

Scooped! Estimating Rewards for Priority in Science*

Ryan Hill[†]

Carolyn Stein[‡]

November 13, 2020

Abstract

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 subsequent publication and career outcomes. To do so, we take advantage of data from structural biology where the nature of the scientific process together with the Protein Data Bank — a repository of standardized research discoveries — enables us to identify priority races and their outcomes. We find that race winners receive more attention than losers, but that these contests are not winner-take-all. Scooped teams are 2.5 percent less likely to publish, are 18 percent less likely to appear in a top-10 journal, and receive 20 percent fewer citations. Getting scooped has only modest effects on academic careers. Finally, we document empirical evidence suggesting that the priority reward system reinforces inequality of attention in science.

*We are very grateful to our advisors Heidi Williams, Amy Finkelstein, Pierre Azoulay, and Josh Angrist for their invaluable mentoring and support. This paper has also benefited from feedback and suggestions from David Autor, Sydnee Caldwell, Jane Choi, Colin Gray, Madeline McKelway, Tamar Oostrom, Christina Patterson, Jim Poterba, Otis Reid, Jon Roth, Adrienne Sabety, Cory Smith, Ariella Kahn-Lang Spitzer, Scott Stern, Liyang Sun, 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 NBER 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. 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.

[†]Northwestern Kellogg School of Management, ryan.hill@kellogg.northwestern.edu. Corresponding author.

[‡]MIT Economics Department, cstein@mit.edu. Both authors contributed equally.

1 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, *Priorities in Scientific Discovery: A Chapter in the Sociology of Science* (1957)

Basic science is a critical input to innovation, but it may be under-provided 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 a 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). Many theoretical papers about innovation races conceptualize the reward structure as winner-take-all (Loury 1979; Fudenberg et al. 1983; Dasgupta and David 1994; Bobtcheff et al. 2016). 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 contribution of this paper is to empirically measure the consequences of getting scooped. 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 the scooped scientist’s subsequent career. We also investigate whether competition for academic attention is a driver of 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 an identical or very similar project. 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 (mostly proteins). 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 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,¹ so in most cases, teams in our data are scooped unexpectedly near the end of their projects.

We construct races using two key dates that are recorded for all PDB projects. First, the deposit date marks when the scientist first uploaded their 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 two to six 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 had deposited a structure discovery for identical macromolecules independently of each other prior to the other competitors’ 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 1,630 races in our data. These races consist of 3,319 separate projects out of 64,018 total projects in our sample period from 1999 to 2017, suggesting that five 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. Our definition of scooped projects focuses only on late-stage races where both teams are on the cusp of publication. Researchers may worry about being scooped earlier in the research process, such as during the design or data collection of an experiment. We cannot systematically identify these events in our data if the first team publishes before the second team deposits. Nevertheless, focusing only on 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.

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

¹In 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 pre-print prior to journal publication.

to the causal interpretation of the estimates. If the winners are positively selected on experience, research ability, or university prestige, our estimates of the scoop penalty will be biased up (in terms of magnitudes). However, we find that the outcome of races — even if not perfectly random — is highly unpredictable. We observe cases of both 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 (Belloni et al. 2014).

We find that getting scooped has a moderate-sized impact on the success of the scooped project. Scooped projects are 2.5 percent less likely to be published. Scooped papers appear in a 0.18 standard deviation lower-ranked journal, and are 19 percent less likely to appear in a top-10 journal. Scooped papers receive 20 percent fewer citations, and are 23 percent less likely to be a “hit” paper, defined as reaching the top 10 percent in citations for that publishing 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 45 percent of the total citations accrued by both papers, a much higher share than the zero percent assumed by a winner-take-all model.

Much of the citation effect is driven by journal placement, with only a five 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.

We also assess the effect of getting scooped on broader measures of attention using alternative outcomes sourced from Altmetric.com. Scooped papers are 45 percent less likely to be downloaded in Mendeley, a popular citation management software. They are 11 percent less likely to appear in a popular press or scientific news story, 4 percent less likely to be cited by a Wikipedia article, and 10 percent less likely to be mentioned on Twitter. Scooped papers receive less attention not just by editors and scientific peers, but the broader scientific community, popular press, and more casual readers.

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 five percent less likely to be actively depositing in the PDB five years after they were scooped, but not less likely to be publishing in life and medical sciences as a whole. We do not find significant effects on intensive margin publication rates. However, scooped scientists receive 17 percent fewer citations to their future work, an effect that is stronger for novice scientists (32 percent) than their veteran co-authors (13 percent).

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 66 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). Our model proposes that readers receive a noisy signal of a paper’s true quality, and therefore place some weight on the authors’ pre-existing reputation. A high-reputation team that wins the race not only receives a premium for priority, but also a boost in citations because of their renown. If a low-reputation team scoops a high team, the winner still receives a priority benefit, but it is fully offset by a penalty for their lower reputation. 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 915 corresponding authors of papers linked to the PDB and pose a hypothetical scenario about getting scooped. The respondents estimate a 25 percent probability of getting scooped between submission and publication, much larger than the three 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 66 percent chance of publishing the paper, again much lower than the 86 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 20 percent penalty we estimate in the PDB data.² These comparisons suggest that scientists may be overly concerned about the probability and cost of getting scooped, and 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 that — to the best of our knowledge — would not be possible in other fields of science. However, a narrow focus on one field naturally raises questions of external validity. Different academic fields have varying norms, institutions, and technology that might lead to different distributions of priority and mechanisms for assigning credit. The scoop penalty may be higher in structural biology than, 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.

In a parallel paper, we focus more broadly on the welfare implications of scientific races (Hill and Stein 2020), using the estimates from the PDB as an empirical benchmark for the returns

²We 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.

to priority. Is the observed difference in priority rewards between winners and losers too large or too small from a social welfare perspective? On one hand, explicitly rewarding priority encourages scientist effort and the timely disclosure of scientific results. On the other hand, sharp priority rewards (or the perception of a scoop penalty) may cause scientists to rush to publication at the expense of the quality of scientific research or the transparency of scientific communication. Indeed, the share of credit given to scooped articles is a salient policy lever for journal editors and funding organizations. Some journals have begun to explicitly offer a grace period where they will consider scooped papers for publication (PLOS Biology Staff Editors 2018, Marder 2017). These policies are aimed toward easing concerns about priority and reducing the risk that scientists face when embarking on competitive projects.

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 pre-suppose a winner-take-all reward, which 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; Dasgupta and David 1994; Stephan 1996). This reward structure again has implications for the pace of research and the strategic interaction of teams (Bobtcheff et al. 2016). Despite these models’ influence on our understanding of innovation systems, there is very little empirical evidence about the actual distribution of rewards in R&D 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) studies the disk drive industry in the 1970s and 1980s to test predictions about competing firms’ strategic behavior, and finds that firms lagging behind the leader are most likely to innovate. Most related to our work, Thompson and Kuhn (2017) 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 \(2013\)](#) 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 estimates also contribute to work in sociology and economics about how academic reputation interacts with future success. The Matthew Effect, first described by [Merton \(1968\)](#), is a model of path-dependent advantage, whereby success begets future success through increased name recognition, resources, and opportunities. Recent empirical work has documented evidence of the Matthew Effect in science. [Azoulay et al. \(2013\)](#) find that life scientists who win a prestigious award experience a “boost” in citations to their pre-award work relative to similar scientists. [Hill \(2019\)](#) finds that astronomers who experience exogenous bad-weather shocks during their telescope observations publish at lower rates in the future, with larger effects for novice researchers. [Jacob and Lefgren \(2011\)](#) and [Bol et al. \(2018\)](#) find that narrowly winning a post-doc grant early in the career can increase profile and accelerate productivity relative to applicants who were narrowly rejected. On the other hand, [Wang et al. \(2019\)](#) find that near-miss rejections from R01 NIH grants lead scientists to produce more impactful and creative work. They attribute this effect to “grit” or other internal motivation to overcome professional setbacks, which is also a possible response to being scooped that would counter the negative effects of being scooped in the long run. Although scientists may value attention and prestige intrinsically, journal placement and citations also translate to monetary gain in the form of grants, tenure promotions, and salary increases ([Hamermesh and Pfann 2012](#); [Ellison 2013](#)). Our estimates of the long-run consequences of getting scooped confirms that there is some amplification of citations after a successful project in our setting. The evidence we present of asymmetric credit for high- and low-reputation teams also agrees with the notion that superstar scientists may be rewarded as much for their past productivity as for their current output.

The remainder of the paper proceeds as follows. Section 2 provides some scientific background and a description of our data. Section 3 describes the empirical design and identification. Section 4 presents results for the short-run impact on publication, journal placement, citations, and alternative attention metrics 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 5 describes a model of academic attention and reports results for heterogeneity of the scoop penalty by pre-existing reputation. Section 6 benchmarks the size of our estimates against the beliefs of surveyed structural biologists about the probability and cost of getting scooped. Section 7 concludes.

2 Background and Data Construction

2.1 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 three-dimensional structure of biological macromolecules. These macromolecules include deoxyribonucleic acid (DNA), ribonucleic acids (RNA), and, most commonly, proteins. Proteins contribute to almost every process inside the body. They transport oxygen in blood (hemoglobin), trigger muscle contractions (actin and myosin), and regulate blood sugar (insulin). In many ways, the form or structure of a protein determines 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 (National Institute of General Medical Sciences 2017). Protein folding and structure has important applications, particularly in medicine, and fifteen 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. 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.³ New structures are usually solved by academic researchers at universities or research centers, although 15 percent of the scientists in our sample work at non-profit research laboratories or private companies.

2.2 The Protein Data Bank

We focus on structural biology because the Protein Data Bank (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.⁴ 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 ten percent annually (Berman et al. 2000; Burley et al. 2019). 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

³The 2013 Supreme Court ruling on the *Association for Molecular Pathology versus Myriad Genetics Inc.* case 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 3D structure of proteins were rare and difficult to obtain (Seide and Russo, 2002; Shimbo et al., 2004).

⁴The remaining types of molecules in the PDB are DNA, RNA, or a complex of protein, DNA, and/or RNA.

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, they upload the structure, experiment, and discovery details to the PDB. This typically happens shortly before or after they submit an academic paper describing their 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 one 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 is assigned a unique ID. For example, PDB ID 4HHB, deposited in 1984, is the structure of human deoxyhemoglobin, the form of hemoglobin without oxygen, which 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, etc., 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 very similar protein. The second key piece of information 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 when 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.

2.3 Identifying Priority Races: Challenges and Solutions

Identifying priority races in scientific data is difficult for three reasons. First, questions should be well-defined and have 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?), which are often published in close succession (for example, Jardim et al. 2018 and Cengiz et al. 2019). 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 the 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: 85 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 the question, methods, and conclusions. Second, as mentioned in Section 2.2, the PDB measures how biologically similar different proteins are to one another. This allows us to link projects based on objective measures of scientific proximity rather than text similarity or citation behavior. 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, having 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.⁵

2.4 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 “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 ([Dessailly et al. 2009](#); [Moult 2005](#)). In other words, the first deposit within these 50 percent similarity clusters are highly cited because they provide a novel structure model that other crystallographers can build on to solve very similar proteins. For robustness, we can restrict to scoops by proteins within the same 100 percent cluster, and find similar results which we report in Appendix Table A2.⁶ The PDB assigns ID numbers to clusters of similar proteins, and we say that the first deposit released in that cluster is the “priority” deposit. There are often many subsequent deposits that report similar structure coordinates as the priority

⁵[Thompson and Kuhn \(2017\)](#) are able to identify patent applications that were engaged in a patent race by finding patents that were rejected for lack of novelty. [Bikard \(2013\)](#) identifies paper “twins” using papers that are frequently co-cited, but this approach precludes cases where one team captured the outsized share of citations by construction, or cases where a project is abandoned.

⁶If 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 Appendix B for details on the data construction.

deposit. These follow-on deposits are either scooped projects, replication projects of the same protein by future teams, or new projects that solve the structure for closely related proteins from different organisms or bonded with different macromolecules in a novel way.⁷

We use the timing 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 outline the timeline of each project and help us determine whether two teams are working concurrently: the deposit date and release date.⁸ 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 submitted their 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 one 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, leading to potential bias if we condition on publication as a requirement for treatment assignment. Lastly, PDB releases tend to be publicly salient dates that the community pays attention to, so we are comfortable using these dates to mark priority. Appendix Section A.4 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 are scooped and fail to publish. We allow the priority project to scoop more than one team, and 5.8 percent of the races we identify have three or more competitors. Appendix Section B provides a more detailed description of the data work necessary to construct these races in practice.

An important caveat to our approach is that we can only identify races that were “close” enough that both teams had already completed a structure determination and were preparing to publish.

⁷For example, there are 30,154 clusters of proteins in the PDB that are 50 percent similar, and each cluster has an average of 7.8 deposits, only some of which are eligible to be considered racing according to our definition.

⁸The 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.

Some scientists may claim they were “scooped” if they were working on an incomplete project when another team published a solution first. We cannot observe their setback if they abandoned the project before completion, nor can we infer their counterfactual probability of success had they not been scooped. Therefore our approach specifically identifies the cost of being scooped when both teams are near the finish line. This effect may be smaller or larger than the effect of being scooped earlier in the scientific process.

