Scooped! Estimating Rewards for Priority in Science*

Ryan Hill† Carolyn Stein‡

June 22, 2020

Click here to download the latest version.

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. As a share of total citations, we estimate that scooped papers receive a credit share of 45 percent. This is larger than the theoretical benchmark of zero percent suggested by classic models of innovation races. We conduct a survey of structural biologists which suggests that active scientists are more pessimistic about the cost of getting scooped than can be justified by the data. Much of the citation effect can be explained by journal placement, suggesting editors and reviewers are key arbiters of academic priority. Getting scooped has only modest effects on academic careers. Finally, we present a simple model of statistical discrimination in academic attention to explain how the priority reward system reinforces inequality in science, and document empirical evidence consistent with our model. On the whole, these estimates inform both theoretical models of innovation races and suggest opportunities to re-evaluate the policies and institutions that affect credit allocation 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 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. We apologize to any authors that were inadvertently scooped by this paper; we hope that they also receive their due share of recognition.

†MIT Economics Department, ryanhill@mit.edu

‡MIT Economics Department, cstein@mit.edu

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). According to conventional wisdom, 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 (Hagstrom 1974). Many theoretical papers about innovation races conceptualize the reward structure as winner-take-all (Loury 1979; Fudenberg et al. 1983; Harris and Vickers 1985; 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 (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, we want 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 measure 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 determine whether two teams are working on the same question. 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 analyze 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 140,000 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, and are otherwise redundant. Second, projects are grouped 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 5 percent of all structural biology projects are involved in a late-stage race to publication. These races are composed of a diverse set of scientific teams from different countries, institutional prestige, and experience.

Our definition of scoops focuses only on late-stage races where both teams are on the cusp of publication. Some researchers 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 had deposited. 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 and credibly estimate the causal effect of getting scooped. We estimate the effect of winning a race using the naturally-occurring variation in the priority

[1] In a survey of structural biologists we conduct, 80 percent of the respondents say they rarely if ever circulate their findings in a working paper or pre-print prior to journal publication.

ordering of races. Therefore, omitted variables bias is a threat to the causal interpretation of the estimates. If the winners are positively selected in experience, research ability, or university prestige, our estimates will be upward biased. 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 18 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 5 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 2 percent less likely to be actively depositing in the PDB five years after they were scooped. We do not find significant effects on intensive margin publication rates. However, scooped scientists receive 20 percent fewer citations to their future work, an effect that is stronger for novice scientists (28 percent) than their veteran co-authors (17 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 rank 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. Then a high-reputation team that wins the race not only receives a premium for priority, but also a boost in citations because of their high reputation. 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 3 percent chance we document in the PDB data. We then ask them to predict the probability of publication and expected citations if they are scooped by a competitor's paper. They predict that they only have a 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.

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 some 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). These models pre-suppose a winner-take-all reward,

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

which has implications for the outcome of innovation tournaments and the strategic behavior of the participants (Fudenberg et al. 1983; Harris and Vickers 1985, 1987). 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. 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 data collection 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. 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. Unlike most academic disciplines, structural biology has very detailed and organized project-level data, which allows us to systematically track soon-to-be-published projects that are at risk of being scooped. 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). By understanding a protein's three-dimensional structure, scientists can better understand how the protein functions. 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 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, although 15 percent of the scientists in our sample work at non-profit research laboratories or private companies.

[3]The 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).

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 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 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.

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

2.3 Identifying Priority Races: Challenges and Solutions

Identifying priority races in scientific data is difficult for three distinct 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 (Jardim et al., 2018; 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 likely 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. 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, most 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 the previous section, 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 get published. Given that scientists might abandon papers that get scooped, having this record of projects that are never published 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

[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.

"cluster" in the PDB).[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 deposit. These follow-on deposits are either scooped projects that proceeded to publication, 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 was 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

[6] If a protein is scooped by more than one other protein, we give preference to the protein that is biologically closer (i.e. in the "higher" cluster). See Appendix B for details on the data construction. For robustness, we can restrict to scoops by proteins within the same 100 percent cluster, and find similar results.

[7] For 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.

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

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. 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: 1ZIW and 2A0Z. Both are structures of the Toll-like Receptor 3, a protein involved in the identification of pathogens and other immune functions. Figure 3 shows the nearly identical biological assembly models that each team deposited confidentially to the PDB. The scientists at Scripps Research Institute (1ZIW) collected their data a few months after the scientists at the National Institutes of Health (2A0Z) (March 13, 2005 versus October 24, 2004). However, by the time of deposit, Scripps had pulled ahead, depositing two months before NIH (April 27, 2005 versus June 27, 2005). Ultimately, Scripps won the priority race, with their structure being released one month before NIH (June 28, 2005 versus August 2, 2005). Importantly, when NIH deposited their structure on June 27, 2005, Scripps had not yet released their structure. This means that NIH 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 (1ZIW) and the loser (2A0Z), we observe that the winning paper was more successful. It was published in a better journal (*Science*, with an impact factor of 30.9, versus *The Proceedings of the National Academy of Sciences*, with an impact factor of 10.2) and received about 50 percent more citations.

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 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.[9] Second, we drop projects that are published in a paper that is linked to 15 or more other structures.

[9]Proteins 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.

