Scooped! Estimating Rewards for Priority in Science*

Ryan Hill† Carolyn Stein‡

September 5, 2024

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 publication and career outcomes. To do so, we analyze data from structural biology where the nature of the scientific process together with the Protein Data Bank enables us to identify priority races and their outcomes. We find that scooped teams are less likely to publish in top journals and receive 21 percent fewer citations. We further study the implications of priority racing on research strategy, academic inequality, and scientist beliefs.

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

†Northwestern University, ryan.hill@kellogg.northwestern.edu. Corresponding author.

‡University of California, Berkeley, carolyn_stein@berkeley.edu. Both authors contributed equally.

1 Introduction

> "In short, property rights in science become whittled down to just this one: the recognition by others of the scientist's distinctive part in having brought the result into being."
>
> – Robert K. Merton, *Priorities in Scientific Discovery: A Chapter in the Sociology of Science (1957)*

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

Therefore, the contribution of this paper is to empirically measure this priority premium. We analyze the impact of getting scooped on the losing project (in terms of probability of publication, journal placement, and citations) as well as on the scooped scientist's subsequent career. We also investigate whether competition for academic attention contributes to inequality within scientific disciplines.

Conceptually, our goal is to measure the cost of getting scooped by constructing comparisons in which multiple teams of scientists are working independently and concurrently on an identical or very similar project. In practice, these races are challenging to identify for three reasons. First, many academic fields use a variety of methods and seek to answer fairly open-ended questions, and so finding near-identical projects is difficult. Second, even if the questions are well-defined, it is difficult — especially without expertise in a given scientific field — to quantify the intellectual distance between two papers in topic space. Third, scooped projects are often abandoned, making them impossible to track in publication data. We tackle these challenges by analyzing project-level data from the field of structural biology. Specifically, we examine projects in the Protein Data Bank (PDB), a repository for structural coordinates of biological macromolecules. The PDB is a centralized, curated, and searchable database of biological details contributed by the worldwide research community, and contains over 150,000 macromolecule structures.[1] Several features of the PDB allow us to make headway on the key empirical challenges described above. First, structural biology papers have a well-defined objective, which is to describe the three-dimensional shape of a known protein. Once the first paper about a protein structure is published, any follow-up publications serve mostly to confirm the result of the first. Second, projects are grouped by the PDB according to molecular similarity, which allows

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

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,[2] 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 her findings to the PDB. Scientists typically deposit their findings shortly after a manuscript has been submitted for publication. The second is the release date, which closely corresponds to the date of publication and is usually 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 1,611 races in our data. These races consist of 3,279 separate projects out of 67,297 total projects in our sample period from 1999 to 2017, suggesting that five percent of all structural biology projects are involved in a late-stage race to publication. These races are composed of a diverse set of scientific teams from different countries, institutional prestige, and experience. In the main analysis of this paper, our definition of scooped projects focuses only on late-stage races where both teams are on the cusp of publication. Focusing primarily on these late-stage scoops is advantageous for the economic interpretation of our results. Since both projects had been completed independently prior to publication, we can infer that the second-place team *would have* published the priority paper in the counterfactual where they had not been scooped. The estimated difference in observed outcomes therefore isolates the premium for novelty awarded by editors and readers. The downside of focusing on these narrow post-deposit scoops is that the scientists are passive at this point. The research has been largely completed and the timing of release is in many ways out of their hands, so these races offer little insight into the strategic interactions between racing teams, a central topic in the economics of R&D racing. Therefore, as an extension in Section 5, we study a sample of teams that were scooped after they had begun their experiments, but before they had deposited their final project, in order to learn more about these strategic interactions.

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

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

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 2.6 percent less likely to be published. Scooped papers appear in a 0.19 standard deviation lower-ranked journal, and are nearly 20 percent less likely to appear in a top-10 journal. Scooped papers receive 21 percent fewer citations, and are 24 percent less likely to be a "hit" paper, defined as reaching the top 10 percent in citations for 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 44 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 four percent difference in citations once we control for journal fixed effects. We provide suggestive evidence that editors and reviewers have a strong taste for novelty. Papers that are scooped prior to submission to a top journal are rarely, if ever, accepted for publication. Some scooped papers do appear in top journals, but only if they were far along in the review process on the date they are scooped.

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

The main analysis focuses only on scoops where the losing team had already deposited and was therefore limited in its opportunity to change its research. When considering cases of pre-deposit scoops (i.e., scoops that occur *before* the losing team has deposited their work), we find that scientists are able to strategically respond to being scooped by adjusting the scope and direction of their project, and also by integrating insights from the winning publication. We identify this subsample of races using the "collection date" feature of the PDB, which allows us to find teams that had done their initial experiments but had not yet deposited their findings in the PDB. When scooped in this intermediate stage, they take 1.4 years longer from collection to deposit than teams that are scooped after depositing. In that time, they tend to include additional structure deposits in their paper, and shift the focus of their writing away from just describing the structure itself and toward more analysis of protein function. They are also more likely than our main sample of scoops to build on the priority findings using a technology called molecular replacement. Although some of these strategic responses to getting scooped slow the scientists down, they also help offset the growing scoop penalty.

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

Finally, we benchmark the size of the scoop penalty by comparing it to the perceptions of active structural

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

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

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

The remainder of the paper proceeds as follows. Section 1.1 offers a brief literature review. Section 2

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

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, and citations, as well as the long-run career results. We also discuss the role of editors and the timing of races for the distribution of priority rewards. Section 5 studies the strategic response to being scooped in races where the scooped team had not yet completed the project. Section 6 describes a model of academic attention and reports results for heterogeneity of the scoop penalty by pre-existing reputation. Section 7 benchmarks the size of our estimates against the beliefs of surveyed structural biologists about the probability and cost of getting scooped. Section 8 concludes.

1.1 Related Literature

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

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

Theoretical models of patent races have considered how racing affects the amount of R&D investment (Loury 1979; Lee and Wilde 1980) as well as the pace of research and the amount of risk-taking induced by the structure of races (Dasgupta and Stiglitz 1980). Many of these models pre-suppose a winner-take-all reward, which has implications for the outcome of innovation tournaments and the strategic behavior of the participants. The conventional wisdom in the sciences — and the assumption underlying much of the theoretical economics work on the topic — is that the process of scientific discovery is also a winner-take-all tournament, even if the prize is priority recognition rather than a patent. (Merton 1957; Stephan 1996). Dasgupta and David (1994) explain that a discontinuous priority reward might arise in science because of a fundamental verification problem. Because of the public goods nature of new knowledge, a team that tries to publish the second paper cannot credibly prove to the community that they would have successfully completed the project absent the help of the priority paper. Even if it would be socially optimal to share more credit with teams who were working in parallel, these information frictions might make credit-sharing difficult. The discontinuous priority reward structure has implications for the pace of research and the strategic interaction of teams (Bobtcheff et al., 2017). 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). Most related to our work, Thompson and Kuhn (2020) document that winners of patent races do more innovation in the future, and that this innovation is more likely to be related to the original patent. The authors identify patent races by looking for patents that were rejected for lack of novelty. Bikard (2020) studies the

phenomenon of simultaneous discovery in science, and documents many cases of papers that are similar in content, are published around the same time, and are frequently cited together. However, our method of using biological details to link competing papers allows us to find simultaneous discoveries where one paper goes unpublished or is cited infrequently in the future.