2.4.1 An Example

To help understand our procedure, consider an 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 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 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 versus *Journal of Molecular Biology*, with an impact factor of 4.0) and received about 30 percent more citations over the next five years (Tripathi et al. 2013; Hiruma et al. 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)

9

2.4.2 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 exceedingly complex. Some very large proteins are composed

⁹Overall, 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 one month after the winner, fewer than 10 percent cite the winning paper. That probability increases to 60 percent for races with an eight month gap between release dates. See Appendix Figure A1.

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.¹⁰ 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. Multi-structure 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 two 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.¹¹

2.5 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.” Let $Cites_{t,t-k}^j$ be the number of citations that journal j received in year t for articles written in year $t - k$. Let $Articles_{t-k}^j$ 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:

$$JIF_t^j = \frac{Cites_{t,t-1}^j + Cites_{t,t-2}^j}{Articles_{t-1}^j + Articles_{t-2}^j}. \quad (1)$$

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

¹⁰Proteins are often composed of sub-units 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 Section B describes in more detail how we deal with multi-entity structures in the data.

¹¹The PDB only releases structures once per week, which can also make very close scoops ambiguous in terms of which truly came first. Our two week restriction helps eliminate these cases but has a minimal impact on our results. See Appendix Section A.4 for more details on the correspondence between the PDB release date and publication date.

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 five-year span. The list contains top-ranked general interest journals as well as top-ranked life science journals.¹²

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 five 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 five-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 co-author 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 A.6.

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 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. We mark an individual as having exited academia if there is a hiatus of at least five years in their publication record that begins in the five years after the priority date. Similarly, we identify individuals that exited structural biology (either changed fields or left academia) as those that have a hiatus of publishing PDB-linked papers in the following five years.

Altmetric.com Getting scooped may not only affect traditional publication outcomes like journal placement and citations, but also the overall engagement with the research by the academic

¹²Top-ten 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*.

community and general public. There have been many recent efforts to measure broader sources of academic impact by counting metrics such as news and social media engagement, patent citations, and online downloads and readership. We link the PubMed papers in our sample to data provided by Altmetric.com. In Section 4.2, we examine the effect of getting scooped in recent years on these non-traditional measures, including Mendeley downloads (a popular citation management software), news article citations, Wikipedia citations, patent citations, Twitter.com mentions, and a composite measure of attention called the Altmetric Attention Score.

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 4.4, 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 19 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 915 responses, for a total response rate of 10.2 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 C.

2.6 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 “non-racing” sample). Table 2 shows summary statistics at the structure level for both of these samples. Just over five 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.

3 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.

3.1 Baseline Specification

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

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

where Y_{ip} is an outcome, such as publication, journal impact factor, or citations. Scooped_{ip} is an indicator for losing a priority race, \mathbf{X}_{ip} is a vector of covariates, and γ_p is a protein (i.e. race) fixed effect. 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 4.3, 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 co-authored deposit i that was in a priority race over protein p , we estimate

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

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 five 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 (corresponding to the protein from the initial priority race). The long-run outcomes are calculated as the sum of each outcome in the five 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.

3.2 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 scientist 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 high 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 five years prior to the racing deposits. These include the number of PDB deposits, publications, and publications in top-ranked journals.¹³

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 authors with slightly less experience. The prior productivity of these labs is more balanced, with both the first and last authors having almost identical numbers of deposits and publications. 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 3.91.

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 model of citations using a Lasso¹⁴ regression of five-year citation counts on the battery of team covariates. This model is estimated only in the sample of non-racing deposits. We then take the selected variables and estimated coefficients to predict citations in the racing sample in a post-Lasso OLS procedure. The covariates we include are counts of publications, citations, and journal placements in the five years prior to the deposit for the first and last author, as well as the squares of these variables. We also

¹³We 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.

¹⁴Least Absolute Shrinkage and Selection Operator (Tibshirani 1996).

use the career age of the first and last authors, the rank of the first author’s institution in ten-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 Appendix 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 teams are omitted here). A perfectly balanced sample would be centered around zero and symmetric. If winners were systematically better-resourced, higher reputation, or 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 minimal, with an average difference in predicted citations of 0.21 citations (p-value of 0.587). We can also compare the distributions with a Kolmogorov-Smirnov test and calculate a test statistic of 0.040 with a p-value of 0.240. Therefore we fail to reject the hypothesis that the difference between these two histograms is different than zero. While winners and losers of priority races are not identical in observables, their differences appear to have very little systematic effect on our measures of project success.

3.3 Control Selection Using Post-double-selection 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 post-double-selection Lasso (PDS-Lasso) proposed by Belloni et al. (2014) to optimally select controls variables. Consider a partially linear model similar to equation 2

$$Y_{ip} = \alpha + \beta Scooped_{ip} + \mathbf{g}(\mathbf{Z}_{ip}) + \gamma_p + \epsilon_{ip} \quad (4)$$

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 in the balance Table 3 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.

4 Results

4.1 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 journal impact factor (standardized within year) (3) an indicator for publishing in a top-10 journal as measured by impact factor, (4) total citations accrued in five years, transformed with the inverse hyperbolic sine function¹⁵, and (5) an indicator for becoming one of the top 10 percent of publications measured by five-year citation counts. 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 five 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 3.3. 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.

Focusing on panel C, scooped projects are 2.5 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 differently, a 21 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 4.4.

In column 2, we estimate a statistically significant penalty in journal impact factor. Scooped papers are published in journals with impact factors 0.18 standard deviations below winning papers. In column 3, this translates to a 6 percentage point (18 percent) decrease in the probability of

¹⁵The inverse hyperbolic sine transform is a standard way of dealing with a right-skewed distribution that has zeroes and/or negative numbers (Burbidge et al. 1988; Bellemare and Wichman 2019). The transformation is given by $asinh(x) = \log(x + \sqrt{x^2 + 1})$. The coefficients on variables transformed by the hyperbolic sine function can be interpreted similarly to logs (i.e. proportionally).

publishing in a top-ten 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 five years. The scooped projects receive 20 percent fewer citations in the same time span. Column 5 suggests that this means scooped projects are 3.5 percentage points (23 percent) less likely to be one of the top 10 percent of papers in that publication year ranked by five-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. As a further robustness check, we reproduce this table using a sub-sample of races that have projects with 100 percent similar sequence structure according to the algorithm used by the PDB. Appendix Table A2 shows that the magnitudes are very similar for all outcomes, even if statistical precision is lower due to the smaller sample size.

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 44.5 percent of the total citations accrued to both papers, a much larger share of credit than zero percent for the winner as is typically assumed by classic models of innovation races.¹⁶

4.2 Alternative Measures of Attention

Scooped projects may not only be penalized 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 5 shows results of project-level regressions using outcomes sourced from Altmetric.com. In these regressions, we restrict the sample period to 2011-2017 since many of these outcomes are only relevant in the recent internet era. All outcomes are count variables again transformed with the inverse hyperbolic sine function to deal with skewness and facilitate proportional interpretation of the effects. Regression results are again reported with the three different control strategies used in Table 4.

Column 1 of Table 5 reports the effect of getting scooped on Mendeley readership. Mendeley is a popular citation manager used by many researchers. Downloading a paper on Mendeley can be interpreted as a proxy for popularity of a paper among readers, and especially those readers that might consider citing the paper at some point. Focusing on panel C, getting scooped leads to an approximately 45 percent decline in Mendeley downloads, which is quite a bit larger than the citation penalty reported in Table 4. News stories covering the academic articles fall by 11 percent for scooped papers, and Wikipedia citations fall by 3.5 percent. There is no detectable effect on patent

¹⁶The estimated share of 44.5 percent is calculated by dividing the mean citations of the losing teams, $28.9 * (1 - 0.197)$ by the implied total citations ($28.9 + 28.9 * (1 - .197)$) based on the estimate of the percent citation penalty from column 4, panel C.

citations. Mentions of a paper on Twitter fall by 10 percent, although this estimate is only marginally significant and not robust to all control strategies. Altmetric.com provides a comprehensive score of alternative attention (Huang et al. 2018), which falls by 24 percent for scooped papers. These results suggest that getting scooped has different effects for different audiences. The large effect on readership proxied by Mendeley suggests that scientists who casually interact with the research are more prone to focus on only the race winners. This is likely driven in part by journal placement, where some scientists stay abreast of advances in various fields by only reading papers that appear in the top general interest or field journals. Science reporters in the news tend to be less responsive to priority ordering, suggesting that they might be more likely to cover both papers about a topic instead of just the first paper. Some of the most specialized readers, such as Wikipedia contributors and patent citers seem to be the least responsive, suggesting that they do a much deeper literature search when citing academic papers.

4.3 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 6 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, which are defined as authors who had seven years or less since their first publication at the time of the scooping event.¹⁷ Panel C restricts to veterans, which are all scientists not defined as novices.¹⁸

Getting scooped has no statistically-significant effect on the probability of remaining in academia in the five years after the race. Column 1 shows that both novices and veteran scientists that get scooped are not more likely to stop publishing after the race. However, in column 2 we do find evidence that both novices and veterans are less likely to still be actively publishing PDB-linked articles after being scooped. In the full sample, 64 percent of authors remain active in structural biology for at least five years following the priority date, and the scooped scientists are 2.9 percentage points less likely to persist than the winning scientists. The negative effect is twice as large in percentage point terms for novices (5.5 percentage points) than for veterans (2.8 percentage points), suggesting that novices might have a more malleable research agenda. Getting scooped appears to not be enough of an obstacle to derail academic careers, but it might cause enough discouragement to redirect researchers toward different areas of study.

We find no significant changes to publishing 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 column 3 and 4, and they are not more or less likely to publish in top-10 journals. However, we do estimate significant penalties in citations for all categories of authors. In the full author sample, the scooped individuals receive 17 percent fewer citations

¹⁷Seven years is the 30th percentile of the distribution of years since first publication.

¹⁸The 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 requires us to restrict to races that have at least one novice (or veteran) in the winning and losing team of each race.

(measured by inverse hyperbolic sine citation-weighted publications) in the next five years, where citations are counted up to three years after each paper’s publication. This effect falls particularly hard on novices, who receive 32 percent fewer citations, while veterans receive only 13 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 0.42 fewer hit papers in the five years following a scoop event. The negative effect is lower for novices in levels (0.10 papers versus 0.58 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 (an eight percent decline versus a six percent decline). We also consider outcomes in the following three years in Appendix Table A3 and ten years in Appendix Table A4. The results are similar in the three 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 ten years of our sample window.

4.4 Mechanisms: Role of Scoop Timing in the Publication Process

Scooped projects receive 20 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 unless they were 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.

4.4.1 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 7 replicates the citation penalty effect from Table 4, column 4, but uses a subsample of races where 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 journal impact factor, first as a linear term and then as a cubic polynomial. The citation effect falls to 11 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 five percent. These results suggest that at least two thirds 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.

4.4.2 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: “[R]ecommend 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.”¹⁹ 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 instead use the timing of submission and acceptance to present suggestive evidence that editors at top journals are reticent to publish scooped papers.

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 6 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, with a one standard deviation higher ranked journal being eight 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. Some of these lower-ranked journals, such as *PLOS Biology*, have 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

¹⁹See 2019 *Science* Instructions for Reviewers of Research Articles: <https://www.sciencemag.org/sites/default/files/RAinstr19.pdf>

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.

4.4.3 Time Lag and the Scoop Penalty

The severity of the scoop penalty is correlated with the time lag between when the winning and losing projects are released. In Figure 5, we plot the difference in outcomes separately for three 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 journal impact factor 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 journal impact factor penalty is 0.1 standard deviations in the first three to four months, then drops to 0.3 standard deviations by eight months. Similarly, projects released within one month of each other have no difference in citations. The scoop penalty grows to 50 percent for scooped projects with an eight 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. Teams likely make strategic decisions to rush to publish, revise and delay, or abandon altogether, so the delay times should be viewed as potentially selected on team or project characteristics. These results suggest, however, that the delay time between projects is relevant for editors and readers, perhaps because the community can more clearly attribute priority credit with more time separating similar projects.

5 Reputation and the Scoop Penalty

In this section we show that academic recognition is affected not only by priority, but also by the preexisting reputation of winners and losers. When a high-status team scoops a low-status team, they receive 66 percent of the total citations, but when a low-status team scoops a high-status team in a comparable race, they only receive 46 percent of the the total citations. This asymmetry in attention suggests that the distribution of priority rewards is not formulaic and may be affected by the institutions and norms of the academic community. We propose a model of academic attention based on a standard statistical discrimination model (Aigner and Cain, 1977) and 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.

5.1 A Model of Academic Attention

5.1.1 Setup

Editors, reviewers, and authors read new academic papers. In doing so, they receive a noisy signal of the paper’s quality. The notion that paper quality is only partially observed by readers is similar to the setup in [Card and DellaVigna \(2019\)](#) and may arise from inattention or uncertainty about the importance of the contribution. The signal, s , is a function of the paper’s true underlying quality (q) as well as a noise term, u :

$$s = q + u$$

where $u \sim N(0, \sigma_u^2)$ is independent of $q \sim N(\alpha, \sigma_q^2)$. Following the standard statistical discrimination model, readers will use both the signal and the average quality to infer the paper’s quality:

$$\hat{q}(s) = E[q|s] = \lambda s + (1 - \lambda)\alpha$$

where $\lambda = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_u^2}$ is the signal-to-noise ratio. Intuitively, expected quality is a weighted average of the observed signal and mean quality. Readers put more weight on the signal when λ is large, i.e. when the signal is informative relative to the noise term.