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 paper has a strong claim of priority and that the order of release is more likely to correspond to the order of publication.[10]

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 publication that evaluates journal influence using a metric called "journal impact factor". Let $Cites^j_{t,t-k}$ be the number of citations that journal j received in year t for articles written in year $t-k$. Let $Articles^j_{t-k}$ be the number of articles published by journal j in year $t-k$. Then journal j's impact factor in year t is given by:

$$JIF^j_t = \frac{Cites^j_{t,t-1} + Cites^j_{t,t-2}}{Articles^j_{t-1} + Articles^j_{t-2}}.$$

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 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[11]

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. We link the PDB to the Web of

[10]The 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. See Appendix Section A.4 for more details on the correspondence between the PDB release date and publication date.

[11]Top-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*

Science using PubMed identifiers, which are unique IDs assigned to research papers in the medical and life sciences. 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 this outcome to 1999-2013 to allow for time for publications to accrue citations we can observe.

We also construct career histories of outcomes before and after the priority date of each race for control variables and long-run outcomes. We sum publications, top-10 publications, citation-weighted publications, and "hit" papers for the years preceding and following the treatment date. The PDB, PubMed, and Web of Science data offer only limited information about careers, and authors are not linked administratively over time. However, since journal publications are the primary productive output of scientific careers, we can use these publication histories to infer the time-span of participation and intensity of scientific activity. Besides analyzing the effects of race outcomes on the intensive margin of publication, we also consider the extensive margin of exit from publishing PDB related papers altogether. We mark an individual as having exited structural biology if there is a hiatus of at least five years in their publication record. If this hiatus began in the five years following the priority date, the individual is marked as having exited. More details about the panel of papers and name disambiguation is provided in Appendix A.

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 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, Public Library of Science (PLOS), and one publisher that requested to remain anonymous. This subsample covers 23 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 5 percent of the structures in our sample are involved in a priority race. We look at both team characteristics and deposit outcomes. Teams involved in priority races tend to be smaller, younger, and more likely to come from a top university. The racing scientists were also more likely to work in Asia, and less likely in North America. The deposit outcomes suggest that proteins involved in priority races are scientifically more important. Proteins in the racing sample are more likely to be published, appear in higher-ranked journals, and receive more citations.

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

$$Y_{ip} = \alpha + \beta Scooped_i + \mathbf{X_i'}\delta + \gamma_p + \epsilon_{ip} \tag{1}$$

where Y_{ip} is an outcome, such as publication, journal impact factor, or citations. *Scooped* is an indicator for losing a priority race, $\mathbf{X}$ is a vector of covariates, and γ is the coefficient on 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* 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 1, 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

$$Y_{isp} = \alpha + \beta Scooped_{is} + \mathbf{X_{is}'}\delta + \gamma_p + \epsilon_{isp} \tag{2}$$

where $Scooped_{is}$ is a dummy equal to one if scientist s was scooped on project i. $\mathbf{X_{is}}$ 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 is the coefficient on the 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.

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.[12]

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

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

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. 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 Lasso[13] 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. We also 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

[13]Least Absolute Shrinkage and Selection Operator (Tibshirani 1996).

et al. (2014) to optimally select controls variables. Consider a partially linear model similar to equation 1

$$Y_{ip} = \alpha + \beta Scooped_i + \mathbf{g}(\mathbf{Z_i}) + \gamma_p + \epsilon_{ip} \tag{3}$$

where $\mathbf{Z_i}$ is a large set of control variables. Also, 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 $\mathbf{g}(\mathbf{Z_i})$. 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 $Scooped_i$ on $\mathbf{Z_i}$ to select a set of regressors that are predictive of treatment. Then it uses a second Lasso regression of Y_{ip} on $\mathbf{Z_i}$ 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 $Scooped_i$. 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 function[14], 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, 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 1 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

[14]The 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 $asinh(x) = \log\left(x + \sqrt{x^2+1}\right)$. The coefficients on variables transformed by the hyperbolic sine function can be interpreted similarly to logs (i.e. proportionally).

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.

The baseline publication rate for winning projects is high at 88 percent. Scooped projects are 2.5 percentage points less likely to publish their results. This small 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 low-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 non-scooped papers. In column 3, this translates to a six percentage point (18 percent) decrease in the probability of 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. 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 strongly 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.[15]

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 forms 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

[15]The 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.

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. Getting scooped leads to a 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 citations. Mentions of a paper on Twitter fall by 12 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. The regressions follow equation 2 and are run at the level of the individual scientist. 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. 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 two years or less since their first publication at the time of the scooping event. Panel C restricts to veterans, which are all scientists not defined as novices.[16]

Overall, scientists are less likely to still be actively publishing PDB-linked articles if they have been scooped. In the full sample, 80 percent of authors persist for at least five years following the priority date, and the scooped scientists are 2.2 percentage points less likely to persist than the winning scientists. This effect is only statistically significant for veteran scientists. The effect

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

among novices is similar in sign and magnitude, but we cannot reject the effect is different than zero, in part because of the smaller sample size. We can only calculate this extensive margin effect using PubMed publications that are linked to PDB projects. It may be that getting scooped is not enough to push scientists all the way out of academia, but could push them to other fields of research in biology that do not involve structure determination. Perhaps the losing authors are discouraged by the competitive nature of races in crystallography and decide to dedicate themselves to other questions in biology. [17]

The negative extensive margin result does not translate into significant changes in intensive margin publishing for novices or veterans. Losing teams have no statistically significant differences in publications in the following years as shown in column 2, 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 20 percent fewer citations to their subsequent 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 28 percent fewer citations, while veterans receive only 17 percent fewer citations. The effect on "hit" papers is reported in column 5 and also suggests that getting scooped decreases attention to future work. The full sample of scientists are 16% less likely to have a hit paper following a scoop event on average. The negative effect in this case is larger in percent terms for veterans (-16%) than it is for novices (-12%), although veterans have a higher base rate of hit papers than novices. We also consider outcomes in the following three years in Appendix Table A2 and ten years in Appendix Table A3. 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 ask 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 very soon after they deposit their findings receive a much larger penalty than teams that are scooped 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.