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

2 Background and Data Construction

2.1 Scientific Primer: Structural Biology and the Role of Proteins

In this section we provide a primer on the field of structural biology, a setting particularly conducive to studying scientific races. Structural biology is the study of the three-dimensional structure of biological macromolecules. These macromolecules include deoxyribonucleic acid (DNA), ribonucleic acids (RNA), and, most commonly, proteins. Proteins contribute to almost every process inside the body. They transport oxygen in blood (hemoglobin), trigger muscle contractions (actin and myosin), and regulate blood sugar (insulin). In many ways, the form or structure of a protein determines its function. For example, antibodies are Y-shaped immune system proteins that bind to foreign molecules (like viruses or bacteria) with two of their arms, while recruiting other immune system proteins with the remaining arm. It is exactly this Y shape that allows the antibody to function (National Institute of General Medical Sciences 2017). Protein folding and structure has important applications, particularly in medicine, and fifteen Nobel Prizes have been awarded for advances in structural biology (Wlodawer et al. 2008; Martz et al. 2019).

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

2.2 The Protein Data Bank

We focus on structural biology because the Protein Data Bank (PDB) contains detailed, organized, and comprehensive project-level data that is publicly available. The PDB is a worldwide repository of biological macromolecule structures, 95 percent of which are proteins.[5] The PDB was established in 1971 at Brookhaven

[4]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).

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

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).

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

Because of these strict public disclosure policies, we believe the PDB represents a near-complete census of macromolecule structure discoveries. Whenever a structural biologist completes a project, she uploads the structure, experiment, and discovery details to the PDB. This typically happens shortly before or after she submits an academic paper describing her findings for publication. An important feature of this process is that the uploaded data is confidential. No other user of the PDB can access the data or see that the deposit has been created. Even the editor and reviewers only receive a receipt of deposit from the PDB and author, and they do not see the underlying structure data until the date of publication. Only at the point of publication is the data released to the public. If any project goes unpublished, the data is released by default after 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 deposit is assigned a unique ID. For example, PDB ID 4HHB, deposited in 1984 by Giuilio Fermi and co-authors, reports the structure of human deoxyhemoglobin, the form of hemoglobin without oxygen which is the predominant protein in red blood cells (Fermi et al. 1984).

The PDB provides three key pieces of information that we will use in our analysis. The first is a measure of similarity between proteins. This is calculated by comparing how similar a protein's amino acid chain is to other proteins in the PDB. For a given protein, the PDB uses an algorithm to construct a list of other proteins that are 100 percent similar, 90 percent similar, etc., all the way down to 30 percent similar. These groupings, or "clusters," allow us to determine whether two structure deposits from different teams correspond to the same or very similar protein. The second key piece of information the PDB provides is a list of dates for the structure deposit, including when the data was deposited and when it was released. This allows us to construct a timeline for the projects and identify cases when two or more teams were working simultaneously on the same protein. Finally, each PDB structure is linked to the academic paper

that the structure was published in (if any). This link includes the PubMed ID, which we link to PubMed bibliographic data and Web of Science citation data.

2.3 Identifying Priority Races: Challenges and Solutions

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

The PDB enables us to make significant progress on these three obstacles. First, the questions in structural biology are well-defined, because scientists are typically trying to solve the structure of a known protein. Moreover, the methods are consistent: 91 percent of proteins are solved using x-ray crystallography. This means that if we observe two papers that study the structure of the same protein, these two papers are likely to be very similar in terms of the question, methods, and conclusions. Second, as mentioned in Section 2.2, the PDB measures how biologically similar different proteins are to one another. This allows us to link projects based on objective measures of scientific proximity rather than text similarity or citation behavior. Finally, scientists are required to deposit their structures in the PDB *prior* to publication. This gives us the ability to observe some projects that never reach publication. Given that scientists might abandon projects that get scooped, having this record of unpublished projects is a key feature of our data. We will discuss the timeline in more detail in the next section. To the best of our knowledge, we are the first to measure scientific races in a data-driven manner.[6]

2.4 Defining Priority Races

Broadly speaking, we define a priority race as an instance where two or more teams are working on the same protein independently and concurrently and are likely uncertain about the identity or progress of their competitors. Following Brown and Ramaswamy (2007), we define "same protein" as meaning two proteins within the same 50 percent or higher sequence similarity group (called a "cluster" in the PDB). This is a conservative cutoff, as 30 percent has been suggested as sufficient similarity for building homology models (Dessailly et al. 2009; Moult 2005). In other words, the first publication within these 50 percent similarity clusters is often highly cited because it provides a novel structure model that other crystallographers can

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

build on to solve very similar proteins.[7] The PDB assigns ID numbers to clusters of similar proteins, and we say that the first structure released in that cluster is the "priority" structure deposit. There are often many subsequent deposits that report similar structure coordinates as the priority deposit, only some of which we define as being scooped. These follow-on deposits appear for a variety of reasons, including concurrent projects by authors that were racing to be first but were scooped, 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.[8]

We use project timelines reported in the PDB to determine whether a follow-on deposit qualifies as scooped by the priority deposit. The PDB provides two key dates at the structure level that help us determine whether two teams were working concurrently: the deposit date and release date.[9] The deposit date corresponds to the date that the scientist uploaded her solved structure to the PDB. Importantly, the structure is not yet visible to the public. Nearly all scientific journals require that authors upload their structures to the PDB prior to publication, so deposit typically occurs slightly before or after the date that the scientist first 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 are publicly salient events that the community pays attention to, so the release dates are therefore good markers of priority order. Appendix Section A.4 discusses implications and presents evidence about the concordance between release dates and publication dates in greater detail.