5.1.2 The Priority Premium

When making decisions about which paper to publish or cite, scientists care about both quality and priority. Consider two papers which answer the same question, with inferred qualities \hat{q}_1 and \hat{q}_2 . Let the numeric subscript index the order of publication, so that \hat{q}_1 was published before \hat{q}_2 , and let $f > 0$ denote the priority premium. A scientist will cite the first paper if $\hat{q}_1 + f \geq \hat{q}_2$. On the other hand, a scientist will cite the second paper if $\hat{q}_1 + f < \hat{q}_2$.

5.1.3 Lab Types

Suppose there are two types of labs, H and L . H labs are “high-reputation” labs, known for producing papers of high average quality, while L labs are “low-reputation” labs, known for producing papers of low average quality. In other words, q is drawn from a different distribution depending on the lab type. For H labs, $q^H \sim N(\alpha^H, \sigma_q^2)$ while for L labs, $q^L \sim N(\alpha^L, \sigma_q^2)$. The key distinction between the two lab types is that $\alpha^H > \alpha^L$. We will assume that variances are equal.

When two labs each write a paper on the identical topic (or in our case, protein), the true qualities of the two papers are the same. However, if the labs have different reputations, the

inferred qualities will be different, even if the signals are identical:

$$\begin{aligned}\hat{q}^H(s) &= \lambda s + (1 - \lambda)\alpha^H \\ \hat{q}^L(s) &= \lambda s + (1 - \lambda)\alpha^L.\end{aligned}$$

Ultimately, this gives rise to two distinct effects when competing labs publish on the same protein. The “priority effect” leads scientists to cite the earlier paper, since this paper receives a premium, as described above. On the other hand, the “reputation effect” leads scientists to cite the paper from the higher-reputation lab, since this paper will have higher inferred quality. This insight leads us to two propositions.

Proposition 1. *If labs are the same type, then the lab that publishes first is more likely to be cited. In other words,*

$$P(\hat{q}_1^H + f \geq \hat{q}_2^H) = P(\hat{q}_1^L + f \geq \hat{q}_2^L) > \frac{1}{2}.$$

Proof. See Appendix D. The intuition is that if the labs are the same type, there is no differential reputation effect. Therefore, citations are driven solely by the priority effect.

Proposition 2. *If the lab that publishes first is H-type and the lab that publishes second is L-type, then the lab that publishes first is more likely to be cited. Moreover, the difference in citations will be greater than if the labs were the same type. Conversely, if the lab that publishes first is L-type and the lab that publishes second is H-type, it is ambiguous which lab is more likely to be cited. However, the difference in probability of citation will certainly be less than if the labs were the same type. This means that we can rank the probability of citation in all four scenarios:*

$$P(\hat{q}_1^H + f \geq \hat{q}_2^L) > P(\hat{q}_1^H + f \geq \hat{q}_2^H) = P(\hat{q}_1^L + f \geq \hat{q}_2^L) > P(\hat{q}_1^L + f \geq \hat{q}_2^H).$$

Proof. See Appendix D. The intuition is that if the first lab is H-type and the second lab is L-type, then the priority effect and the reputation effect work in the same direction. However, if the first lab is L-type and the second lab is H-type, then the priority effect and the reputation effect are working in opposite directions. Therefore, the net effect on citation behavior is ambiguous.

5.2 Priority and Academic Reputation

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 $CitationShare_A = Citations_A / (Citations_A + Citations_B)$. This citation share maps to the probability of citation outlined in the model above.²⁰

We proxy for the pre-existing “reputation” of each lab using the Lasso-estimated predicted citations from the non-racing data sample as described in Section 3.1.1. Labs with above-median

²⁰The model does not include the possibility of co-citations, where both papers are cited together, but the empirical results are proportional to an analysis where co-citations are excluded.

predicted citations correspond to the H labs, while teams below median correspond to the L labs. In Figure 7 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 pre-existing 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 7 conveniently partition the sample into four sub-samples 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 H teams produce higher quality or more convincing results, then the additional citations they receive may not only be caused by their high-profile reputation. 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 3.2. Appendix Table A5 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 H and L teams 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 8 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 H versus H matchups and the L versus L matchups (winning team receives 55 percent of the total citations). This is consistent with the prediction from proposition 1.²¹

Panel B shows the unevenly matched races. When an H lab scoops an L 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 (66 percent). Conversely, when an L lab scoops an H 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

²¹The 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.

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.

6 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 55 percent of the total citations, a far cry from 100 percent as is 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 C. 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 and compares them to the magnitudes estimated in the PDB data. 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,²² 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 double the 8 percent scoop probability in the comparable PDB sample. Six percent of respondents report that they would abandon the project, but only 70 percent think they would succeed at publishing conditional on submitting, suggesting a 66 percent unconditional probability of publishing. This is much lower than the 86 percent of scooped papers that are actually published in the PDB data, and the 97 percent that are published in the comparable subsample. Scientists are very pessimistic about the potential journal placement of scooped papers, expecting that the

²²Specifically, we sort races by the time elapsed between the loser deposit date and the winner release date and keep the quarter of race losers that were scooped earliest in the process.

best journal they could publish in would be almost three 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 20 percent penalty or a 45 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 5. 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 seven 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 to get about five more citations (nine percent) than low-reputation scientists. 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 is driven by statistical discrimination of editors, reviewers, and readers as explained in the model in Section 5. 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.

7 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 linking adjacent projects using biological measures of similarity, we reconstruct races and compare the outcomes of winners and losers, even in cases where the losing project goes unpublished. We find that losing a priority race decreases the probability of publishing by 2.5 percentage points. Conditional on publishing, the scooped papers are less likely to appear in a top journal and receive 20 percent fewer citations than the winning papers. The effect of getting scooped lingers along some

dimensions in the years following the event. We find no effect on exiting academia, but a small increase in the probability of exiting the field of structural biology. We also observe that citations decrease for scooped scientists in subsequent work, particularly for novices. Priority rewards are in part dependent on pre-existing reputation. In cases where a high-reputation team is racing against a low-reputation team, priority rewards are unevenly distributed. High-reputation winners receive much more attention than losers. And in cases where the high-reputation team is scooped, the winning low-reputation team receives no more citations than their high-reputation rival.

Given the moderate estimated cost of losing a race, especially in the long run, are scientists overly concerned about the threat of being scooped? There has been scant evidence on scientist beliefs about the threat of being preempted. The best evidence we can find comes from a survey conducted by Hagstrom (1974) who finds that 29 percent of experimental biologists are moderately or very concerned that they will be scooped on their current research. We update these survey results in the field of structural biology, and find that scientists may be overly concerned about getting scooped. In the survey we conduct, scientists perceive a higher likelihood of being scooped than we see in the PDB data, and conditional on being scooped, they believe the penalty in terms of publication and citations is higher than we estimate.

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, or consider alternative systems or policies. An important concern raised in popular and academic writing is the potential “dark side” of priority, where novelty may be pursued at the expense of openness and 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 (2020), 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. Future work should also focus on how competition affects the openness of science, ease of collaboration, and free transmission of knowledge between scientists. 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 in a few months after being scooped, providing opportunities to establish priority for early-stage work through pre-prints, 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.²³

²³Ryan Hill: Northwestern Kellogg School of Management, and Carolyn Stein: MIT Economics Department

References

- Aigner, Dennis J. and Glen G. Cain**, “Statistical Theories of Discrimination in Labor Markets,” *ILR Review*, 1977, *30* (2), 175–187.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman**, “Basic Local Alignment Search Tool,” *Journal of Molecular Biology*, 1990, *215* (3), 403–410.
- Arrow, Kenneth J.**, “Economic Welfare and the Allocation of Resources for Invention,” in “The Rate and Direction of Inventive Activity: Economic and Social Factors,” Princeton University Press, 1962.
- Azoulay, Pierre, Toby Stuart, and Yanbo Wang**, “Matthew: Effect or Fable?,” *Management Science*, 2013, *60* (1), 92–109.
- Barinaga, Marcia**, “The Missing Crystallography Data,” *Science*, 1989, *245* (4923), 1179.
- Bellemare, Marc F. and Casey J. Wichman**, “Elasticities and the Inverse Hyperbolic Sine Transformation,” *Oxford Bulletin of Economics and Statistics*, 2019.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, “Inference on Treatment Effects After Selection Among High-Dimensional Controls,” *The Review of Economic Studies*, 2014, *81* (2), 608–650.
- Berman, Helen, Kim Henrick, Haruki Nakamura, and John L Markley**, “The Worldwide Protein Data Bank (wwPDB): Ensuring a Single, Uniform Archive of PDB Data,” *Nucleic Acids Research*, 2006, *35*, D301–D303.
- Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T.N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne**, “The Protein Data Bank,” *Nucleic Acids Research*, January 2000, *28* (1), 235–242.
- , **Stephen K. Burley, Gerald J. Kleywegt, John L. Markley, Haruki Nakamura, and Sameer Velankar**, “The Archiving and Dissemination of Biological Structure Data,” *Current Opinion on Structural Biology*, 2016, *40*, 17–22.
- Bikard, Michaël**, “Simultaneous Discoveries as a Research Tool: Method and Promise,” *SSRN Working Paper*, 2013.
- Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti**, “Researcher’s Dilemma,” *The Review of Economic Studies*, 2016, *84* (3), 969–1014.
- Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt**, “The Matthew effect in science funding,” *Proceedings of the National Academy of Sciences*, 2018, *115* (19), 4887–4890.

- Brown, Eric N. and S. Ramaswamy**, “Quality of Protein Crystal Structures,” *Acta Crystallographica Section D*, 2007, *63*, 941–950.
- Burbidge, John B., Lonnie Magee, and A. Leslie Robb**, “Alternative Transformations to Handle Extreme Values of the Dependent Variable,” *Journal of the American Statistical Association*, 1988, *83* (401), 123–127.
- Burley, Stephen K., Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M. Duarte, Shuchismita Dutta et al.**, “RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy,” *Nucleic Acids Research*, January 2019, *47* (D1), D464–D474.
- Campbell, Philip**, “New Policy for Structural Data,” *Nature*, July 1998, *394* (6689), 105.
- Card, David and Stefano DellaVigna**, “What Do Editors Maximize? Evidence from Four Economics Journals,” *The Review of Economics and Statistics*, 2019, *forthcoming*.
- Cengiz, Doruk, Arindrajit Dube, Atilla Lindner, and Ben Zipperer**, “The Effect of Minimum Wages on the Total Number of Jobs: Evidence from the United States Using a Bunching Estimator,” Working Paper 25434, National Bureau of Economic Research 2019.
- Dasgupta, Partha and Joseph Stiglitz**, “Uncertainty, Industrial Structure, and the Speed of R&D,” *The Bell Journal of Economics*, Spring 1980, *11* (1), 1–28.
- **and Paul A. David**, “Toward a New Economics of Science,” *Research Policy*, 1994, *23*, 487–521.
- Dessailly, Benoît H, Rajesh Nair, Lukasz Jaroszewski, J Eduardo Fajardo, Andrei Kouranov, David Lee, Andras Fiser, Adam Godzik, Burkhard Rost, and Christine Orengo**, “PSI-2: structural genomics to cover protein domain family space,” *Structure*, 2009, *17* (6), 869–881.
- Ellison, Glenn**, “How Does the Market Use Citation Data? The Hirsch Index in Economics,” *American Economic Journal: Applied Economics*, July 2013, *5* (3), 63–90.
- Fermi, Giulio, Max F. Perutz, Boaz Shaanan, and Roger Fourme**, “The Crystal Structure of Human Deoxyhaemoglobin at 1.74 Å resolution,” *Journal of Molecular Biology*, May 1984, *175* (2), 159–174.
- Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole**, “Preemption, Leapfrogging and Competition in Patent Races,” *European Economic Review*, 1983, *22* (1), 3–31.
- Goodsell, David S.**, “Methods for Determining Atomic Structures,” Technical Report, Protein Data Bank: PDB-101 2019.