[17] The full PubMed database could shed light on full exit from academia, but due to data availability and challenges with name disambiguation, we are not able to analyze these outcomes yet. See Appendix A for a discussion of name disambiguation.

4.4.1 Decomposing the Citation Effect

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 5 percent and is no longer statistically significant. 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 Getting Scooped and the Publication Timeline

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."[18] 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 in almost all cases. 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 5 further shows that this pattern varies greatly by the impact factor of the journal that eventually publishes the scooped paper. For lower ranked journals, such as *PLOS One*, only 60 percent of scooped papers had been received by the journal on the date they were scooped, and just

[18] See 2019 *Science* Instructions for Reviewers of Research Articles: https://www.sciencemag.org/sites/default/files/RAinstr19.pdf

over 20 percent had been accepted. Among the twelve large journals for which we have received and accepted data, 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 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 were only willing to provide the data anonymously, stating their concern about presenting public evidence that they publish scooped papers. Figure 5 shows that this anonymous editor may have realistic grounds for concern, as they are much more likely to accept scooped papers than other large journals in our sample.

5 Reputation and the Scoop Penalty

In this section we document large heterogeneity in priority rewards between race winners and losers depending on their preexisting reputation. 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 heterogeneity 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 or to compare the merits of two similar papers. 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:

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

where $\gamma = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_u^2}$ 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 γ 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, $q^H \sim N\left(\alpha^H, \sigma_q^2\right)$ while for L labs, $q^L \sim N\left(\alpha^L, \sigma_q^2\right)$. 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:

$$\hat{q}^H(s) = \gamma s + (1 - \gamma)\alpha^H$$
$$\hat{q}^L(s) = \gamma s + (1 - \gamma)\alpha^L.$$

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 Heterogeneity by Academic Reputation

To test our model, we measure the share of total citations received by winning and losing labs, and compare these shares by the reputations of the racing labs. 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.

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 predicted citations correspond to the H labs, while teams below median correspond to the L labs. In Figure 6 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 6 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.

Figure 7 shows the average citation counts by matchup type, as well as the citation shares. Panel A shows the evenly matched races, where we have shut down the reputation effect and isolated 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 whether we are in a H versus H matchup (winning team receives 55 percent of the total citations) or an L versus L matchup (winning team receives 55 percent of the total citations). This is consistent with the prediction from proposition 1.[19]

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 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 6. 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 inform them 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

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

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?" This hypothetical scenario was designed to match the instances of racing that we have in our data. Two projects are near completion, and one is preempted by the other. We can therefore compare our magnitudes to the average beliefs of the survey respondents.

Table 8 reports the average responses of the biologists and compares them to the magnitudes estimated in the PDB data. Because we tried to pose the survey questions as a concrete hypothetical for clarity and consistency, the racing situation does not exactly match the average situation in the data. In particular, 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,[20] 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 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 29 percent as a share 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 7 percentage points lower probability that they will be able to publish their scooped paper. Perhaps surprisingly, both types of respondents had similar expectations for

[20]Specifically, 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.

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 (9 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 estimate a small increase in the probability of exiting the field of structural biology, and find 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, there is very little impact on their citation rates, and 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 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 an infeasible 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 ongoing research, we analyze objective measures of the quality of crystal diffraction data and corresponding structure models to study how racing in science affects quality outcomes. 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.

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.

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, Giuilio, 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.

Harris, Christopher and John Vickers, "Perfect Equilibrium in a Model of a Race," *Review of Economic Studies*, April 1985, *102* (2), 193–209.

_ **and** _ , "Racing with Uncertainty," *Review of Economic Studies*, January 1987, *54* (1), 1–21.

Hill, Ryan, "Searching for Superstars: Research Risk and Talent Discovery in Astronomy," *Working Paper*, 2019.

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.

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.

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.

