Proc. Nat. Acad. Sci. USA* 115 (19): 4887–90.

Brown, Eric N., and S. Ramaswamy. 2007. "Quality of Protein Crystal Structures." *Acta Crystallographica* D63:941–50.

Burbidge, John B., Lonnie Magee, and A. Leslie Robb. 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *J. American Statis. Assoc.* 83 (401): 123–7.

Burley, Stephen K., Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, et al. 2019. "RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy." *Nucleic Acids Res.* 47 (D1): D464–D474.

Campbell, Philip. 1998. "New Policy for Structural Data," *Nature* 394 (6689): 105.

Cengiz, Doruk, Arindrajit Dube, Atilla Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *Q.J.E.* 134 (3): 1405–54.

Dasgupta, Partha, and Paul A. David. 1994. "Toward a New Economics of Science." *Res. Policy* 23:487–521.

Dasgupta, Partha, and Joseph Stiglitz. 1980. "Uncertainty, Industrial Structure, and the Speed of R&D." *Bell J. Econ.* 11 (1): 1–28.

Dessailly, Benoît H., Rajesh Nair, Lukasz Jaroszewski, J. Eduardo Fajardo, Andrei Kouranov, David Lee, Andras Fiser, Adam Godzik, Burkhard Rost, and Christine Orengo. 2009. "PSI-2: Structural Genomics to Cover Protein Domain Family Space." *Structure* 17 (6): 869–81.

Doraszelski, Ulrich. 2003. "An R&D Race with Knowledge Accumulation." *RAND J. Econ.* 34 (1): 20–42.

Fermi, Giuilio, Max F. Perutz, Boaz Shaanan, and Roger Fourme. 1984. "The Crystal Structure of Human Deoxyhaemoglobin at 1.74 Å Resolution." *J. Molecular Biology* 175 (2): 159–74.

Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole. 1983. "Preemption, Leapfrogging and Competition in Patent Races." *European Econ. Rev.* 22 (1): 3–31.

Gilbert, Richard, and David M. Newbery. 1982. "Preemptive Patenting and the Persistence of Monopoly." *A.E.R.* 72 (3): 514–26.

Goodsell, David S. 2019. "Methods for Determining Atomic Structures." Technical Report, Protein Data Bank. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure.

Hagstrom, Warren O. 1974. "Competition in Science." *American Sociological Rev.* 39 (1):1–18.
Hill, Ryan. 2019. "Searching for Superstars: Research Risk and Talent Discovery in Astronomy." Working paper, Kellogg Graduate School of Management, Northwestern Univ.
Hill, Ryan, and Carolyn Stein. 2024a. "Race to the Bottom: Competition and Quality in Science." Working paper.
———. 2024b. "Replication Data for: 'Scooped! Estimating Rewards for Priority in Science.'" Harvard Dataverse https://doi.org/10.7910/DVN/TJ5VCW.
Hill, Ryan, Carolyn Stein, and Heidi Williams. 2020. "Internalizing Externalities: Designing Effective Data Policies." *AEA Papers and Proc.* 110:49–54.
Hiruma, Yoshitaka, Mathias A. S. Hass, Yuki Kikui, Wei-Min Liu, Betül Ölmez, Simon P. Skinner, Anneloes Blok, Alexander Kloosterman, Hiroyasu Koteishi, Frank Löhr, et al. 2013. "The Structure of the Cytochrome P450cam–putidaredoxin Complex Determined by Paramagnetic nmr Spectroscopy and Crystallography." *J. Molecular Biology* 425 (22): 4353–65.
Hopenhayn, Hugo, and Francesco Squintani. 2021. "On the Direction of Innovation." *J.P.E.* 129 (7): 1991–2022.
Horner, Johannes. 2004. "A Perpetual Race to Stay Ahead." *Rev. Econ. Studies* 71:1065–88.
Jacob, Brian, and Lars Lefgren. 2011. "The Impact of NIH Postdoctoral Training Grants on Scientific Productivity." *Res. Policy* 40 (6): 864–74.
Jardim, Ekaterina, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wething. 2022. "Minimum-Wage Increases and Low-Wage Employment: Evidence from Seattle." *American Econ. J.: Econ. Policy* 14 (2): 263–314.
Judd, Kenneth L. 1985. "Closed-Loop Equilibrium in a Multi-Stage Innovation Race." Discussion Paper no. 647, Center for Mathematical Studies in Economics and Management Science, Northwestern Univ.
Kim, Soomi. 2023. "Shortcuts to Innovation: The Use of Analogies in Knowledge Production." Working paper, Columbia Business School.
Klebel, Thomas, Stefan Reichmann, Jessica Polka, Gary McDowell, Naomi Penfold, Samantha Hindle, and Tony Ross-Hellauer. 2020. "Peer Review and Preprint Policies Are Unclear at Most Major Journals." *PLoS One* 15 (10): e0239518.
Lee, Tom, and Louis L. Wilde. 1980. "Market Structure and Innovation: A Reformulation." *Q.J.E.* 94 (2): 429–36.
Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND J. Econ.* 28 (2): 228–47.
Loury, Glenn C. 1979. "Market Structure and Innovation." *Q.J.E.* 93 (3): 395–410.
Marder, Eve. 2017. "Scientific Publishing: Beyond Scoops to Best Practices." *Elife* 6:e30076.
Martz, Eric, Wayne Decatur, Joel L. Sussman, Michal Harel, and Eran Hodis. 2019. "Nobel Prizes for 3D Molecular Structure." https://proteopedia.org/wiki/index.php/Nobel_Prizes_for_3D_Molecular_Structure.
Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Rev.* 22 (6): 635–59.
———. 1968. "The Matthew Effect in Science." *Science* 159 (3810): 56–63.
Moult, John. 2005. "A Decade of Casp: Progress, Bottlenecks and Prognosis in Protein Structure Prediction." *Current Opinion in Structural Biology* 15 (3): 285–9.
Nelson, Richard R. 1959. "The Simple Economics of Basic Scientific Research." *J.P.E.* 67 (3): 297–306.