- Hagstrom, Warren O.**, “Competition in Science,” *American Sociological Review*, February 1974, 39 (1), 1–18.
- Hamermesh, Daniel and Gerard Pfann**, “Reputation and Earnings: The Roles of Quality and Quantity in Academe,” *Economic Inquiry*, January 2012, 50 (1), 1–16.
- Hill, Ryan**, “Searching for Superstars: Research Risk and Talent Discovery in Astronomy,” *Working Paper*, 2019.
- **and Carolyn Stein**, “Race to the Bottom: Competition and Quality in Science,” *Working Paper*, 2020.
- Hiruma, Yoshitaka, Mathias AS Hass, Yuki Kikui, Wei-Min Liu, Betül Ölmez, Simon P Skinner, Anneloes Blok, Alexander Kloosterman, Hiroyasu Koteishi, Frank Löhr et al.**, “The structure of the cytochrome P450cam–putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography,” *Journal of molecular biology*, 2013, 425 (22), 4353–4365.
- Huang, Wenya, Peiling Wang, and Qiang Wu**, “A Correlation Comparison Between Altmetric Attention Scores and Citations for Six PLOS Journals,” *PloS One*, 2018, 13 (4).
- Jacob, Brian and Lars Lefgren**, “The Impact of NIH Postdoctoral Training Grants on Scientific Productivity,” *Research Policy*, 2011, 40 (6), 864–874.
- Jardim, Ekaterina, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wething**, “Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle,” Working Paper 23532, National Bureau of Economic Research 2018.
- Lee, Tom and Louis L. Wilde**, “Market Structure and Innovation: A Reformulation,” *Quarterly Journal of Economics*, March 1980, 94 (2), 429–436.
- Lerner, Josh**, “An Empirical Exploration of a Technology Race,” *RAND Journal of Economics*, Summer 1997, 28 (2), 228–247.
- Loury, Glenn C.**, “Market Structure and Innovation,” *Quarterly Journal of Economics*, August 1979, 93 (3), 395–410.
- Marder, Eve**, “Scientific Publishing: Beyond scoops to best practices,” *Elife*, 2017, 6, e30076.
- Martz, Eric, Wayne Decatur, Joel L. Sussman, Michal Harel, and Eran Hodis**, “Nobel Prizes for 3D Molecular Structure,” February 2019.
- Merton, Robert K.**, “Priorities in Scientific Discovery: A Chapter in the Sociology of Science,” *American Sociological Review*, December 1957, 22 (6), 635–659.
- , “The Matthew Effect in Science,” *Science*, 1968, 159 (3810), 56–63.

- Milojević, Staša**, “Accuracy of simple, Initials-Based Methods for Author Name Disambiguation,” *Journal of Informetrics*, 2013, 7 (4), 767–773.
- Moult, John**, “A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction,” *Current opinion in structural biology*, 2005, 15 (3), 285–289.
- National Institute of General Medical Sciences**, “Structural Biology,” Technical Report October 2017.
- Nelson, Richard R.**, “The Simple Economics of Basic Scientific Research,” *Journal of Political Economy*, June 1959, 67 (3), 297–306.
- Phelps, Edmund S.**, “The Statistical Theory of Racism and Sexism,” *American Economic Review*, 1972, 62 (4), 659–661.
- PLOS Biology Staff Editors**, “The Importance of Being Second,” *PLOS Biology*, 2018, 16 (1).
- Ramakrishnan, Venki**, *Gene Machine: The Race to Decipher the Secrets of the Ribosome*, Basic Books, 2018.
- Seide, Rochelle K. and Alicia A. Russo**, “Patenting 3D Protein Structures,” *Expert Opinion on Therapeutic Patents*, 2002, 12 (2), 147–150.
- Shimbo, Itsuki, Rie Nakajima, Shigeyuki Yokoyama, and Koichi Sumikura**, “Patent Protection for Protein Structure Analysis,” *Nature Biotechnology*, 2004, 22 (1), 109–112.
- Stephan, Paula E.**, “The Economics of Science,” *Journal of Economic Literature*, 1996, 34 (3), 1199–1235.
- Sussman, Joel L.**, “What’s New at the PDB,” *Quarterly Newsletter published by Brookhaven National Laboratory Protein Data Bank*, April 1998, 84, 1.
- Thompson, Neil C. and Jeffrey M. Kuhn**, “Does Winning a Patent Race Lead to More Follow-on Innovation?,” *SSRN Working Paper*, January 2017.
- Tibshirani, Robert**, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58 (1), 267–288.
- Torvik, Vetle I. and Neil R. Smalheiser**, “Author name disambiguation in MEDLINE,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2009, 3 (3), 11.
- , **Marc Weeber, Don R. Swanson, and Neil R. Smalheiser**, “A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation,” *Journal of the American Society for Information Science and Technology*, 2005, 56 (2), 140–158.
- Tripathi, Sarvind, Huiying Li, and Thomas L Poulos**, “Structural basis for effector control and redox partner recognition in cytochrome P450,” *Science*, 2013, 340 (6137), 1227–1230.

Wang, Yang, Benjamin F Jones, and Dashun Wang, “Early-career setback and future career impact,” *Nature communications*, 2019, *10* (1), 1–10.

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

wwPDB, “wwPDB Policies and Processing Procedures Document, Release of PDB Entries,” 2019.

Yong, Ed, “In Science, There Should Be a Prize for Second Place,” *The Atlantic*, February 2018.

Tables and Figures

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 date	September 14, 2012	February 3, 2012
Deposit date	March 27, 2013	April 3, 2013
Release date	June 19, 2013	August 21, 2013
First author affiliation	University of California, Irvine	Leiden University
Journal	<i>Science</i>	<i>Journal of Molecular Biology</i>
Journal impact factor	31.5	4
Five Year Citations:	52	39

Notes: This table presents an example of a racing pair identified in the Protein Data Bank using the scoop rules outlined in Section 2.4. See Figure 3 for the image of the structure models deposited by each team.

Table 2: Summary Statistics for Structure-Level Data

Variable	Racing (1)	Not racing (2)	Difference (race - not race) (3)	Std. error of difference (4)
<i>Panel A. Team characteristics</i>				
Number of authors	7.134	7.454	-0.319	(0.078) ***
Affiliation in North America	0.292	0.351	-0.058	(0.008) ***
Affiliation in Europe	0.151	0.158	-0.007	(0.006)
Affiliation in Asia	0.190	0.133	0.057	(0.007) ***
Top 50 university	0.251	0.241	0.010	(0.008)
Rank 51-200 university	0.238	0.260	-0.022	(0.008) ***
Other affiliation	0.511	0.499	0.013	(0.009)
Industry or non-profit affiliation	0.154	0.170	-0.016	(0.006) **
First author experience (years)	5.462	5.986	-0.524	(0.109) ***
Last author experience (years)	7.410	7.813	-0.403	(0.119) ***
<i>Panel B. Project outcomes</i>				
Published	0.867	0.752	0.115	(0.006) ***
Standardized impact factor	0.114	-0.045	0.158	(0.021) ***
Top ten journal	0.354	0.281	0.073	(0.009) ***
Five-year citation counts	26.370	17.245	9.125	(0.739) ***
Top 10% in five-year citations	0.132	0.132	0.000	(0.000) ***
<i>Panel C. Project altmetrics</i>				
Mendeley downloads	33.838	24.032	9.806	(1.400) ***
News stories	0.300	0.214	0.086	(0.059)
Wikipedia citations	0.178	0.091	0.088	(0.009) ***
Patent citations	0.906	0.661	0.246	(0.089) ***
Twitter mentions	1.855	1.691	0.165	(0.196)
Altmetric attention score	5.262	3.875	1.387	(0.621) **
Observations	3,319	64,018		

Notes: This table presents summary statistics for the racing and non-racing samples. Observations are at the structure level. Column 1 shows the means of the racing sample and column 2 shows the means of the non-racing sample. Column 3 shows the difference between the racing and non-racing projects, and column 4 shows the heteroskedasticity-robust standard error of the difference.

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

Table 3: Covariate Balance Between Winning and Losing Teams

Variable	Not racing (1)	Racing: losers (2)	Racing: winners (3)	Difference: (lose - win) (4)	Std. error of difference (5)
<i>Panel A. Team characteristics</i>					
Number of authors	7.454	7.193	7.074	0.119	(0.204)
Affiliation in North American	0.351	0.264	0.321	-0.057	(0.022) ***
Affiliation in Europe	0.158	0.133	0.170	-0.038	(0.018) **
Affiliation in Asia	0.133	0.223	0.155	0.068	(0.018) ***
Top 50 university	0.241	0.222	0.280	-0.058	(0.020) ***
Rank 51-200 university	0.260	0.247	0.228	0.019	(0.020)
Other affiliation	0.499	0.531	0.491	0.039	(0.023) *
Industry or non-profit affiliation	0.170	0.156	0.152	0.004	(0.018)
First author experience (years)	5.986	5.785	5.127	0.658	(0.278) **
Last author experience (years)	7.813	7.510	7.306	0.203	(0.311)
<i>Panel B. First author productivity (prior five years)</i>					
Deposits	12.362	4.168	5.473	-1.304	(0.734) *
Publications	2.893	2.677	3.138	-0.461	(0.464)
Top-10 publications	0.649	0.706	0.666	0.040	(0.064)
Top-5 publications	0.222	0.265	0.242	0.023	(0.032)
<i>Panel C. Last author productivity (prior five years)</i>					
Deposits	44.284	30.772	28.922	1.850	(4.288)
Publications	9.909	12.423	13.511	-1.088	(2.233)
Top-10 publications	4.007	4.617	4.569	0.048	(0.505)
Top-5 publications	1.419	1.638	1.784	-0.146	(0.188)
<i>Panel D. Project quality metrics</i>					
Resolution (Å)	2.244	2.328	2.317	0.011	(0.062)
R-free goodness-of-fit	0.236	0.245	0.243	0.002	(0.002)
Observations	64,018	1,689	1,630	<i>F</i> -stat:	3.911 ***

Notes: 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 non-racing sample, column 2 shows the means of the losing projects in the racing sample, and column 3 shows the means of the winning projects in the racing sample. Column 4 shows the difference between the losing and winning projects, and column 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* < 0.1, *p* < 0.05, ****p* < 0.01.







Table 4: Effect of Getting Scooped on Project Outcomes

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
<i>Panel A. No controls</i>					
Scooped	-0.027* (0.015)	-0.187*** (0.044)	-0.065*** (0.020)	-0.243*** (0.070)	-0.037** (0.014)
<i>Panel B. Base controls</i>					
Scooped	-0.026** (0.013)	-0.176*** (0.044)	-0.062*** (0.020)	-0.208*** (0.063)	-0.028** (0.014)
<i>Panel C. PDS-Lasso selected controls</i>					
Scooped	-0.025*** (0.010)	-0.178*** (0.032)	-0.060*** (0.014)	-0.197*** (0.045)	-0.035*** (0.010)
Winner Y mean	0.880	-0.031	0.318	28.918	0.150
Observations	3,319	3,319	3,319	2,546	2,546

Notes: This table presents regression estimates of the scoop penalty, following equation 1 in the text. 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 as 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. Column 4 regression uses $\text{asinh}(\text{five-year citations})$ as the dependent variable, but Winner Y Mean is reported in levels for ease of interpretation.

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

Table 5: Effect of Getting Scooped on Alternative Measures of Attention

Dependent variable: All transformed with asinh()	 Mendeley downloads (1)	 News stories (2)	 Wikipedia citations (3)	 Patent citations (4)	 Twitter mentions (5)	 Altmetric attention (6)
<i>Panel A. No controls</i>						
Scooped	-0.452*** (0.152)	-0.107** (0.042)	-0.037** (0.018)	-0.007 (0.028)	-0.114 (0.077)	-0.240** (0.094)
<i>Panel B. Base controls</i>						
Scooped	-0.425*** (0.144)	-0.092** (0.043)	-0.030 (0.020)	0.001 (0.031)	-0.087 (0.074)	-0.199** (0.090)
<i>Panel C. PDS-Lasso selected controls</i>						
Scooped	-0.453*** (0.105)	-0.108*** (0.032)	-0.035** (0.014)	-0.008 (0.021)	-0.101* (0.054)	-0.237*** (0.066)
Winner Y mean	42.874	0.641	0.104	0.260	3.982	9.137
Observations	1,339	1,339	1,339	1,339	1,339	1,339

Notes: Attention outcomes are sourced from Altmetric.com. Sample restricted to years 2011-2017. 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 as 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. All outcomes are cumulative counts of the metrics summed over time between the publication date to August 2019. All counts are transformed with the inverse hyperbolic sine transformation. The Altmetric Attention Score is a composite measure of all metrics used by Altmetric.com.

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

Table 6: Effect of Getting Scooped on Five-Year Productivity

Dependent variable	Active in PubMed 5 years later (1)	Active in PDB 5 years later (2)	Total count five years after race				
			PubMed Publications (3)	PDB Publications (4)	Top-ten publications (5)	Citation-weighted publications (6)	Top-10% cited publications (7)
<i>Panel A. All scientists</i>							
Scooped	-0.010 (0.008)	-0.029** (0.015)	-1.122 (1.039)	-0.079 (0.218)	-0.116 (0.100)	-0.171*** (0.044)	-0.415** (0.179)
Winner Y mean	0.834	0.639	45.750	7.123	3.603	497.310	7.749
Observations	4,648	4,648	8,700	8,700	8,700	6,531	6,531
<i>Panel B. Novices</i>							
Scooped	-0.030 (0.024)	-0.055** (0.025)	-0.017 (0.273)	0.006 (0.167)	0.108 (0.067)	-0.317*** (0.103)	-0.097 (0.109)
Winner Y mean	0.464	0.332	4.228	1.882	0.614	75.359	1.162
Observations	1,097	1,097	2,049	2,049	2,049	1,539	1,539
<i>Panel C. Veterans</i>							
Scooped	-0.008 (0.005)	-0.028* (0.017)	-1.219 (1.544)	-0.176 (0.304)	-0.202 (0.143)	-0.131*** (0.042)	-0.584** (0.250)
Winner Y mean	0.981	0.763	61.490	9.216	4.775	667.393	10.396
Observations	3,142	3,142	5,870	5,870	5,870	4,411	4,411

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than eight years of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

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

Table 7: Decomposing Citation and Journal Effect

Dependent variable	Five-year citations			
	(1)	(2)	(3)	(4)
Scooped	-0.164*** (0.032)	-0.114*** (0.028)	-0.107*** (0.028)	-0.047* (0.026)
Journal controls	None	Linear JIF	Cubic JIF	Journal FE
Winner Y mean	34.8	34.8	34.8	34.8
Observations	1,917	1,917	1,917	1,917

Notes: This table reports the scooped coefficients in regressions with five-year citations as the outcome where we control for journal impact factor. 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 re-estimates the Table 1, column 4 regression in this subsample. Column 2 and 3 add linear and then cubic controls for journal impact factor. Column 4 includes fixed effects for journal. All regressions also include PDS-Lasso selected controls and protein fixed effects.