Figure 2 illustrates how we define a scoop event. Consider two projects, A and B, authored by two distinct teams working on the same protein. Suppose project A is a priority project in one of the similarity clusters. We say that project A scoops project B if (i) A is released before B is released, but (ii) after B has deposited to the PDB. Condition (i) guarantees that A finishes first, while condition (ii) guarantees that B did not know about A until after the structure was deposited in the PDB. Since B had already deposited a completed structure, they likely would have been the priority deposit had they not been scooped by A. Requiring that B has deposited before A is released ensures that we observe abandoned projects, since all deposited structures appear in our data even if they are scooped and fail to publish. We allow the priority project to scoop more than one team, and 5.6 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. In our main analysis, we exclusively focus on these clean, but narrowly defined scoops that occur after B has already deposited. However, in Section 5 we expand our analysis to include earlier-stage scoops,

[7] Appendix Figures A1 and A2 provide evidence that at each level of similarity above 50 percent, paper pairs in our sample have very similar titles and are have similar rates of citation between the scooped and winning paper. For robustness, we can restrict to scoops by proteins within the same 100 percent cluster, and find similar results which we report in Appendix Table A5. 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.

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

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

that occur before B deposits.

2.4.1 An Example

To help understand our procedure, consider an example outlined in Table 1. The table shows two structures: 4JWS and 3W9C. Both are structures of the Cytochrome P450cam protein complexed with its redox partner, putidaredoxin (Pdx-P450cam complex). This enzyme is involved in metabolism and clearing toxins, such as in the human liver. Figure 3 shows the nearly identical biological assembly models that each team deposited independently and confidentially to the PDB. The scientists at Leiden University (3W9C) collected their data a few months before the scientists at University of California, Irvine (4JWS) (February 3, 2012 versus September 14, 2012). However, by the time of deposit, the UC Irvine team had pulled ahead, depositing one week before the Leiden team (March 27, 2013 versus April 3, 2013). Ultimately, UC Irvine won the priority race, with their structure being released two months before Leiden (June 19, 2013 versus August 21, 2013). Importantly, when Leiden deposited their structure on April 3, 2013, UC Irvine had not yet released their structure. This means that Leiden was likely unaware of their competitor's progress or results when they were preparing their publication and depositing the structure. Comparing the outcomes of the winner (4JWS) and the loser (3W9C), we observe that the winning paper was more successful. It was published in a better journal (*Science*, with an impact factor of 31.5 versus *Journal of Molecular Biology*, with an impact factor of 4.0) and received about 30 percent more citations over the next five years (Tripathi et al. 2013; Hiruma et al. 2013). In this case, the Leiden authors became aware that they were scooped during the manuscript review. In the conclusion of their paper, they write, "While this manuscript was under review, Tripathi et al. published the crystal structure of the Pdx–P450cam complex that was obtained via cross-linking of the two proteins. It is interesting to compare our complex with those reported in that study. Tripathi et al. found a position and orientation of Pdx relative to P450cam that is essentially identical with ours." (Hiruma et al. 2013)[10]

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.[11] Second, we drop projects that are published in a paper that is linked to 15 or more other structures. Among the set of papers included in our final analysis sample, 46 percent are linked to more than one structure, and the average number of structures per paper is 1.9. 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

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

[11] 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.

the two papers were released closer than two weeks apart from each other. We make this restriction to help ensure that the first project has a clear claim of priority and that the order of release is more likely to correspond to the order of publication.[12]

2.5 Additional Data Sources

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

Journal Citation Reports Journal Citation Reports is an annual report published by Clarivate Analytics that evaluates journal influence using a metric called "journal impact factor." Let $Cites^j_{t,t-k}$ 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:

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

In words, the journal impact factor attempts to capture a journal's rolling average citations per article. We standardize the impact factors within a year t to account for the fact that impact factors have been rising over time as the rate of publishing within the life sciences has increased. We also use the journal impact factor to create a list of "top-10 journals." In order to focus on journals that are both high impact and also relevant to structural biology, we restrict to a potential 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.[13]

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

We construct career histories of variables before and after the priority date of each race to serve as control variables and long-run outcomes. Reconstructing publication records for individual authors is difficult because names are not disambiguated in the PubMed or PDB. We use a dataset called Author-ity, which groups PubMed IDs into distinct author identifiers using co-author and topic patterns (Torvik et al. 2005;

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

[13] 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.*

Torvik and Smalheiser 2009). However, because not all PDB deposits are published, it is hard to link unpublished deposits to the correct name identity in Author-ity. Therefore, in the long-run results section, we restrict to a subset of authors that have uncommon names and uniquely match to an individual in Author-ity. We also use simple name-matching techniques within the PDB to construct control variables of team productivity prior to treatment, which we can do for all deposits including those that are not published. We describe the name disambiguation procedures in detail in Appendix A.6.

For long-run outcomes, we count PubMed publications, PDB-linked publications, top-10 publications, citation-weighted publications, and "hit" publications for the years following the treatment date. Besides analyzing the effects of race outcomes on the intensive margin of publication, we also consider the extensive margin of exit from publishing PubMed papers and PDB-linked papers altogether.

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

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

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 under five percent of the structures in our sample are involved in a priority race. We look at both team characteristics and deposit outcomes. Teams involved in priority races tend to be smaller, younger, and more

likely to come from a top university. The racing scientists were also more likely to work in Asia, and less likely in North America. The deposit outcomes suggest that proteins involved in priority races are scientifically more important. Proteins in the racing sample are more likely to be published, appear in higher-ranked journals, and receive more citations.

3 Empirical Design

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

3.1 Baseline Specification

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

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

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

In Section 4.2, we consider the long-run effect of getting scooped on academic career outcomes. The regression specification is similar to equation 2, but the unit of observation is a scientist, rather than a project. For scientist s who co-authored deposit i that was in a priority race over protein p, we estimate

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

where $Scooped_{isp}$ is a dummy equal to one if scientist s was scooped on project i. $\mathbf{X_{isp}}$ is a vector of scientist-project covariates, such as the number of publications accumulated by scientist s in the five years before the priority date associated with project i. We also include cubic controls for career age, which is defined

[14] Covariates include all variables listed in Table 2, excluding resolution and R-free. Variables in Panel A are included for both first and last author. We also control for variables in Panel B and C calculated over the full career (in addition to the counts calculated over 5 years). Lastly, we control for indicators that tag first and last authors that have common last names as defined in Appendix A.6.

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

as the number of years since the author's first publication in the PDB, as well as the university rank of the first author affiliation and the continent where the first author is located. Again, γ_p is a protein fixed effect. The long-run outcomes are calculated as the sum of each outcome in the five years following the priority date. Importantly, we exclude the publication that is linked to the structure ID of the PDB projects that were involved in the race. These outcomes therefore represent productivity in other projects not including the winning or losing paper in each race. Although each scientist may win or lose races multiple times, we include each appearance as a separate treatment event, and consider the subsequent outcomes for all scoop events.