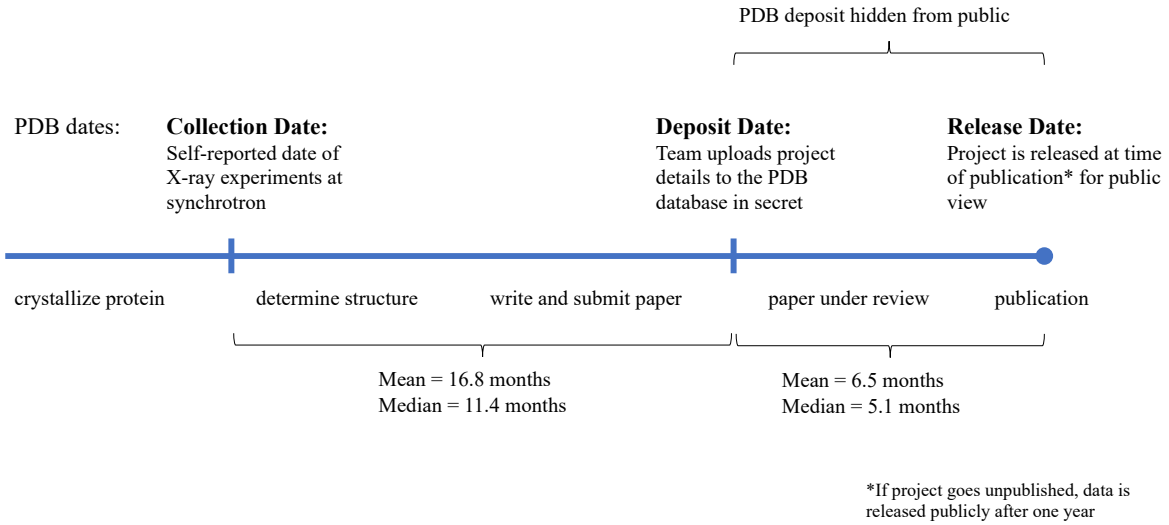
Wlodawer, Alexander, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski, "Protein Crystallography for Non-Crystallographers, or How to Get the Best (But Not More) From Published Macromolecular Structures," *FEBS Journal*, January 2008, *275* (1), 1–21.

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.

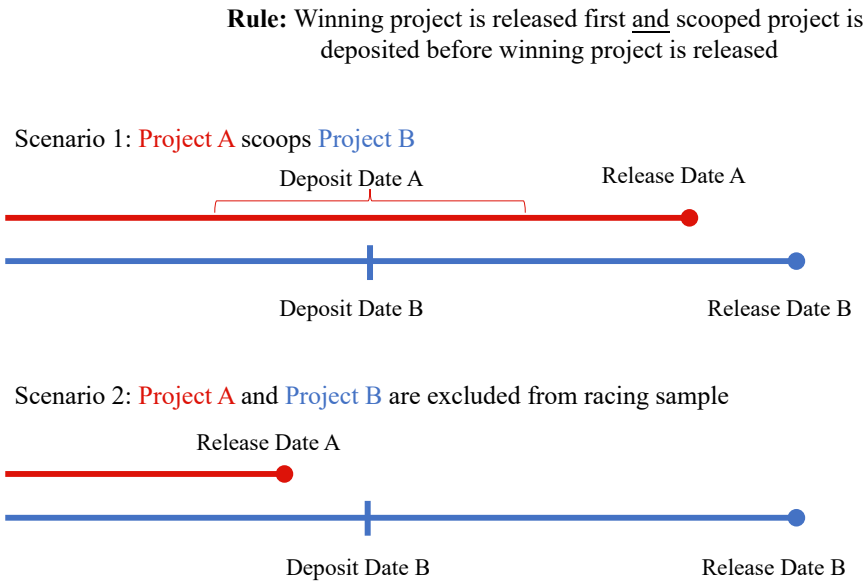
Figures and Tables

Figure 1: Project Timeline and Key Dates



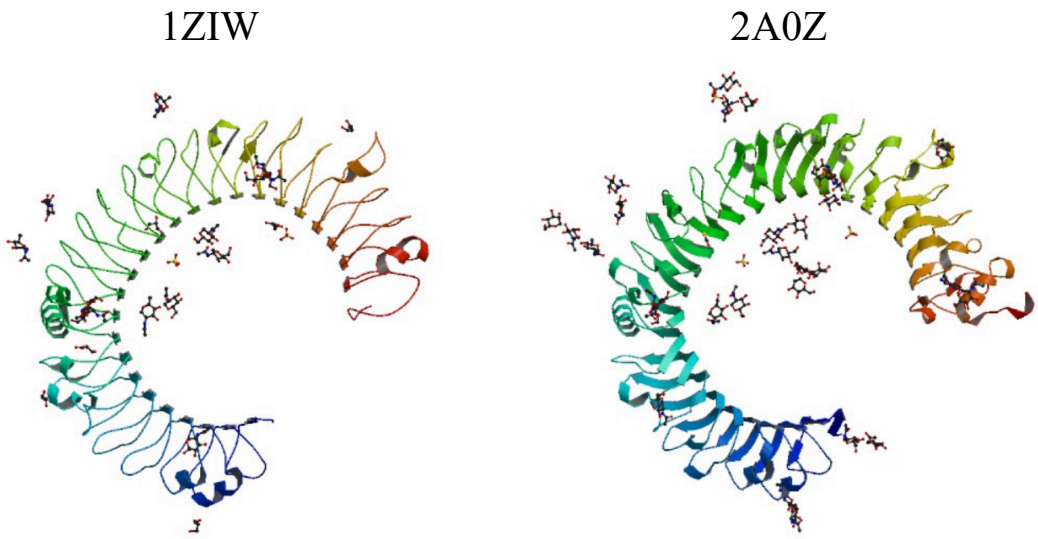
Notes: This figure shows the timeline of a PDB project. Dates in bold above the line are observed in our data. Events listed below the timeline are the approximate timing of other project events including the submission and review process. Deposit event and structure data is hidden from public until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