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

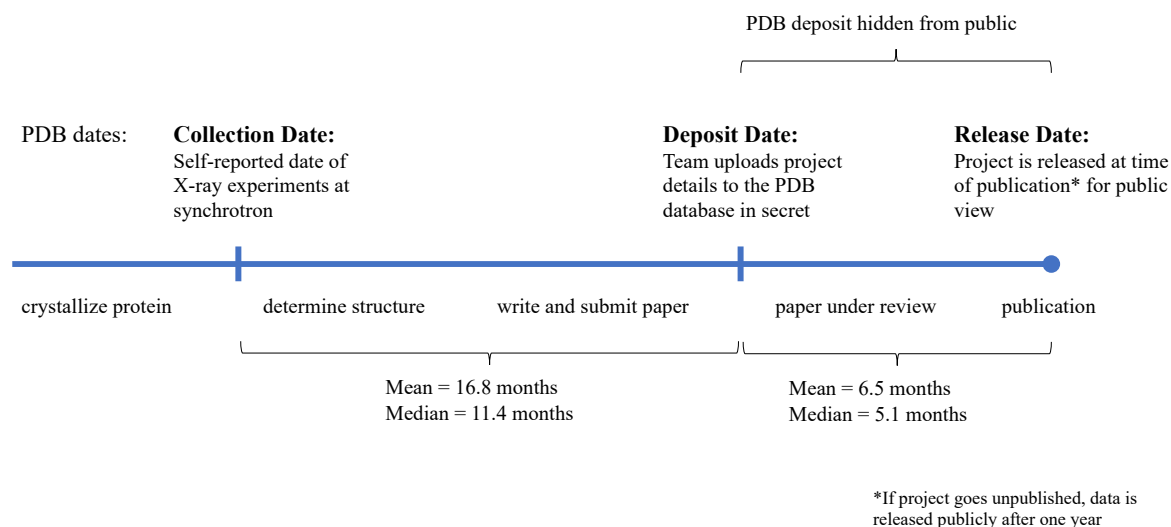
Table 8: Survey Benchmark of Scoop Penalty

	PDB estimate		Survey estimate			
	Full sample (1)	Comparable subsample (2)	All respondents (3)	Below-median reputation (4)	Above-median reputation (5)	Column (4) - (5) difference (6)
<i>Prob</i> (Scoop)	0.029	0.081	0.266	0.268	0.264	0.004 (0.016)
<i>Prob</i> (Publication)	0.853	0.976	0.665	0.628	0.703	-0.075*** (0.022)
Journal impact factor penalty	-0.18	-1.23	-2.92	-2.95	-2.89	-0.055 (0.084)
Citation penalty	-0.197	-0.150	-0.594	-0.620	-0.568	-0.052** (0.024)
Scooped citation share	0.445	0.459	0.257	0.241	0.274	-0.033*** (0.011)

Notes: This table reports the responses to a survey of 915 structural biologists. The survey asked respondents to estimate the probability and consequences of getting scooped on a hypothetical project. See Appendix C for full survey text. Estimates from the PDB main regressions are reported in column 1. Comparable subsample PDB estimates in column 2 restrict to PDB races where one racer published in *Science*, *Nature*, or *Cell*, and losing team was scooped early in the process (quarter of sample with the shortest time between loser deposit and winner release). In column 4 and 5, respondents were divided into two groups, high- and low-reputation using the predicted citations measure used for heterogeneity in Section 6 of the text. Column 6 reports the difference in response means between columns 4 and 5 and reports the heteroskedastic-robust standard error in parentheses.

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

Figure 1: Project Timeline and Key Dates

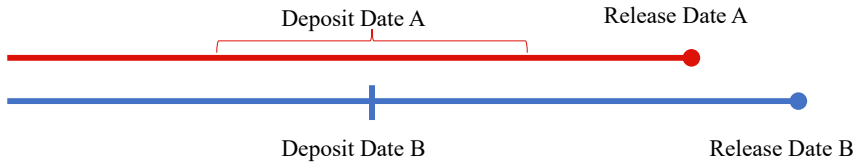


Notes: This figure shows the timeline of a typical PDB project. Dates in bold above the line are observed in our data. Events listed below the timeline are the approximate timing of other project events including the submission and review process. Deposit event and structure data is hidden from public until the structure is released.

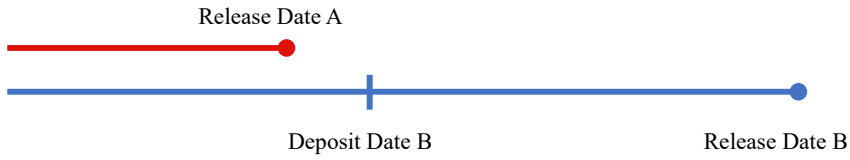
Figure 2: Defining Priority Races

- Rules:**
1. Take two projects that have identical sequence and different authors.
 2. Assert that both projects are deposited before the first project is released.
 3. Call the first to release the winner, call the second project “scooped.”

Scenario 1: **Project A** scoops **Project B**

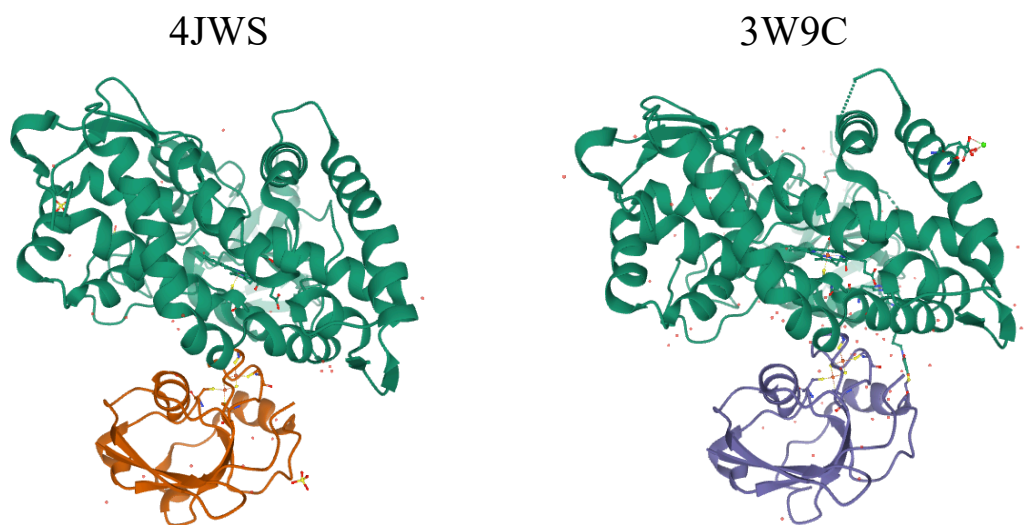


Scenario 2: **Project A** and **Project B** are excluded from racing sample



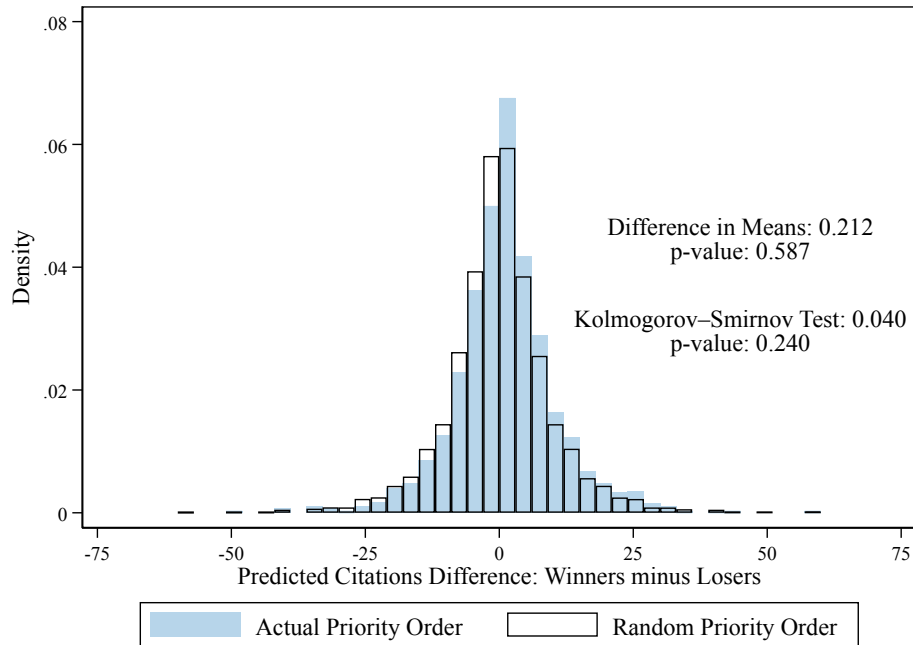
Notes: This figure shows visually the timing rule we use to define scoops. In the first example, Project *A* scoops Project *B* according to the rules, and therefore this example enters our regression sample. In the second scenario, Project *A* releases before Project *B*, but Project *B* had not yet deposited their data at the time of Project *A*'s release. Therefore this example would be excluded from our regression sample. We do not include these cases because Team *B* had full information about being scooped before they decided to deposit, and could therefore have decided to abandon the project without ever entering the data.

Figure 3: Example Priority Race — Pdx-P450cam Complex



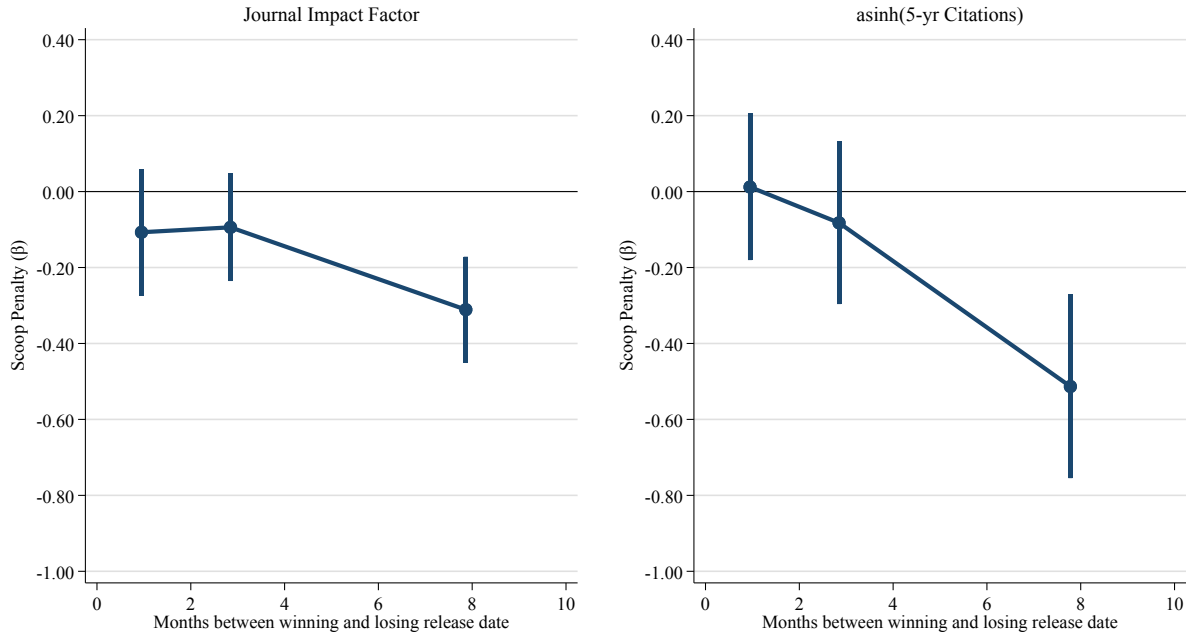
Notes: This figure presents a side-by-side comparison of the biological assembly models of the Pdx–P450cam complex protein deposited by two independent racing teams. According to the scoop definition in Section 2.4, structure deposit 4JWS scooped structure deposit 3W9C. See Table 1 for more details.