3.2 Identification and Balance

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

On the other hand, skill, experience, or resources could provide an advantage to certain teams that would allow them to systematically start earlier or work faster and therefore win priority races. This is a threat to identification because these characteristics may simultaneously increase the probability of winning and improve project outcomes. For example, suppose a technological breakthrough marks the starting point of a race that many diverse teams enter. If one team from Harvard has exceptional resources to adopt the technology and complete the project first, we will observe them win the race and receive many citations. But since Harvard is a high-reputation university and has a track record of success, they would likely have high citations even in the counterfactual where their competitor won the race. Therefore, we rely on the assumption that well-resourced or otherwise high-reputation teams are not able to systematically win priority races, and we test this using observable characteristics of each team.

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

Table 3 shows the mean values of each covariate for the winning and losing teams, as well as for the teams in the non-racing sample, for reference. We report test statistics for the difference in means between the

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

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

Unbalanced covariates lead to biased estimates only if they are systematically correlated with the outcome variable. Therefore, to further assess potential selection bias, we visually inspect the difference in expected citations between winners and losers. We estimate a paper-level model of citations using a Lasso[17] regression of three-year citation counts on the battery of team covariates. This model is estimated only in the sample of non-racing deposits. We then take the selected variables and estimated coefficients to predict citations in the racing sample in a post-Lasso OLS procedure. The covariates we include are counts of publications, citations, and journal placements in the five years prior to the deposit for the first and last author, as well as the squares of these variables. We also 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 small. The difference in means between the two distributions is 0.68 predicted citations with a p-value of 0.065 (for reference, the sample average is about 12 citations, so this represents a six percent difference). This slight lack of balance motivates our control strategy discussed in the next section.

3.3 Control Selection Using Post-double-selection Lasso

In light of potential treatment imbalance, we rely on an identification assumption that treatment is exogenous conditional on observable control variables. There are many potential control variables in our data, so we use a method called post-double-selection Lasso (PDS-Lasso) proposed by Belloni et al. (2014) to optimally

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

select controls variables. Consider a partially linear model similar to equation 2

$$Y_{ip} = \alpha + \beta Scooped_{ip} + \mathbf{g}(\mathbf{Z_{ip}}) + \gamma_p + \epsilon_{ip} \tag{4}$$

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

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

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,[18] and (5) an indicator for becoming one of the top 10 percent of publications measured by five-year citation counts. Not all projects are published, and if they are, they may not be published in a ranked journal. We count unpublished papers as having zero citations. If the project is not published in a ranked journal, we impute the impact factor of their publications as being equivalent to the minimum journal ranking in the regression sample. The sample is restricted in columns 4 and 5 to projects released before 2014 to allow a full five years of data coverage to count citations in that window before our citation data ends in 2018. We present regression results from three different specifications. Panel A shows the results from a simplified version of equation 2 with no control variables. Panel B adds all controls listed in Table 3, and panel C uses controls selected from the PDS-Lasso procedure described in Section 3.3. The results across all five outcomes suggest that covariates have very little impact on the coefficients between panel A and panel C, assuaging concerns about omitted variables bias. We will use panel C as the preferred specification to report our estimates throughout the paper. To further test for selection bias on unobservables, we implement a robustness check following Oster (2019) in Appendix Table A2.[19]

[18] 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).

[19] Adding controls and protein fixed effects increases the R^2 from less than 0.01 to over 0.60 in all regressions, suggesting that most of the variance in the outcome is explained by treatment and observable controls. Implementing the suggested bias adjustment, we conservatively assume a maximum $R^2 = 1$ and $\delta = 1$ (unobservables are equally important for treatment selection as observables), and find that the adjusted coefficients are almost identical to our baseline findings. Further, the δ

Scooped projects are 2.6 percentage points less likely to be published off of a baseline publication rate for winning projects of 88 percent. This represents a three percent decrease in probability of publishing, or framed differently, a 20 percent increase in the probability of abandoning the project. This modest discouragement rate is likely driven by the low cost of publishing once the project has already been deposited in the PDB (recall that in our sample, all scooped projects have already been deposited in the PDB when they learn that they have been scooped). In many cases, the scooped teams may be well into their submission and revision process at the time of being scooped, and therefore will persist to publication. Even if they are rejected from a journal, there are many lower-ranked outlets that may be more willing to accept scooped papers, a mechanism we explore in Section 4.3.

In column 2, we estimate a statistically significant penalty in journal impact factor. Scooped papers are published in journals with impact factors 0.19 standard deviations below winning papers. In column 3, this translates to a six percentage point (20 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 21 percent fewer citations in the same time span. Column 5 suggests that this means scooped projects are 3.6 percentage points (24 percent) less likely to be one of the top 10 percent of papers in that publication year ranked by 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 (see Appendix Table A3). Appendix Table A4 shows results are robust to the exclusion of protein (i.e., race) fixed effects. As a further robustness check, we reproduce the regressions using a sub-sample of races that have projects with 100 percent similar sequence structure according to the algorithm used by the PDB. Appendix Table A5 shows that the magnitudes are very similar for all outcomes, even if statistical precision is lower due to the smaller sample size.

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

Taken together, these results suggest that there is a significant penalty for being scooped, both in the likelihood of publication, the journal rank of publication, and the number of citations accrued in the early life cycle. However, these results also indicate that the rewards for priority are not winner-take-all. Losing teams receive a smaller, but still substantial share of the credit as measured by publication and citations. Translating the citation penalty to shares of total citations, losing projects receive approximately 44 percent of the total citations accrued to both papers, a much larger share of credit than zero percent for the winner as is typically assumed by classic models of innovation races.[20]

4.2 Long-run Effect on Authors

In this section we analyze the long-run consequences of being scooped on the careers of the various authors of scooped papers following equation 3. Table 5 reports the results of the long-run outcomes regression. Panel A contains results for regressions in the full sample of authors. Panel B restricts to novices only, which are

needed to reduce the estimate to zero ranges from 8 to 60 across all specifications, meaning there would need to be an unrealistic degree of selection on unobservables to threaten the robustness of the results.