Figure 2: Defining Priority Races



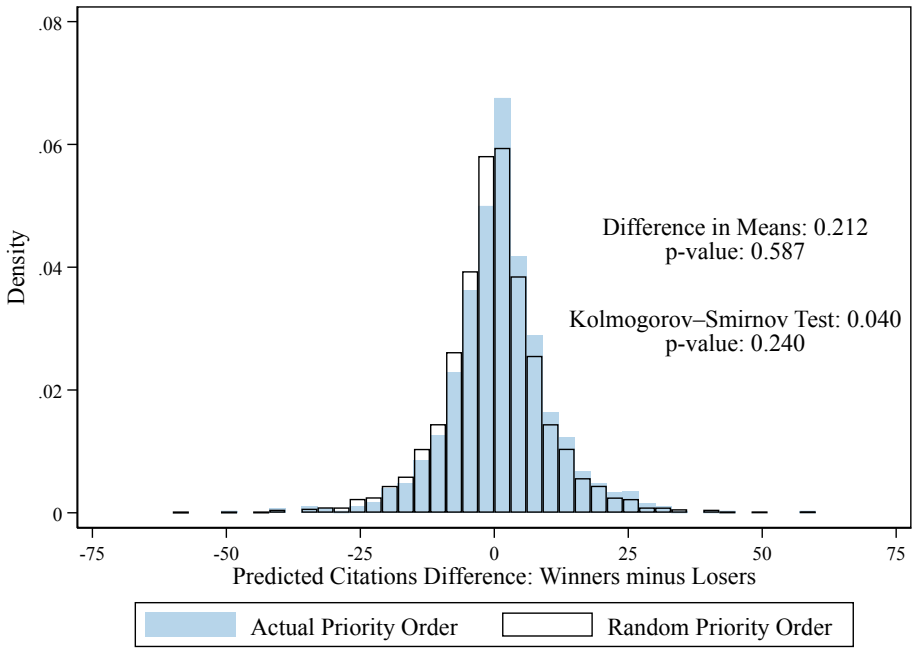
Notes: This figure shows visually the timing rule we use to define scoops: (1) the winning project is released first and (2) the scooped project is deposited before the winning project is released. In the first example, Project A scoops Project B according to the rule, and therefore this example enters 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 — Toll-like Receptor 3



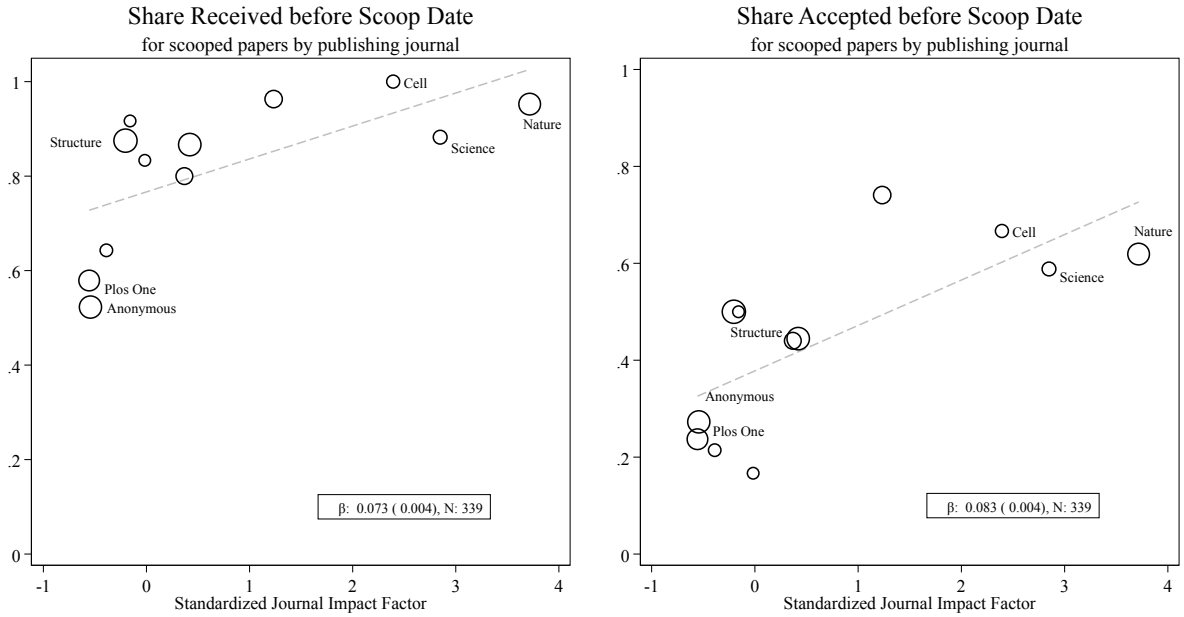
Notes: This figure presents a side-by-side comparison of the biological assembly models of the Toll-like Receptor 3 protein deposited by two independent racing teams. According to the scoop definition in Section 2.4, structure deposit 1ZIW scooped structure deposit 2A0Z. See Table 1 for more details.