Figure 4: Histogram of Team Reputation Difference



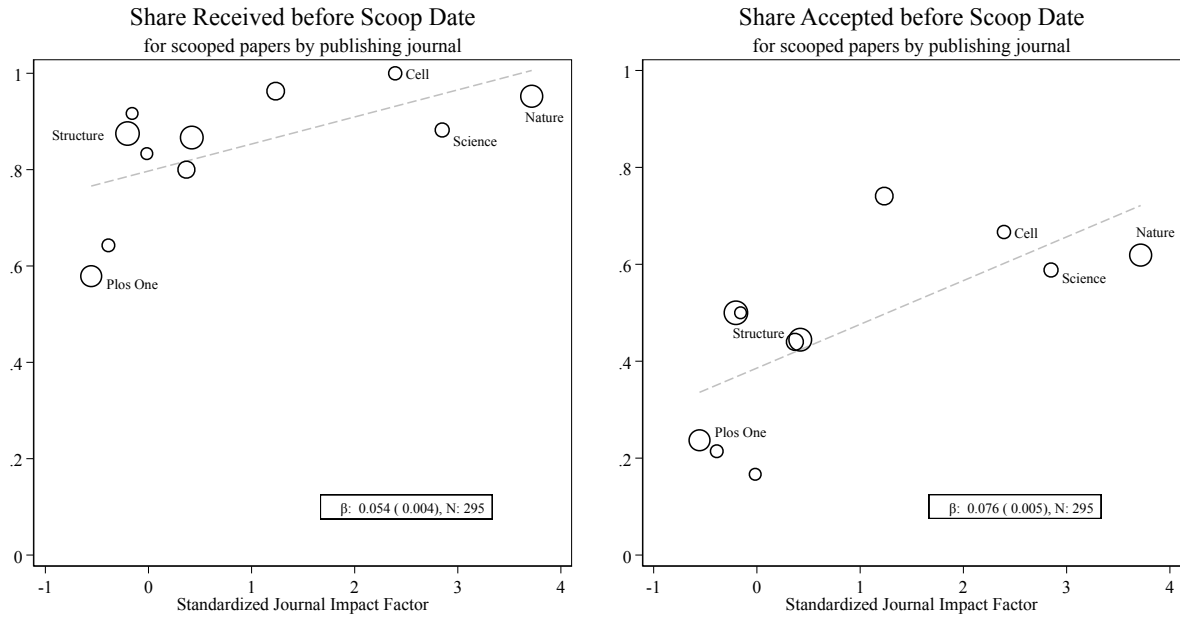
Notes: An observation in this figure is a racing pair. The blue distribution shows the actual difference in predicted citations. Bars to the right of zero represent instances when the winning team had higher predicted citations than the losing team, and bars to the left of zero represent instances when the winning team had lower predicted citations than the losing team. The white distribution outlined in black shows the difference in predicted citations if the winning and losing team 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.

Figure 5: JIF and Citation Penalty by Scooped Project Release Delay



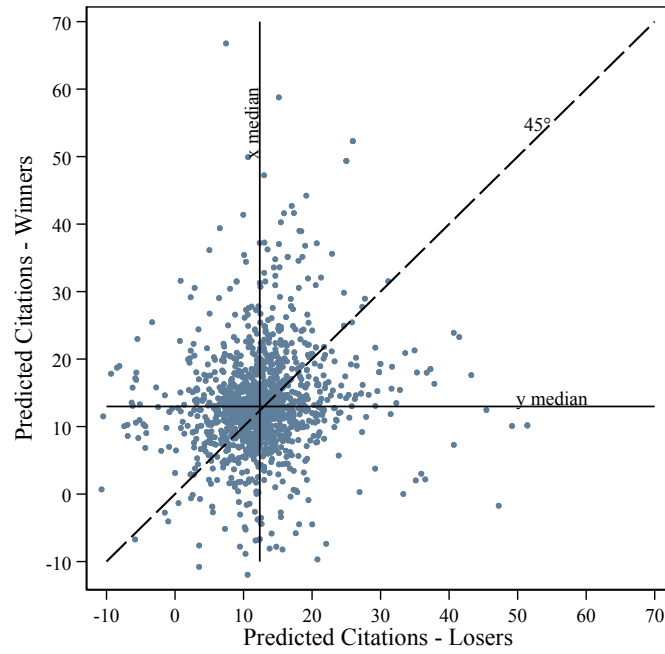
Notes: The sample of races is divided into three terciles along the distribution of time between winning and losing release date. Races are positioned along the x-axis at the average scoop release delay within each group. Projects released in close proximity are to the left, and those with a long delay are to the right. The y-axis shows the difference in journal impact factor and citations between the winner and loser in the left and right panel respectively.

Figure 6: Journal Placement and Timing of Scoops



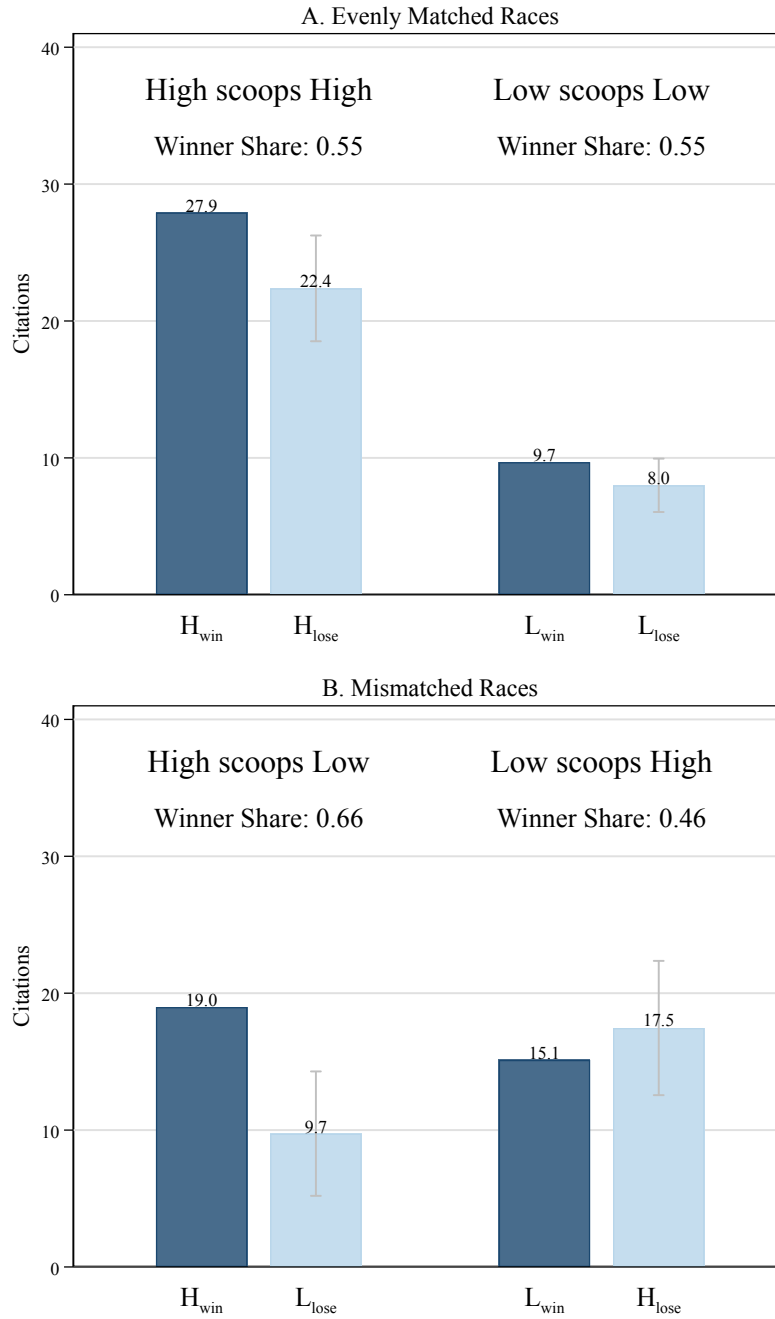
Notes: The figure reports the share of scooped papers that were received and accepted before the scoop date at different journals. Each circle represents one of the eleven largest journals that we collected supplemental data on the editorial timeline. Journals are arranged along the x-axis by their standardized journal impact factor. The size of the circles is proportional to the number of scooped papers published in each one.

Figure 7: Scatter Plot of Team Reputation Difference



Notes: An 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. Perfectly matched teams would lie on the 45-degree line. 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.

Figure 8: Priority Effect by Reputation Match-up



Notes: We divide the sample of races from Figure 7 into four quadrants, depending on whether the winners and losers are above- or below-median in expected 3-year citations defined by the Lasso estimation. In each panel, the dark bars represent the actual citations of the winning team and the light bars of the losing team. Panel A reports the comparison between evenly matched races, H scoops H or L scoops L. Panel B reports the comparison between mismatched races, H scoops L or L scoops H. The winner's share of total citations are reported above each set of bars.

A Data Appendix

A.1 Protein Data Bank

The Protein Data Bank (PDB) is the main source of project data we use to construct priority races. The first iteration of the PDB started in 1971, and the current archive is a global collaboration run by a non-profit organization called the World Wide Protein Data Bank (wwPDB). The wwPDB is a union of four existing data banks from around the world, including the Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB), Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), and Biological Magnetic Resonance Data Bank (BMRB). The data has been standardized and currently represents the universe of discoveries deposited in each of these archives. All new discoveries deposited to any database are transferred to, processed, standardized, and archived by the RCSB (Berman et al. 2006) at Rutgers University. Details about the PDB data can be found on their website.²⁴

We access the data directly from the RCSB Custom Report Web Service.²⁵ The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

- Structure Summary: structure ID, structure title, structure authors, deposit date, release date.
- Citation: PubMed ID, publication year, and journal name.
- Cluster Entity: entity ID, chain ID, sequence similarity clusters (BLAST algorithm for 90 percent and 100 percent sequence similarity, see section B below)
- Data Collection Details: collection date (the self-reported date the scientists generated diffraction data at a major synchrotron or in a home lab).

Additional data on cluster entities was accessed through a separate raw file archive at RCSB²⁶ on December 14, 2018. These files provided additional cluster groupings for the BLAST algorithm at 50 percent and 70 percent sequence similarity.

A.2 Citations and Journal Impact Factor

We use the journal names from the PDB extracts to link data to the Journal Citations Reports for journal impact factor and the Web of Science for citations.²⁷ We link the Journal Citations Reports using the journal name listed in the PDB. Each journal has an impact factor in each year and is calculated as the average number of citations per paper in the preceding two years. We standardize

²⁴<http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction>

²⁵<https://www.rcsb.org/pdb/results/reportField.do>

²⁶<ftp://resources.rcsb.org/sequence/clusters/> clusters50.txt and clusters70.txt

²⁷Both data sources were owned by Thompson Reuters at the time of access, but have since been sold to Clarivate Analytics.

impact factor in each year within the set of PDB-linked publications in our extracts each year. The citation data from the Web of Science and is restricted to citations from papers linked to PubMed IDs,²⁸ and self-citations are excluded. Citations are aggregated for each cited paper by publication year of the citing paper. When we report five-year citations, it represents the total number of citations in the publishing year and the subsequent five calendar years.

A.3 Altmetric.com Data

We use data from Altmetric.com to measure alternative forms of attention for academic research.²⁹ One limitation of the Altmetric data extract we use is that it only reports cumulative counts from the time of publication to the present (date of access: August 2nd, 2019). We account for the fact that scooped papers are published later and have less time to accumulate attention scores, using information about the change in score in recent time periods. The Altmetric.com data reports the change in attention in the past week, month, etc. We can therefore restrict the regression sample to races in which both teams had not accrued any additional attention in the amount of time that had passed between publications. For example, if paper A was released two months before paper B, we do not include this race in the analysis if paper A or paper B had accrued any additional attention in the most recent two months. This allows paper B to have the same window of time to accrue attention despite starting two months late. Because races in our sample end across a wide range of years, the regression coefficients are interpreted as the percent difference in outcomes for papers of an average vintage.

A.4 Editorial Dates

We access the received, accepted, and published dates from the websites of publications of Science, Nature Journals, Cell Press, and Public Library of Science. These data are used to compare the scoop date to the timeline of the journal review process as reported in Section 4.4.

We also use these data to look at the correspondence between the journal publication date and the release date. Appendix Figure A2 reports the correspondence between the PDB release date and the publication date for the 625 articles in the racing sample for which they are available. This correspondence is not exact for a few reasons. First, according to PDB policy, scientists are allowed to release their findings immediately after deposit, which could potentially come before the publication date. In typical practice, the scientists prefer to wait until publication so that other scientists cannot use the information for follow-on work until after publication. In fact, scientists prefer to wait for release as long as possible to maintain a competitive advantage, which was the motivation behind the 1998 policy change to align release and publication (Campbell 1998). Another reason that release may come earlier than publication is because of the policy that all data is released after one year. If a team takes more than one year to publish results after the deposit, they would be

²⁸Because structural biology falls squarely within the life sciences, restricting to citations with PubMed IDs does not have a large effect on citation counts.

²⁹<https://help.altmetric.com/support/solutions/articles/6000190631-using-altmetric-data-for-altmetrics-research>

forced to release at the one year point even if they eventually publish. Release sometimes happens after publication, but these cases should be rare and only be delayed for a few weeks. Any longer delays for release is either due to data errors or non-compliance with PDB policies.

Overall, 49 percent of the release dates are within two weeks of publication. This may lead to concerns about potential measurement error in the definition of the priority ordering. Throughout the paper, we always define the order of PDB release as the rule for being scooped. The community tracks public PDB releases carefully, so we believe this is a valid definition of priority. Publication dates are also complicated in recent years by the practice of online publication, which sometimes comes weeks before the print edition is published. But even if we prefer to consider only the publications as a claim to priority, our release date definition appears to usually correspond to the publication date ordering. In the 102 races where we have journal publication dates for the winner and loser, the priority ordering as defined by deposit corresponds with the priority ordering as defined by publication 82 percent of the time. To the degree that this is interpreted as measurement error, the scooped estimate will be somewhat attenuated.