[20] The estimated share of 44 percent is calculated by dividing the mean citations of the losing teams, $28.8 * (1 - 0.208)$ by the implied total citations $(28.8 + 28.8 * (1 - .208))$ based on the estimate of the percent citation penalty from column 4, panel C.

defined as authors who had seven years or less since their first publication at the time of the scoop event.[21] Panel C restricts to veterans, which are all scientists not defined as novices.[22]

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

Despite a significant extensive margin effect, we find no significant changes to publication counts on the intensive margin for novices or veterans. Losing teams have no statistically significant differences in publications or PDB-linked publications in the following years as shown in column 3 and 4, and they are not more or less likely to publish in top-10 journals. This difference in intensive and extensive margin effects might mirror a similar dynamic documented by Wang, Jones and Wang (2019), where scientists that persevere through setbacks (in their case being denied a grant), do not experience negative productivity effects in the long run (perhaps due to grit or psychological persistence). However, we do estimate significant penalties in citations for all categories of authors. In the full author sample, the scooped individuals receive 20 percent fewer citations (measured by inverse hyperbolic sine citation-weighted publications) in the next five years, where citations are counted up to three years after each paper's publication. This effect falls particularly hard on novices, who receive 34 percent fewer citations, while veterans receive only 16 percent fewer citations. The effect on "hit" papers is reported in column 7 and also suggests that getting scooped decreases attention to future work. The full sample of scientists publish 0.59 fewer hit papers in the five years following a scoop event. The negative effect is lower for novices in levels (0.18 papers versus 0.82 papers for veterans), and not statistically significant for novices. However, if we scale the effect size by the average number of hit papers, the effect is larger for novices (a 16 percent decline versus an 8 percent decline). We also consider outcomes in the following three years in Appendix Table A7 and ten years in Appendix Table A8. 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. Lastly in Appendix Table A9, we restrict to first, middle, and last authors separately because first and last authors are considered to have a larger reputation stake in life science papers, but we find broadly similar effects for all types of authors.

4.3 Mechanisms: Role of Scoop Timing in the Publication Process

Scooped projects receive about 21 percent fewer citations than their winning counterparts, suggesting that academic researchers pay less attention to the projects that are scooped. In this section, we investigate how the editorial process affects the scoop penalty, and we argue that journal placement is a primary driver of the citation penalty. Further, the size of the penalty is highly correlated with the timing of races. Teams that are scooped early (very shortly after they deposit their findings) receive a much larger penalty than

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

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

teams that are scooped late (shortly before publication). We provide evidence that top journal editors are unlikely to accept scooped papers, therefore scooped papers consistently fall to lower-ranked journals unless they were deep into the review process at the time they were scooped. These results suggest that editors and reviewers are key policymakers in determining the distribution of academic credit for novel research.

4.3.1 Decomposing the Citation Effect by Journal

First we show that the citation penalty is largely driven by journal placement. We decompose the citation effect into an editor/reviewer effect and a reader effect by controlling for journal placement. Column 1 of Table 6 replicates the citation penalty effect from 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 10 percent, but remains statistically significant. Finally, in column 4 we include journal fixed effects to control completely for any direct effect of the publication outlet on citations. The effect falls to four percent. These results suggest that nearly three fourths of the citation penalty comes through the channel of the publishing journal. Any remaining effect on citation attention comes through readers differentially citing winning and losing papers in similar journals.

4.3.2 Editors' Role in Priority Credit

We further explore the role of editors in adjudicating priority credit by focusing on the submission, review, and publication timelines of scooped projects submitted to leading science journals. Academic journals compete fiercely to publish the highest quality and most novel scientific articles. Many of these journals have explicit policies for accepting only highly original and novel research. For example, *Science* provides the following guidelines to peer reviewers: "[R]ecommend in your review whether the paper should be published in *Science* and provide a more detailed critique based on the following: ... Novelty: Indicate in your review if the conclusions are novel or are too similar to work already published."[23] Editors and reviewers therefore likely drive much of the scoop penalty if they choose to reject scooped papers when they come across their desk. In this section we look at how the scoop penalty is affected by the timing of journal submissions. Many of the papers in our sample had already been submitted to a journal when they were scooped, and a few papers had already been accepted. Even if an editor would prefer to reject a scooped paper, they may be unable to do so if the paper had already been accepted or was far along in the review process. We use the supplementary data collected from journal websites to examine how the scoop penalty is affected by the timing of the review process. Ideally, we would compare the scoop date to rejection dates at leading journals. But data on rejected papers is not publicly available. Therefore, we instead use the timing of submission and acceptance to present suggestive evidence that editors at top journals are reticent to publish scooped papers.

In our data, scooped papers occasionally appear in top journals like *Science, Nature,* and *Cell*, but 90 percent of those papers were already under review on the date that they were scooped. Furthermore, about 60 percent of those papers were scooped after they had already been accepted. Figure 5 further shows that this pattern varies greatly by the impact factor of the journal that eventually publishes the scooped paper. For lower ranked journals, such as *PLOS One*, only 60 percent of scooped papers had been received by the journal on the date they were scooped, and just over 20 percent had been accepted. Among the

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

11 large journals for which we have information about received and accepted dates, there is a positive and statistically significant relationship between the share accepted before the scoop date and the impact factor, with a one standard deviation higher ranked journal being eight percentage points more likely to have already been accepted on the scoop date. Although we cannot directly observe scooped papers being rejected from these journals, we can infer from this pattern that top journals are less willing to accept papers that were scooped before submission or early in the review process. Many of these scooped papers fall to lower ranked general interest journals or highly specialized structural biology journals.[24] 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 only offered to provide the data anonymously, stating their concern about presenting public evidence that they publish scooped papers.

4.3.3 Time Lag and the Scoop Penalty

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

[24] One possible strategy a team might consider to win a race is to submit to a lower ranked journal that has a faster average review time. Indeed we find that top-ranked journals take about 120 days on average from submission to acceptance while lower ranked journals take about 90 days on average. However, as the results in Table 6 show, the bulk of the scoop penalty is due to journal placement, suggesting that the citation-maximizing strategy is to submit to the best possible journal first, despite the potential for a slightly longer review.

5 Strategic Responses to Getting Scooped Before Project Completion

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

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

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

In the remainder of this section, we study the strategic decisions of a scientist who is scooped early enough in the project's life that she still has an opportunity to re-optimize the path of the project. The key idea is that once a scientist learns that she has been scooped, she faces a tradeoff. She knows that on one hand, she will get more credit if she publishes quickly because the scoop penalty grows with time (as shown in Figure 6). On the other hand, she can expand the project in new directions (for example, by adding additional structures or experiments). This will take time — leading to a larger penalty — but will also make the project more valuable overall. Moreover, because she can now take advantage of informational spillovers from the first paper, it might be easier to expand the project than before that paper was released. We formalize this tradeoff in Appendix C. Broadly speaking, there are three possible cases. In one case, the scientist speeds up when she learns that she has been scooped to minimize the penalty. In a second case, she slows down and improves or broadens her project to maximize its value. Finally, it is possible that the

cost of completing a project is no longer offset by the reward, leading to a third case where she abandons the project upon learning it has been scooped.