Figure 4: Histogram of Team Reputation Difference



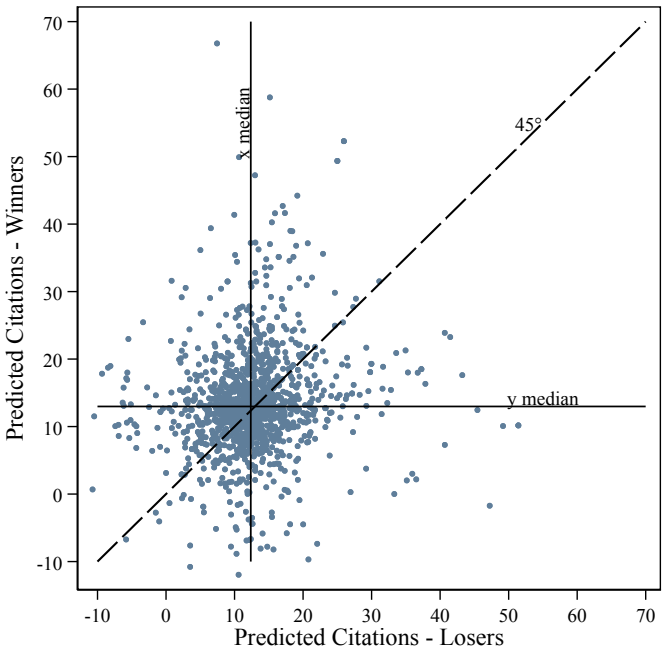
Notes: An observation in this figure is a racing pair. The blue distribution shows the actual difference in predicted citations. Bars the the right of zero represent instances when the winning team had higher predicted citations than the losing team, and bars to the left of zero represent instances when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



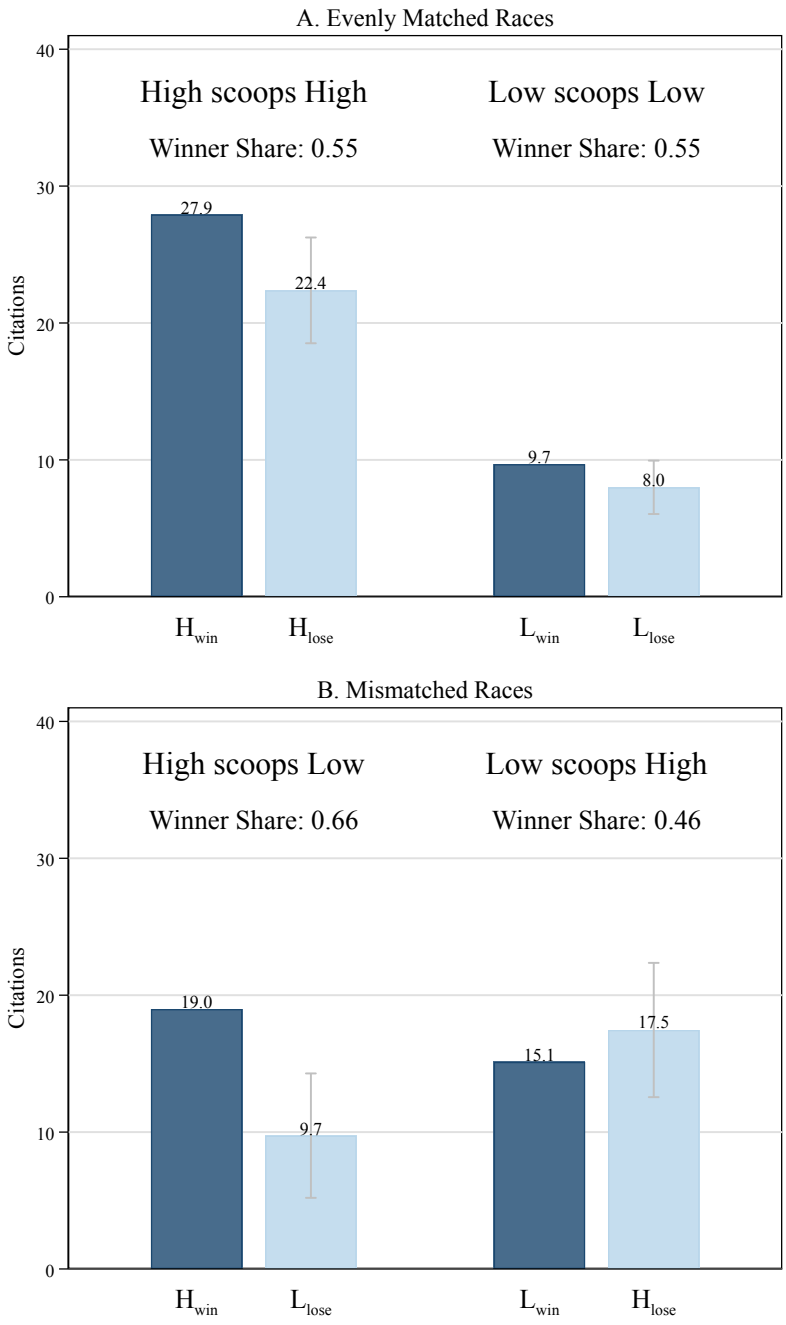
Notes: The figure reports the share of scooped papers that were received and accepted before the the scoop date at different journals. Each circle represents one of the twelve largest journals that we collected supplemental data on the editorial timeline. Journals are arranged along the x-axis by their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

Figure 6: 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 y-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 7: Priority Effect by Reputation Match-up



Notes: We divide the sample of races from Figure 6 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.