NIGMS (National Institute of General Medical Sciences). 2017. "Structural Biology." Technical report. https://web.archive.org/web/20180123003121/https://www.nigms.nih.gov/Education/pages/Factsheet_StructuralBiology.aspx.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *J. Bus. and Econ. Statis.* 37 (2): 187–204.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *A.E.R.* 62 (4): 659–61.

PLoS Biology Staff Editors. 2018. "The Importance of Being Second." *PLoS Biology* 16 (1): e2005203.

Ramakrishnan, Venki. 2018. *Gene Machine: The Race to Decipher the Secrets of the Ribosome.* New York: Basic Books.

Reinganum, Jennifer. 1983. "Uncertain Innovation and the Persistence of Monopoly." *A.E.R.* 73 (4): 741–43.

Seide, Rochelle K., and Alicia A. Russo. 2002. "Patenting 3D Protein Structures." *Expert Opinion on Therapeutic Patents* 12 (2): 147–50.

Shimbo, Itsuki, Rie Nakajima, Shigeyuki Yokoyama, and Koichi Sumikura. 2004. "Patent Protection for Protein Structure Analysis." *Nature Biotechnology* 22 (1): 109–12.

Stephan, Paula E. 1996. "The Economics of Science." *J. Econ. Literature* 34 (3): 1199–235.

Strasser, Bruno J. 2019. *Collecting Experiments: Making Big Data Biology.* Chicago: Univ. Chicago Press.

Sussman, Joel L. 1998. "What's New at the PDB," *Protein Data Bank Newsletter*, April. https://files.wwpdb.org/pub/pdb/doc/newsletters/bnl/news84_apr98/newslttr.pdf.

Thompson, Neil C., and Jeffrey M. Kuhn. 2020. "Does Winning a Patent Race Lead to More Follow-on Innovation?" *J. Legal Analysis* 12:183–220.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *J. Royal Statis. Soc.* B58 (1): 267–88.

Torvik, Vetle I., and Neil R. Smalheiser. 2009. "Author Name Disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data* 3 (3): 11.

Torvik, Vetle I., Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation." *J. American Soc. Information Sci. and Tech.* 56 (2): 140–58.

Tripathi, Sarvind, Huiying Li, and Thomas L. Poulos. 2013. "Structural Basis for Effector Control and Redox Partner Recognition in Cytochrome P450." *Science* 340 (6137): 1227–30.

Wang, Yang, Benjamin F. Jones, and Dashun Wang. 2019. "Early-Career Setback and Future Career Impact." *Nature Communications* 10: 4331.

Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *J.P.E.* 121 (1): 1–27.

Wlodawer, Alexander, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. 2008. "Protein Crystallography for Non-Crystallographers, or How to Get the Best (But Not More) From Published Macromolecular Structures." *FEBS J.* 275 (1), 1–21.

Worldwide Protein Data Bank (wwPDB). 2019. "wwPDB Policies and Processing Procedures: Release of PDB Entries." https://www.wwpdb.org/documentation/policy.

Yong, Ed. 2018. "In Science, There Should Be a Prize for Second Place." *The Atlantic*, February 1, 2018. https://www.theatlantic.com/science/archive/2018/02/in-science-there-should-be-a-prize-for-second-place/552131/.