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

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

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

This delay in release appears to be consistent with scientists electing to add additional experiments and differentiate their project from the race winner. Table 8 presents regression results using the full sample of races associated with 1,778 pre-deposit scoops and 979 post-deposit scoops combined for which we have available data.[25] We regress a series of project characteristics that relate to the margins of adjustment discussed above on a scooped indicator and an interaction between a pre-deposit indicator and the scooped indicator.[26] In column 1, we can see that there is a very large increase in maturation time (time between collection and release) for the pre-deposit scooped teams relative to the post-deposit scooped teams, with the pre-deposit teams spending 1.4 more years on average. Next, we consider how trailing scientists may adjust the scale or scope of their research to offset the scoop penalty. We find in columns 2 and 3 that pre-deposit scooped teams are much more likely to include multiple protein structures in their paper relative to the post-deposit scooped teams, suggesting that pre-deposit scooped teams expand the scope of their papers. Next, we look at how scooped teams may have adjusted the content of their paper by analyzing keywords from the paper titles. Column 4 suggests that pre-deposit scooped teams are much less likely to use the word "structure" or "structural" than post-deposit scooped teams. However, as seen in column 5, they are *more* likely to use words like "function", "mechanism", or "analysis" in the title. It appears that if teams have the flexibility to adjust the direction of their research after being scooped, they choose to shift the focus away from the structure determination itself and toward describing the function or biological mechanisms that the

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

[26]The pre-deposit main effect is absorbed by the protein fixed effect.

structure implies.

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

Finally, we compare the cost of being scooped before and after deposit. We interpret the results of this exercise cautiously because of the endogenous selection into the pre-deposit sample and the additional flexibility that pre-deposit scooped teams have to strategically respond to the scoop. Appendix Table A10 reproduces Table 4 in the pre-deposit sample. We find that the difference in most outcomes between the winners and losers is about 15 to 40 percent larger in pre-deposit scoops compared to our primary post-deposit sample. The citation gap is 28 percent in the pre-deposit scoops compared to 21 percent in the post-deposit (main sample) scoops. The relative reduction in the probability of publication is comparable between the two groups. Scientists who persist despite being (knowingly) scooped are likely a selected set who are determined to publish.

6 Reputation and the Scoop Penalty

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

Priority rewards are allocated by a decentralized set of actors, including journal editors and readers, in a market for academic attention. Because scientists have limited time for reading and reviewing new papers, it may be difficult to determine the quality of new research. Therefore, editors and readers may rely on signals of ability based on the reputation of the researchers or their institution to supplement their judgement of a paper's quality. The model considers cases where two types of teams, high- and low-reputation, publish

identical papers. Readers decide who to cite based on priority and reputation. In cases where teams are of the same type, the priority effect is isolated, and the first team to publish receives more than 50 percent of the total citations. However, in cases where teams are of different types, the priority and reputation effects will either work in the same or opposite direction, depending on which team finishes first. If the high-reputation team wins the race, the two effects reinforce each other, meaning the high ranked team will have an equal or greater share of citations compared to the low-ranked team than they would competing against another high-reputation team. If the low-reputation team scoops the high-reputation team, the net effect is ambiguous. If the reputation effect is stronger than the priority effect, the low-reputation team may receive less than 50 percent of the total citations, despite publishing first.

To test our model, we measure the share of total citations received by winning and losing labs, and compare these shares in races where the reputation varies between the two racing teams. More specifically, if lab A and lab B race to write a paper about the same protein, we compute $CitationShare_A = Citations_A/(Citations_A + Citations_B)$. This citation share maps to the probability of citation outlined in the model above.[27]

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 3.2. Labs with above-median predicted citations are categorized as H labs, while teams below median are called the L labs. In Figure 8 we plot the predicted citations of the losers on the x-axis and the predicted citations of the corresponding winners on the y-axis. Each point on this scatter plot represents the observed match between two racing labs. If all labs were equally matched in 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 8 conveniently partition the sample into four sub-samples that line up with the four types of "matchups" we discuss in our model. The top right and bottom left corners represent subsamples of closely matched races where both labs were either high-reputation or both low-reputation. The top-left and bottom-right subsamples represent mismatched races where an above-median team scooped a below-median team and vice versa.

In mismatched races, we interpret the difference between citations as being caused by an additive effect of priority and reputation. One potential confounder in that interpretation is that high- and low-reputation teams might produce different quality of scientific outputs for the same structure discovery. If H teams produce higher quality or more convincing results, then the additional citations they receive may not only be caused by their high-profile reputation. Although it is difficult to quantify all aspects of paper quality, we examine two important measures of quality reported by the PDB: resolution and R-Free (goodness-of-fit), described in more detail in Section 3.2. Appendix Table A11 compares the average resolution and R-Free of the winning and losing structures in each of the four subsets of races. We find very little evidence of statistical difference in quality metrics between H and L teams engaged in a race. This suggests that any difference in citations is not driven by the quality of science that each team is producing.

Figure 9 shows the average citation counts by matchup type, as well as the citation shares. Panel A shows the evenly matched races, which isolates the priority effect. As predicted by the model, the winning labs receive more citations. Moreover, if we look at the *share* received by the winning team, we see that it is identical in the H versus H matchups and the L versus L matchups (winning team receives about 55

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

percent of the total citations), which is consistent with the model.[28]

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 (65 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.

7 Benchmarking Magnitudes: Survey Results

We estimate that getting scooped causes a decrease in the probability of publication, leads to publication in lower-impact journals, and reduces citations. However, priority races are not winner-take-all. Our citation estimate suggests that winners get 56 percent of the total citations, a far cry from 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 D. First we ask, "Suppose you have just completed a very promising research project...what do you think is the probability that your project will be scooped between now and when it is published?" We next state that their hypothetical project has indeed been scooped by a paper in the journal *Science*. In this scenario, we ask them the following questions: "Would you choose to abandon your manuscript? Assuming you submit, what is the probability the article will eventually be published? What is the best journal that would accept your paper? If your competitor receives 100 citations, how many citations do you expect your publication to receive?"