A.5 Affiliations and University Rankings

Affiliation data is available from PubMed for most PDB deposits that resulted in a publication. Often the affiliation is only available for the first author of those publications, so we assign that affiliation to all authors on the publication. This assumption is more reasonable in structural biology than it is in economics for example, because cross-university collaboration is somewhat unusual in lab-based life sciences. The affiliations are contained in an author- or journal-reported text field that sometimes contains addresses or non-standard abbreviations. We standardize as many of these affiliations as possible using regular expressions and hand classification. We also assign as many affiliations as possible to their continent (Asia, North America, Europe, and other) to use as control variables. Affiliations are also categorized based on whether the affiliation is a university, non-profit research entity, or private corporation (typically a pharmaceutical company). In our full sample of projects (both racing and non-racing), there are 44,167 unique PubMed articles linked to the deposits. Of those papers, we were able to classify 71 percent to a standardized affiliation.

We link the university affiliations to the QS Top Universities Ranking for Life Sciences and Medicines.³⁰ This website provides rankings for 500 top academic programs based on surveys of academics and employers as well as citations per paper and h-index of the scientists affiliated with each department.

A.6 Name Disambiguation and Linked Author Papers in the PDB and PubMed

At various points in our analysis, we construct panel data of individual scientist and team productivity. First, we use measures of past PDB and PubMed productivity as control variables (Tables 3 and 4) and to predict citations as a measure of team reputation (Figures 7 and 8). Second, we use

³⁰<https://www.topuniversities.com/university-rankings/university-subject-rankings/2018/life-sciences-medicine>

a panel of publications to construct long-run outcomes in the years following a scoop event (Table 6). The PDB does not explicitly link authors between deposits, and neither PubMed nor Web of Science have author identifiers across publications. A further challenge is that many PDB deposits are not linked to a publication, so constructing control variables of past productivity is difficult using only publication data. We therefore use two separate approaches for constructing author-level panel variables: 1) Link PDB deposits by simple author name matching for control variables, 2) Use name disambiguation clustering from the Author-ity project (Torvik et al. 2005; Torvik and Smalheiser 2009) to count future publications and citations for long-run outcomes.

Simple Author Name Matching in PDB

In the first approach, we manually create a panel of author deposits and PDB-linked publications by matching last names and initials within the PDB. This name disambiguation procedure requires making assumptions about match reliability, and we follow the suggestions of Milojević (2013). We don't use additional information such as affiliations because they often change throughout a career, and are often only available for one author in the team.

The name disambiguation procedure using only last names and initials is more reliable in a smaller subset of academic papers. We therefore choose to focus the panel only on PubMed papers that are linked to the PDB instead of trying to use the full PubMed archive, which covers all of the medical and life science literature. This choice improves the reliability of our name-matching, but offers less information about academic productivity. Since we can use PDB name matching for unpublished deposits, we use this approach for constructing control variables for our main analysis.

Scientists usually identify themselves on publications with a consistent last name, but are sometimes inconsistent with their use of first and last initials, or first names and nicknames.³¹ According to Milojević (2013), there are two potential matching errors that should be accounted for. First, a given individual may be identified as two or more authors (splitting). Second, two or more individuals may be identified as a single author (merging). We follow the hybrid model they propose to deal with these concerns, using first and second initials to determine whether splitting or merging is likely, especially in cases of very common last names.

To connect names across PDB-linked publications, we use the following procedure:

1. Strip names of non-alphabetic characters and standardize spacing and hyphenation of compound last names.
2. Identify groups of paper-authors that have the same last name and first initial.
3. Look at the second initial to determine potential merging errors. We find that 96.5 percent of the last name/first initial groups have no second-initial conflict, so we treat these as distinct individuals

³¹Changes from maiden names to married names is also a potential source of error which we cannot account for, but this is becoming less common in recent years, especially among academics.

4. If we are unable to differentiate the individual using the second initial, (e.g. JACKSON, P; JACKSON, PA; and JACKSON, PS), we keep them as a merged name, but mark the group as “common.” These make up 3.5 percent of the sample.
5. We include a dummy control variable throughout the analysis that indicates the common names to help account for the possibility that name-matching errors are correlated with treatment.

We also use this panel to assign university rank and location controls. Racing projects sometimes go unpublished, so we cannot use the PDB-linked publication affiliation as a control variable in the main regression. Therefore we assign the most recent affiliation of the first author in the publication panel to improve the coverage of these control variables.

Author-ity Name Disambiguation

For long-run productivity outcomes, we focus on a broader set of PubMed publications. For most authors, structural biology in the PDB is only one part of their scientific portfolio. Since simple name matching is not reliable in the full sample of PubMed publications, we use a dataset called Author-ity (Torvik et al. 2005; Torvik and Smalheiser 2009) to help disambiguate names. The Author-ity project is a large-scale, data-driven effort that incorporates additional information about co-author networks and research topics to separate unique authors within the full PubMed database. Each iteration of an author last name and first initial that appears on a PubMed paper is grouped together with the other papers that the algorithm infers to be the same individual and is assigned a unique person ID. For example, the name JACKSON, P has 293 different person IDs in Author-ity, each with a distinct set of PubMed identified papers.

If all PDB deposits were published, we could simply link the PDB deposits to the associated authors using PubMed IDs. But many of the racing projects are not published, so we need to match PDB author names to Author-ity name clusters and determine which cluster the PDB author belongs to. We first merge the full list of PDB author names to Author-ity using last name and first initial. We then mark every instance where a PDB-linked PubMed ID matches to a PubMed ID cluster within the Author-ity merged name.

These two steps leave us with three distinct groups of author names in the PDB:

1. Names that do not match to any Author-ity cluster (11 percent of racing sample authors). These are individuals who deposit at least once in the PDB, but never publish a paper (e.g. a graduate student that does not pursue academia).
2. Names that have PubMed IDs that match to one and only Author-ity person ID (60 percent of racing sample authors). We take this exclusive matching as evidence that all instances of the name in the PDB is a single person that is represented by the matched Author-ity person ID.

- Names that have PubMed IDs that match to multiple Author-ity person IDs (29 percent of racing sample authors). These are common names that are likely distinct people within the PDB. We drop them from the long run analysis sample because we cannot determine which person is the author of a structure deposit that is not published.

We restrict our long-run analysis sample to the first two groups listed above (71 percent of racing sample authors). In this sub-sample, the individuals either never published a PubMed paper, or if they did, we have confidence that the PDB name represents a single individual.

Although our name disambiguation methods are not perfect, we rely on the assumption that any biases in our measures are equally distributed across winning and losing teams in a race. Given the balance in team characteristics shown in Table 3, we believe the winning teams are no more likely to have common names or mis-calculated productivity variables than losing teams, which should limit potential bias. To the extent that any remaining name matching mistakes create classical measurement error in the right-hand-side variables, it would attenuate our results.

B Protein Similarity and Race Definition

In this section we describe in detail the algorithm used to construct priority races used for our main analysis. Although the main text of the paper describes the basic rules for this sample construction, we report here a number of technical details and decisions that were used to construct the races in practice.

B.1 Sequence Similarity Algorithm

Each protein in the PDB is a chain composed of the 22 different types of proteinogenic amino acids in some combination. The order of these molecules in the chain defines the type of protein, and we use this code to compare the similarity of the proteins that scientists are working on. The PDB provides a clustering algorithm called the Basic Local Alignment Search Tool or BLAST (Altschul et al. 1990) which creates groupings of structure deposits that have identical or similar amino acid chains. The clusters can be defined at different thresholds of similarity, including 100 percent, 90 percent, 70 percent, and 50 percent. One possible approach to defining races would be to only focus on competing projects that determine the structure of proteins that are 100 percent similar. But in many cases, two proteins that are 90 percent similar or lower have many of the same defining features and functions within the same organism or across different species. Therefore, many interesting priority races are between teams working on very similar if not identical proteins. Following the similarity threshold chosen by (Brown and Ramaswamy 2007), we define racing for proteins all the way down to 50 percent similarity. We include races with a broad threshold in part to increase the sample size for our regressions, but also to include races over discoveries that were exceedingly different from any past structure discoveries.

Another tricky feature of the PDB data is that cluster groupings are sometimes defined at a level of granularity that is smaller than our outcome variables, which are defined at the structure

deposit and article level. Proteins are composed of “chains” of amino acids, and large proteins are often characterized in the PDB as a set of distinct chains. Further, chains of amino acids are often grouped as “entities”, and many proteins are combinations of two or more entities. This is relevant to our sample construction because the BLAST similarity algorithm clusters at the entity level rather than the protein level. In simple cases where proteins are made of a single entity, a new structure discovery might directly scoop another team working on the same entity. But in a few cases, a team working on a single entity might scoop a team that is working on a complex protein with multiple entities, only one of which was being worked on by both teams. These deposits will still be linked by the algorithm, but the interpretation of the scooping event is less obvious. We consider these cases to be “partial scoops” where some part of the scientific discovery was overshadowed by the winning team. Since outcomes are defined at the protein and paper level, including these partial scoops will potentially understate the effect of an average “full scoop.” We drop some very large proteins (such as the ribosome) that have more than 15 entities (0.7 percent of the sample). In these cases, the notion of a partial scoop is hard to define, as many different discoveries overlap at the entity level in sometimes complicated directions.

B.2 Procedure for defining races and scoop events

We follow the steps below to define priority races and scoop events. These steps are performed separately for four different similarity thresholds (50 percent, 70 percent, 90 percent, and 100 percent) and then combined in a final step.

1. Keep all clusters that have at least two deposits.
2. Sort the deposits within the clusters by release date, starting with the project that was released earliest. We focus only on cases of novel structure discoveries, so winners must be the first structure release in a given similarity cluster. We call this the priority deposit.
3. Compare the list of structure authors on the priority deposit with the list of authors on all subsequent deposits. Drop any follow-on deposits with one or more author names that were also on the priority deposit.³²
4. Drop all deposits with a deposit date after the release date of the priority deposit. This rule allows for multiple teams to be scooped by the same priority structure. See Section 2.3 for a discussion of this rule.

This procedure identifies a set of races that are defined within 50 percent, 70 percent, 90 percent, or 100 percent similarity clusters. We consolidate to a final analysis sample that minimizes duplicate races and duplicate deposits. Using this procedure leaves us with some proteins that are scooped at multiple levels. For example, protein A may be first and protein B may be second in a 100 percent

³²In a few cases, we see instances where the same team of authors deposited multiple structure discoveries in the same cluster around the same time. We keep only one of those structures per team and give preference to the first deposit that resulted in a publication or the first one deposited if they are never published.

similar cluster but are also the first and second in a 90 percent similar cluster (and 70 percent and 50 percent). To avoid counting this race multiple times, we keep only the instance defined in the 100 percent sample. In more complicated cases, protein A might be scooped by protein B that is 70 percent similar, but also scooped by protein C that is 100 percent similar either before or after protein B is released. In these cases, we always keep the scoop event at the closest similarity. So the race between protein A and protein B is dropped, and the race between protein A and protein C is kept. This leaves us with a final sample of mutually exclusive races where each scooped paper only appears once. Some winning deposits are allowed to scoop more than one protein, sometimes at different similarity levels. In Appendix Table [A2](#), we include robustness results of our main effects for races defined at the 100 percent level, and show that the results are comparable.

C Survey Text

This survey will ask you questions about the experience of being "scooped" as a scientist. Throughout the survey, we define being scooped as a case where a project is near completion and then a different lab publishes an article that is nearly identical. This means that most of the substantive research questions, methods, and findings are the same.

We focus only on cases where the project is near completion and ready for publication. Although some people experience being scooped at earlier stages of the research process, we do not consider those cases in this study.

Suppose you have just completed a very promising research project and you plan to submit it for publication this week.

What do you think is the probability that your project will be scooped between now and when it is published?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Probability of being scooped



Now suppose that just before you submit for publication, another lab publishes an article that is essentially identical to your project. They publish their paper in the journal *Science*. You have been scooped.

Would you choose to abandon your manuscript (meaning you do not submit for publication and drop the project)?

Yes, I would abandon the project

No, I would submit anyway

Assuming you do decide to submit, what do you think is the probability that your article will eventually be published?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Probability of Publication



If your competitor published their paper in *Science*, what do you think is the best journal that would accept your paper?

(list one academic journal)

Suppose your paper is successfully published. If your competitor's *Science* article receives 100 citations, how many citations do you expect your publication to receive?

D Proofs of Propositions

Proof of Proposition 1.

Consider two high-reputation labs, H_1 and H_2 . H_1 publishes before H_2 . The probability that H_1 is cited is:

$$\begin{aligned}
P(\hat{q}_1^H + f > \hat{q}_2^H) &= P((1 - \lambda)\alpha^H + \lambda s_1 + f > (1 - \lambda)\alpha^H + \lambda s_1) \\
&= P(\lambda(q + u_1) + f > \lambda(q + u_2)) \\
&= P(\lambda u_1 + f > \lambda u_2) \\
&= P\left(u_2 - u_1 < \frac{f}{\lambda}\right) \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) \\
&> \frac{1}{2}
\end{aligned}$$

using the fact that $(u_2 - u_1) \sim N(0, 2\sigma_u^2)$ and $f, \lambda > 0$. Similarly, consider two low-reputation labs, L_1 and L_2 . L_1 publishes before L_2 . Analogously, the probability that L_1 is cited is $\Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) > \frac{1}{2}$.