Table 1: Example Priority Race — Toll-like Receptor 3

	Winning project	Scooped project
PDB structure ID	1ZIW	2A0Z
Protein name	Toll-like Receptor 3	Toll-like Receptor 3
Paper title	"Crystal structure of human toll-like receptor 3 (TLR3) ectodomain"	"The molecular structure of the toll-like receptor 3 ligand-binding domain"
Key dates:		
Collection date	March 13, 2005	October 24, 2004
Deposit date	April 27, 2005	June 27, 2005
Release date	June 28, 2005	August 2, 2005
First author affiliation	Scripps Research Institute	National Institutes of Health
Journal	*Science*	*Proc. of the Natl. Acad. of Sciences (PNAS)*
Journal impact factor	30.9	10.2
Citations:		
Three years	140	90
Five years	196	129
Ten years	311	212

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)
Panel A. Team characteristics				
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) ***
Panel B. Project outcomes				
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) ***
Panel C. Project altmetrics				
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)	
Panel A. Team characteristics						
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)	
Panel B. First author productivity (prior five years)						
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)	
Panel C. Last author productivity (prior five years)						
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)	
Panel D: Project quality metrics						
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	F-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)
Panel A. No controls					
Scooped	-0.027*	-0.187***	-0.065***	-0.243***	-0.037**
	(0.015)	(0.044)	(0.020)	(0.070)	(0.014)
Panel B. Base controls					
Scooped	-0.026**	-0.176***	-0.062***	-0.208***	-0.028**
	(0.013)	(0.044)	(0.020)	(0.063)	(0.014)
Panel C. PDS-Lasso selected controls					
Scooped	-0.025***	-0.178***	-0.060***	-0.197***	-0.035***
	(0.010)	(0.032)	(0.014)	(0.045)	(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 asinh(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)	Atltmetric attention (6)
Panel A. No controls						
Scooped	-0.452***	-0.107**	-0.037**	-0.007	-0.114	-0.240**
	(0.152)	(0.042)	(0.018)	(0.028)	(0.077)	(0.094)
Panel B. Base controls						
Scooped	-0.425***	-0.092**	-0.030	0.001	-0.087	-0.199**
	(0.144)	(0.043)	(0.020)	(0.031)	(0.074)	(0.090)
Panel C. PDS-Lasso selected controls						
Scooped	-0.453***	-0.108***	-0.035**	-0.008	-0.101*	-0.237***
	(0.105)	(0.032)	(0.014)	(0.021)	(0.054)	(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

		Total count five years after race			
Dependent variable	Active in PDB 5 years later (1)	Publications (2)	Top-ten publications (3)	Citation-weighted publications (4)	Top-10% cited publications (5)
Panel A. All scientists					
Scooped	-0.023**	0.325	0.047	-0.199***	-0.159***
	(0.010)	(0.264)	(0.105)	(0.053)	(0.052)
Winner Y mean	0.797	13.541	4.441	187.129	1.529
Observations	6,642	12,488	12,488	9,297	9,297
Panel B. Novices					
Scooped	-0.013	-0.054	-0.076	-0.282**	-0.121***
	(0.022)	(0.183)	(0.065)	(0.112)	(0.039)
Winner Y mean	0.587	2.667	0.929	50.692	0.440
Observations	2,273	3,554	3,554	2,868	2,868
Panel C. Veterans					
Scooped	-0.024**	0.504	0.118	-0.167***	-0.165**
	(0.009)	(0.373)	(0.145)	(0.051)	(0.071)
Winner Y mean	0.930	19.042	6.221	263.894	2.143
Observations	4,027	8,251	8,251	5,913	5,913

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 two 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

	Five-year citations			
Dependent variable	(1)	(2)	(3)	(4)
Scooped	-0.164***	-0.114***	-0.107***	-0.047*
	(0.032)	(0.028)	(0.028)	(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	Comparable subsample	All respondents	Below-median reputation	Above-median reputation	Column (4) - (5) difference
	(1)	(2)	(3)	(4)	(5)	(6)
Prob (Scoop)	0.029	0.081	0.266	0.268	0.264	0.004
						(0.016)
Prob (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.
*$p<0.1$, **$p<0.05$, ***$p<0.01$.

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.[21]

We access the data directly from the RCSB Custom Report Web Service.[22] 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).

Additional data on cluster entities was accessed through a separate raw file archive at RCSB[23] 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.[24] 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

[21] http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction

[22] https://www.rcsb.org/pdb/results/reportField.do

[23] ftp://resources.rcsb.org/sequence/clusters/ clusters50.txt and clusters70.txt

[24] Both 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, 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.[25] 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. Another large publisher did not have accessible data on their website, and they were willing to share the data with us anonymously. 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 A1 reports the correspondence between the PDB release date and the publication date for the 750 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

[25] https://help.altmetric.com/support/solutions/articles/6000190631-using-altmetric-data-for-altmetrics-research

after one year. If a team takes more than one year to publish results after the deposit, they would be 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, 52 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 122 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 81 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.[26] 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

Some of our analysis relies on using panel data of scientist productivity, including control variables for past productivity and long-run outcomes in the years following a scoop event. The PDB does not

[26] https://www.topuniversities.com/university-rankings/university-subject-rankings/2018/life-sciences-medicine

explicitly link authors between deposits, and neither PubMed or Web of Science have author identifiers across publications. Therefore we manually create a panel of author publications by matching last names and initials across projects and papers. 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.

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.[27] This choice improves the reliability of our name-matching, but offers less information about academic productivity. For example, observing an author that stops publishing PDB-linked papers may indicate that they have shifted their focus to a related field rather than dropped out of academia altogether.

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.[28] 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

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. In all regressions using long-run outcomes, we control for a dummy variable for the common names to help account for potential measurement error of the left-hand side variables.