Table 8 reports the average responses of the biologists in columns 3 to 6 and compares them to the magnitudes estimated in the PDB data in columns 1 and 2. The hypothetical scenario in the survey was designed to match the instances of racing that we have in our data. However, because we tried to pose the survey questions as concretely as possible for clarity, the racing situation does not exactly match the average situation in the PDB. In particular, in the survey the losing team is scooped early in the submission process, and the project is very high-quality, with an expected journal placement in *Science*. Therefore we report estimates in column 2 from a subset of the PDB data where (1) the losing team is scooped soon after they deposit their data,[29] and (2) one of the teams published in one of the three highest impact journals (*Science, Nature,* or *Cell*). These restrictions make some of the PDB estimates smaller or larger, but we still consistently find evidence of pessimism among respondents. Surveyed scientists report a 27 percent chance of being scooped between submission and publication, more than three times the eight percent scoop

[28] 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.

[29] 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.

probability in the comparable PDB sample.[30] Six percent of respondents report that they would abandon the project, but only 68 percent think they would succeed at publishing conditional on submitting, implying a 67 percent unconditional probability of publishing as shown in column 3. This is much lower than the 85 percent of scooped papers that are actually published in the PDB data, and the 98 percent that are published in the comparable subsample. Scientists are very pessimistic about the potential journal placement of scooped papers, expecting that the best journal they could publish in would be almost three standard deviations below *Science*, which has a standardized impact factor of about three in most years. Finally, we ask about expected citation effects. When asked to guess the number of citations they would receive compared to the hypothetical winner's 100 citations, the average guess was only 41 citations, which translates to a 59 percent penalty, or a share of 29 percent of the total citations. The corresponding estimate in the PDB is no more than a 21 percent penalty or a 44 percent share. Ultimately, PDB scientists expect much worse consequences from being scooped than can be found in the data.

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

8 Conclusion

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

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

biology. We find that rewards are far from winner-take-all; rather, our preferred estimates suggest a 56-44 split in citations between the winning and losing paper.

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

In this paper, we establish that priority is a relevant incentive in science, but we do not analyze the overall welfare implications of the priority system and size of the priority premium, nor do we consider alternative systems or policies. How would a larger or smaller priority premium affect the efficiency of science? There are many margins to consider, including how it would affect effort, project selection, collaboration, and even participation in science. A particularly interesting concern raised in popular and academic writing is that priority may be pursued at the expense of quality. Racing to complete projects may stimulate effort and hasten the pace of discovery, but it may lead scientists to cut corners on the quality of the results that they disclose. If the incentives for replication are low and the costs of replication are high, science as a whole may suffer as quick and sloppy research becomes the norm. In Hill and Stein (2024), we analyze objective measures of the quality of crystal diffraction data and corresponding structure models to study how racing in science affects quality outcomes. We find that proteins with high ex-ante potential have more competitors racing to complete the structure, are deposited faster, and are completed with lower quality. This evidence suggests that racing in science does indeed hasten disclosure, but has negative effects on quality. Concerns about the cutthroat nature of racing have led to suggestions of policies that might dampen the strong incentives for novelty. These include allowing a grace period for journal acceptance 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.

Aoki, Reiko, "R&D Competition for Product Innovation: An Endless Race," *American Economic Review*, 1991, *81* (2), 252–256.

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

Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti, "Researcher's Dilemma," *The Review of Economic Studies*, 2017, *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*, 2020, *102* (1), 195–217.

Cengiz, Doruk, Arindrajit Dube, Atilla Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *Quarterly Journal of Economics*, 2019, *134* (3).

Dasgupta, Partha and Joseph Stiglitz, "Uncertainty, Industrial Structure, and the Speed of R&D," *The Bell Journal of Economics*, Spring 1980, *11* (1), 1–28.

_ and Paul A. David, "Toward a New Economics of Science," *Research Policy*, 1994, *23*, 487–521.

Dessailly, Benoît H, Rajesh Nair, Lukasz Jaroszewski, J Eduardo Fajardo, Andrei Kouranov, David Lee, Andras Fiser, Adam Godzik, Burkhard Rost, and Christine Orengo, "PSI-2: structural genomics to cover protein domain family space," *Structure*, 2009, *17* (6), 869–881.

Doraszelski, Ulrich, "An R&D Race with Knowledge Accumulation," *RAND Journal of Economics*, 2003, *34* (1), 20–42.

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.

Gilbert, Richard and David M. Newbery, "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, 1982, *72* (3), 514–526.

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.

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

_ and Carolyn Stein, "Race to the Bottom: Competition and Quality in Science," *Working Paper*, 2024.

_ , _ , and Heidi Williams, "Internalizing externalities: Designing effective data policies," in "AEA Papers and Proceedings," Vol. 110 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2020, pp. 49–54.

Hiruma, Yoshitaka, Mathias AS Hass, Yuki Kikui, Wei-Min Liu, Betül Ölmez, Simon P Skinner, Anneloes Blok, Alexander Kloosterman, Hiroyasu Koteishi, Frank Löhr et al., "The structure of the cytochrome P450cam–putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography," *Journal of molecular biology*, 2013, *425* (22), 4353–4365.

Hopenhayn, Hugo and Francesco Squintani, "On the Direction of Innovation," *Journal of Political Economy*, 2021, *129* (7).

Horner, Johannes, "A Perpetual Race to Stay Ahead," *Review of Economic Studies*, 2004, *71*, 1065–1088.

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 and Low-Wage Employment: Evidence from Seattle," *American Economic Journal: Economic Policy*, 2022, *14* (2).

Judd, Kenneth L., "Closed-Loop Equilibrium in a Multi-Stage Innovation Race," *Working Paper*, 1985.

Kim, Soomi, "Shortcuts to Innovation: The Use of Analogies in Knowledge Production," *Working Paper*, 2023.

Klebel, Thomas, Stefan Reichmann, Jessica Polka, Gary McDowell, Naomi Penfold, Samantha Hindle, and Tony Ross-Hellauer, "Peer review and preprint policies are unclear at most major journals," *PLoS One*, 2020, *15* (10), e0239518.

Lee, Tom and Louis L. Wilde, "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, *94* (2), 429–436.

Lerner, Josh, "An Empirical Exploration of a Technology Race," *RAND Journal of Economics*, Summer 1997, *28* (2), 228–247.

Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93* (3), 395–410.

Marder, Eve, "Scientific Publishing: Beyond scoops to best practices," *Elife*, 2017, *6*, e30076.

Martz, Eric, Wayne Decatur, Joel L. Sussman, Michal Harel, and Eran Hodis, "Nobel Prizes for 3D Molecular Structure," February 2019.

Merton, Robert K., "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *American Sociological Review*, December 1957, *22* (6), 635–659.

_ , "The Matthew Effect in Science," *Science*, 1968, *159* (3810), 56–63.