Proof of Proposition 2.

Consider a high-reputation lab and a low-reputation lab, H_1 and L_2 . H_1 publishes before L_2 . The probability that H_1 is cited is:

$$\begin{aligned}
P(\hat{q}_H + f > \hat{q}_L) &= P((1 - \lambda)\alpha^H + \lambda s_1 + f > (1 - \lambda)\alpha^L + \lambda s_2) \\
&= P((1 - \lambda)\alpha^H + \lambda(q + u_1) + f > (1 - \lambda)\alpha^L + \lambda(q + u_2)) \\
&= P((1 - \lambda)(\alpha^H - \alpha^L) + f > \lambda(u_2 - u_1)) \\
&= P\left(u_2 - u_1 < \frac{(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda}\right) \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&> \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) > \frac{1}{2}
\end{aligned}$$

again using the fact that $(u_2 - u_1) \sim N(0, 2\sigma_u^2)$ and $(1 - \lambda) > 0$, $\alpha_H > \alpha_L$. Similarly, consider a low-reputation lab and a high-reputation lab, L_1 and H_2 . L_1 publishes before H_2 . The probability

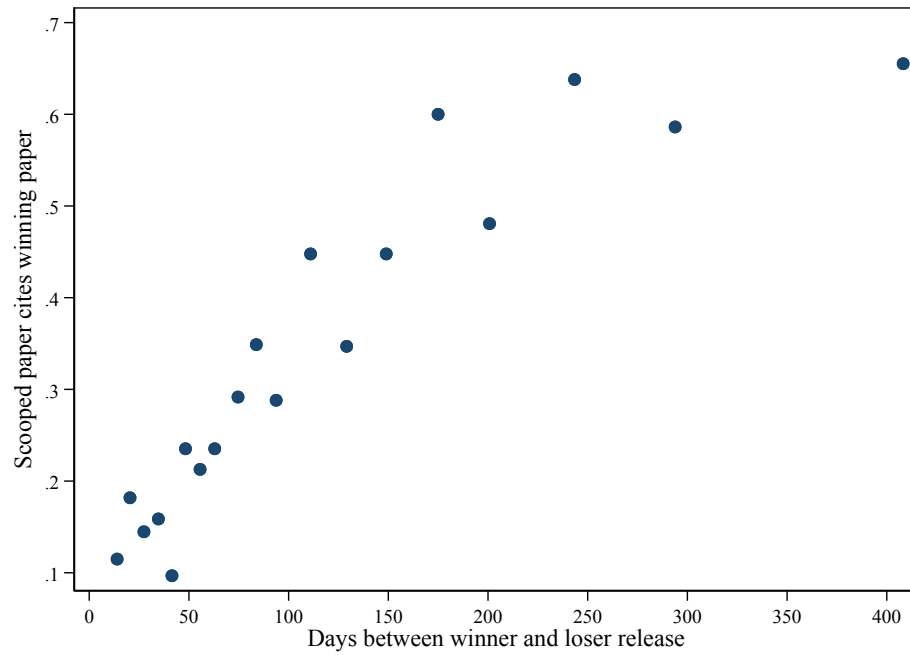
that L_1 is cited is:

$$\begin{aligned}
P(\hat{q}_L + f > \hat{q}_H) &= P((1 - \lambda)\alpha^L + \lambda s_1 + f > (1 - \lambda)\alpha^H + \lambda s_2) \\
&= P((1 - \lambda)\alpha^L + \lambda(q + u_1) + f > (1 - \lambda)\alpha^H + \lambda(q + u_2)) \\
&= P(-(1 - \lambda)(\alpha^H - \alpha^L) + f > \lambda(u_2 - u_1)) \\
&= P\left(u_2 - u_1 < \frac{-(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda}\right) \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{-(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{-(1 - \lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&< \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right).
\end{aligned}$$

Whether the expression is greater or less than $\frac{1}{2}$ depends on the magnitude of $(1 - \lambda)(\alpha^H - \alpha^L)$. More specifically, if $(1 - \lambda)(\alpha^H - \alpha^L) < f$, then $P(\hat{q}_L + f > \hat{q}_H) > \frac{1}{2}$. If $(1 - \lambda)(\alpha^H - \alpha^L) > f$, then $P(\hat{q}_L + f > \hat{q}_H) < \frac{1}{2}$.

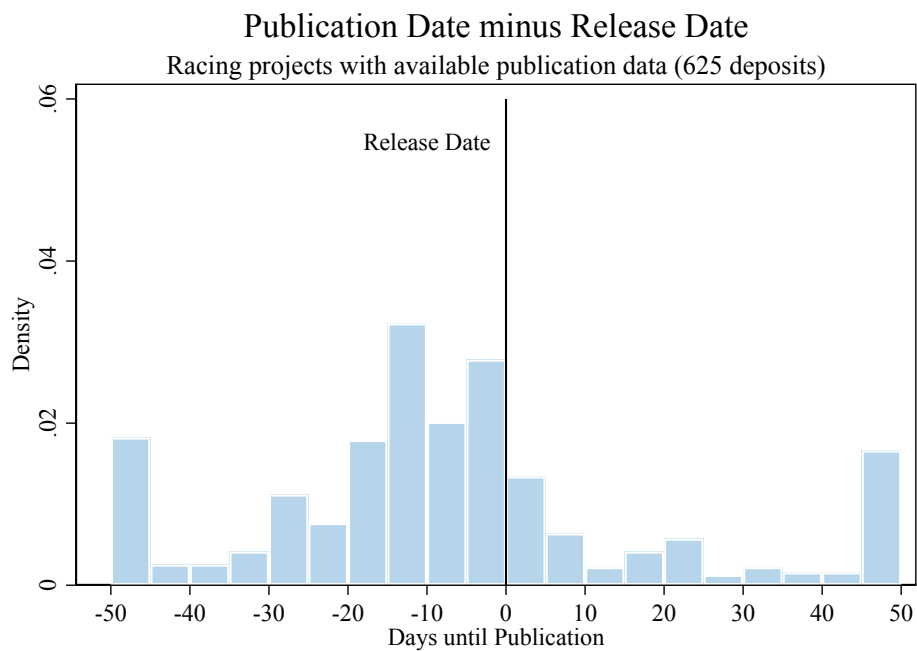
E Appendix Figures and Tables

Figure A1: Probability that Scooped Paper Cites Winning Paper by Release Date Gap



Notes: This binscatter shows the probability that the scooped paper cited the winning paper by the number of days between the release dates of the winning and losing projects. Sample is the set of races where both teams published and had a PubMed ID.

Figure A2: Correspondence Between Release Date and Available Publication Dates



Notes: This histogram shows the correspondence between PDB release date and publication date when publication dates are available from the editorial date supplement. Positive days means the publication came before release, and negative days mean it came after release.

Table A1: Lasso-selected Variables and Coefficients for Predicted Citations

Lasso-selected variables	Post-Lasso OLS coefficients
Number of authors	0.54
Affiliation in North America	1.81
Affiliation in Asia	-3.45
Non-academic affiliation	1.63
First author experience (years)	-0.20
First author PDB deposits, 5 prior years	-0.07
First author top-5 publications, 5 prior years	2.48
First author PDB deposits, all years squared	0.00
First author PDB deposits, 5 prior years squared	0.00
First author publications, 5 prior years squared	0.00
Last author experience (years)	-0.22
Last author PDB deposits, 5 prior years	-0.11
Last author publications, 5 prior years	0.02
Last author top-5 publications, all years	0.20
Last author top-5 publications, 5 prior years	2.16
Last author PDB deposits, all years squared	0.00
Last author PDB deposits, 5 prior years squared	0.00
Last author top-10 publications, 5 prior years squared	-0.01
<i>University rank bins:</i>	
1-10	3.47
71-80	-0.22
81-90	-1.05
101-110	-2.46
111-120	4.96
151-160	-2.81
171-180	-2.23
181-190	-0.42
211-220	-5.25
221-230	-7.14
271-280	-4.24
291-300	-3.11
361-370	-3.81
401-410	-2.79
451-460	-2.88
Constant	10.32
R-squared	0.103
N	58,758

Notes: This table presents results from a Lasso regression of 3-year unconditional citations on observable team characteristics. The model is estimated in the non-racing sample and uses data-driven and heteroskedasticity-robust penalization. Estimated coefficients are from a post-Lasso OLS regression of 3-year citations on selected regressors.

Table A2: Effect of Getting Scooped on Project Outcomes - 100 Percent Sequence Similarity

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
<i>Panel A. No controls</i>					
Scooped	-0.025 (0.025)	-0.177** (0.070)	-0.055* (0.032)	-0.271** (0.112)	-0.046** (0.021)
<i>Panel B. Base controls</i>					
Scooped	-0.034 (0.022)	-0.160** (0.074)	-0.048 (0.034)	-0.280** (0.110)	-0.031 (0.021)
<i>Panel C. PDS-Lasso selected controls</i>					
Scooped	-0.028 (0.018)	-0.176*** (0.052)	-0.054** (0.023)	-0.252*** (0.080)	-0.046*** (0.015)
Winner Y mean	0.882	-0.075	0.289	27.968	0.139
Observations	1,187	1,187	1,187	900	900

Notes: This table presents regression estimates of the scoop penalty comparable to Table 4 in the main text. This version restricts to protein clusters in which the BLAST algorithm classifies the protein sequences as being 100% similar. This sub-sample therefore offers the narrowest definition of a scoop where the racing projects are scientifically identical. See Table 4 notes for regression details.

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

Table A3: Effect of Getting Scooped on Three-Year Productivity

Dependent variable	Total count three years after race				
	PubMed Publications (1)	PDB Publications (2)	Top-ten publications (3)	Citation-weighted publications (4)	Top-10% cited publications (5)
<i>Panel A. All scientists</i>					
Scooped	-0.543 (0.520)	-0.077 (0.114)	-0.015 (0.061)	-0.159*** (0.040)	-0.224* (0.117)
Winner Y mean	27.208	4.274	2.179	297.224	4.650
Observations	10,157	10,157	10,157	7,726	7,726
<i>Panel B. Novices</i>					
Scooped	-0.036 (0.141)	-0.078 (0.096)	0.073* (0.040)	-0.249*** (0.085)	-0.041 (0.063)
Winner Y mean	2.293	1.091	0.334	43.853	0.677
Observations	2,401	2,401	2,401	1,819	1,819
<i>Panel C. Veterans</i>					
Scooped	-0.398 (0.796)	-0.039 (0.160)	-0.036 (0.088)	-0.141*** (0.044)	-0.314* (0.168)
Winner Y mean	36.797	5.556	2.910	399.891	6.253
Observations	6,809	6,809	6,809	5,210	5,210

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 4 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 5 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than eight years of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

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

Table A4: Effect of Getting Scooped on Ten-Year Productivity

Dependent variable	Total count ten years after race				
	PubMed Publications (1)	PDB Publications (2)	Top-ten publications (3)	Citation-weighted publications (4)	Top-10% cited publications (5)
<i>Panel A. All scientists</i>					
Scooped	-2.766 (2.773)	0.254 (0.517)	-0.284 (0.230)	-0.036 (0.071)	-0.920 (0.594)
Winner Y mean	91.467	13.942	7.077	926.740	14.062
Observations	5,373	5,373	5,373	3,124	3,124
<i>Panel B. Novices</i>					
Scooped	0.134 (0.825)	0.303 (0.470)	0.229 (0.181)	-0.125 (0.150)	0.563* (0.306)
Winner Y mean	9.886	3.734	1.299	122.905	1.792
Observations	1,260	1,260	1,260	743	743
<i>Panel C. Veterans</i>					
Scooped	-5.310 (4.043)	-0.890 (0.707)	-0.683** (0.323)	-0.114* (0.063)	-1.736** (0.833)
Winner Y mean	123.981	18.064	9.393	1241.262	18.856
Observations	3,626	3,626	3,626	2,088	2,088

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 4 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 5 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than eight years of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

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

Table A5: Structure Quality Balance in High- and Low-Reputation Match-ups

Matchup subsample	Loser structure quality (1)	Winner structure quality (2)	Difference: (lose - win) (3)	Std. error of difference (4)	Observations (5)
<i>Panel A. Resolution (\AA)</i>					
High scoops High	2.586	2.507	0.078	(0.216)	672
Low scoops Low	2.340	2.227	0.113	(0.128)	467
High scoops Low	2.188	2.205	-0.017	(0.074)	498
Low scoops High	2.158	2.155	0.003	(0.053)	652
<i>Panel B. R-free goodness-of-fit</i>					
High scoops High	0.256	0.249	0.007	(0.004) **	649
Low scoops Low	0.245	0.242	0.002	(0.004)	462
High scoops Low	0.242	0.245	-0.003	(0.004)	490
Low scoops High	0.240	0.239	0.002	(0.004)	650

Notes: This table compares structure quality metrics of winning and losing projects in subsamples of races divided by team reputation as measured by predicted citations. Lower values of resolution and r-free represent better quality. Observations are at the structure level. Column 1 shows the means of the losing projects in the racing sample, and column 2 shows the means of the winning projects in the racing sample. Column 3 shows the difference between the losing and winning projects, and column 4 shows the heteroskedasticity-robust standard error of the difference.

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