Using these linked panels of data, we construct controls and outcome variables, such as the number of publications in the five years prior to a scoop event, or the number of citation-weighted publications

[27] Large-scale, data-driven name disambiguation efforts of the PubMed archive have been undertaken (Torvik et al. 2005; Torvik and Smalheiser 2009). They incorporate additional information about co-author networks and research topics to infer matches. Unfortunately, an updated version of their linked author data is not yet publicly available.

[28] Changes 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.

in the five years after the scoop event. 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.

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 work with the data 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, etc. 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 clear. 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.[29]

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 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.

[29] In 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.

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:

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

using the fact that $(u_2 - u_1) \sim N\left(0, 2\sigma_u^2\right)$ and $f, \gamma > 0$. 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 $\Phi\left(\frac{f}{\gamma\sqrt{2}\sigma_u}\right) > \frac{1}{2}$.

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:

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

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

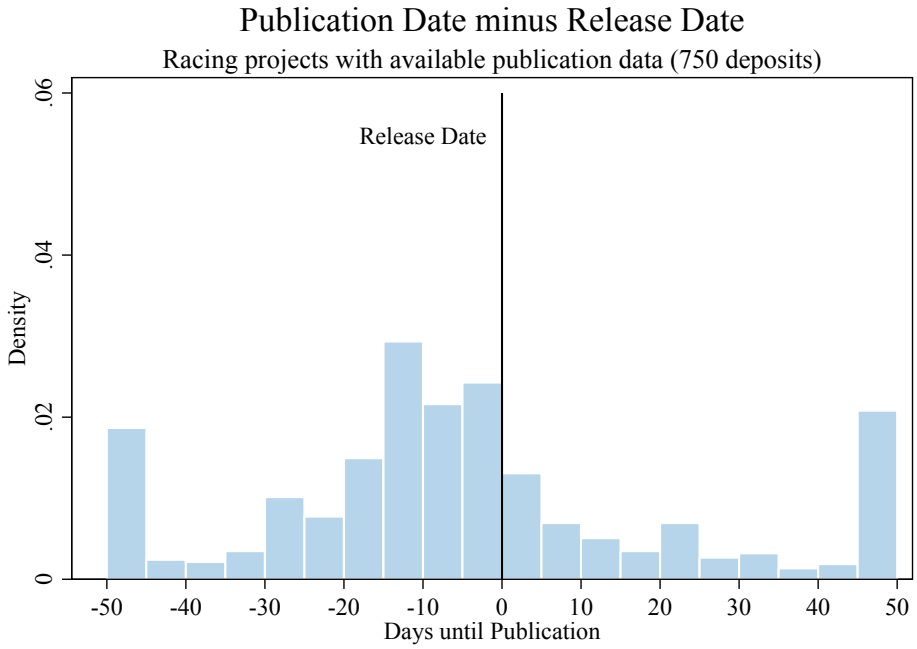
that L_1 is cited is:

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

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

E Appendix Figures and Tables

Figure A1: 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
University rank bins:	
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 Three-Year Productivity

	Total count three years after race			
Dependent variable	Publications (1)	Top-ten publications (2)	Citation-weighted publications (3)	Top-10% cited publications (4)
Panel A. All scientists				
Scooped	0.101	0.035	-0.145***	-0.111***
	(0.143)	(0.064)	(0.049)	(0.035)
Winner Y mean	8.706	2.925	124.312	1.017
Observations	14,768	14,768	11,075	11,075
Panel B. Novices				
Scooped	-0.032	-0.035	-0.277***	-0.104***
	(0.107)	(0.043)	(0.103)	(0.026)
Winner Y mean	1.545	0.538	31.174	0.266
Observations	3,926	3,926	3,241	3,241
Panel C. Veterans				
Scooped	0.140	0.079	-0.132***	-0.116**
	(0.199)	(0.086)	(0.049)	(0.047)
Winner Y mean	12.087	4.053	173.221	1.412
Observations	10,019	10,019	7,225	7,225

Notes: This table presents regression estimates of the three-year long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 3 dependent variable is the total citations accrued in three years to all papers published in the three years after the race. 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 inverse hyperbolic sine function (winner Y means reported in level citations). Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than two 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 A3: Effect of Getting Scooped on Ten-Year Productivity

	Total count ten years after race			
Dependent variable	Publications (1)	Top-ten publications (2)	Citation-weighted publications (3)	Top-10% cited publications (4)
Panel A. All scientists				
Scooped	0.793	0.134	-0.050	-0.004
	(0.675)	(0.251)	(0.087)	(0.145)
Winner Y mean	24.945	7.982	318.123	2.432
Observations	7,673	7,673	4,384	4,384
Panel B. Novices				
Scooped	0.063	-0.016	0.034	-0.044
	(0.448)	(0.147)	(0.148)	(0.082)
Winner Y mean	5.382	1.713	74.943	0.583
Observations	2,494	2,494	1,624	1,624
Panel C. Veterans				
Scooped	0.893	0.300	-0.051	0.083
	(1.002)	(0.373)	(0.079)	(0.217)
Winner Y mean	36.507	11.700	491.753	3.757
Observations	4,769	4,769	2,520	2,520

Notes: This table presents regression estimates of the ten-year long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 3 dependent variable is the total citations accrued in three years to all papers published in the ten years after the race. 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 inverse hyperbolic sine function (winner Y means reported in level citations). Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than two 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.
*$p<0.1$, **$p<0.05$, ***$p<0.01$.