Milojević, Staša, "Accuracy of simple, Initials-Based Methods for Author Name Disambiguation," *Journal of Informetrics*, 2013, *7* (4), 767–773.

Moult, John, "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction," *Current opinion in structural biology*, 2005, *15* (3), 285–289.

National Institute of General Medical Sciences, "Structural Biology," Technical Report October 2017.

Nelson, Richard R., "The Simple Economics of Basic Scientific Research," *Journal of Political Economy*, June 1959, *67* (3), 297–306.

Oster, Emily, "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

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.

Reinganum, Jennifer, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, 1983, *73* (4), 741–743.

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.

Strasser, Bruno J, *Collecting experiments: Making big data biology*, University of Chicago Press, 2019.

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?," *Journal of Legal Analysis*, 2020, *12*, 183–220.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, *58* (1), 267–288.

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

_ , Marc Weeber, Don R. Swanson, and Neil R. Smalheiser, "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation," *Journal of the American Society for Information Science and Technology*, 2005, *56* (2), 140–158.

Tripathi, Sarvind, Huiying Li, and Thomas L Poulos, "Structural basis for effector control and redox partner recognition in cytochrome P450," *Science*, 2013, *340* (6137), 1227–1230.

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

Williams, Heidi L, "Intellectual property rights and innovation: Evidence from the human genome," *Journal of Political Economy*, 2013, *121* (1), 1–27.

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.

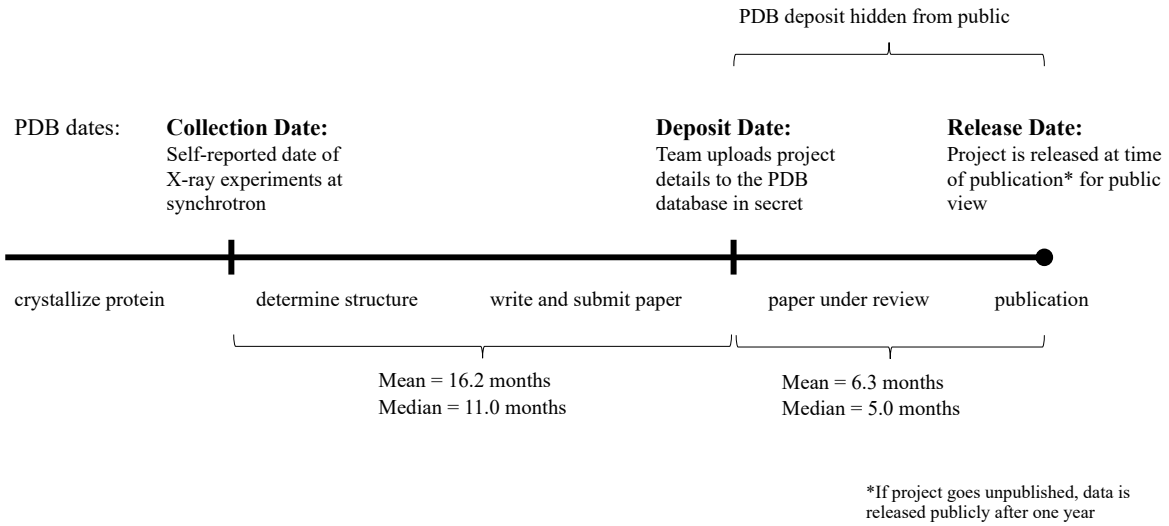
wwPDB, "wwPDB Policies and Processing Procedures Document, Release of PDB Entries," 2019.

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

Zhuo, Ran, "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs," *Working Paper*, 2022.

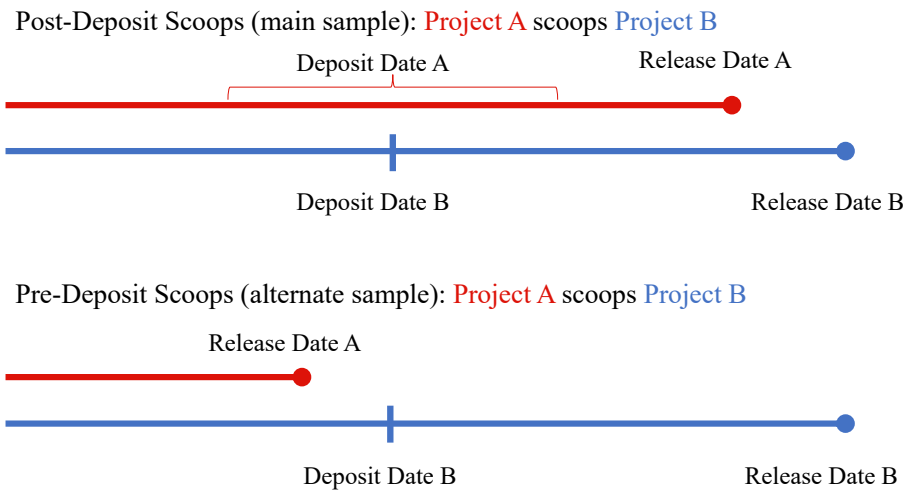
Figures and Tables

Figure 1: Project Timeline and Key Dates



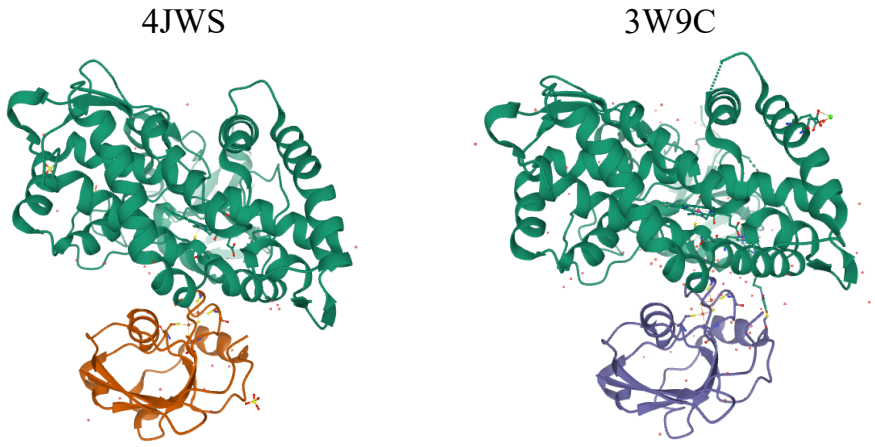
Notes: This figure shows the timeline of a typical PDB project in our regression sample. Dates in bold above the line are observed in our data. Events listed below the timeline are the approximate timing of other project events including the submission and review process. Deposit event and structure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

Figure 2: Defining Priority Races



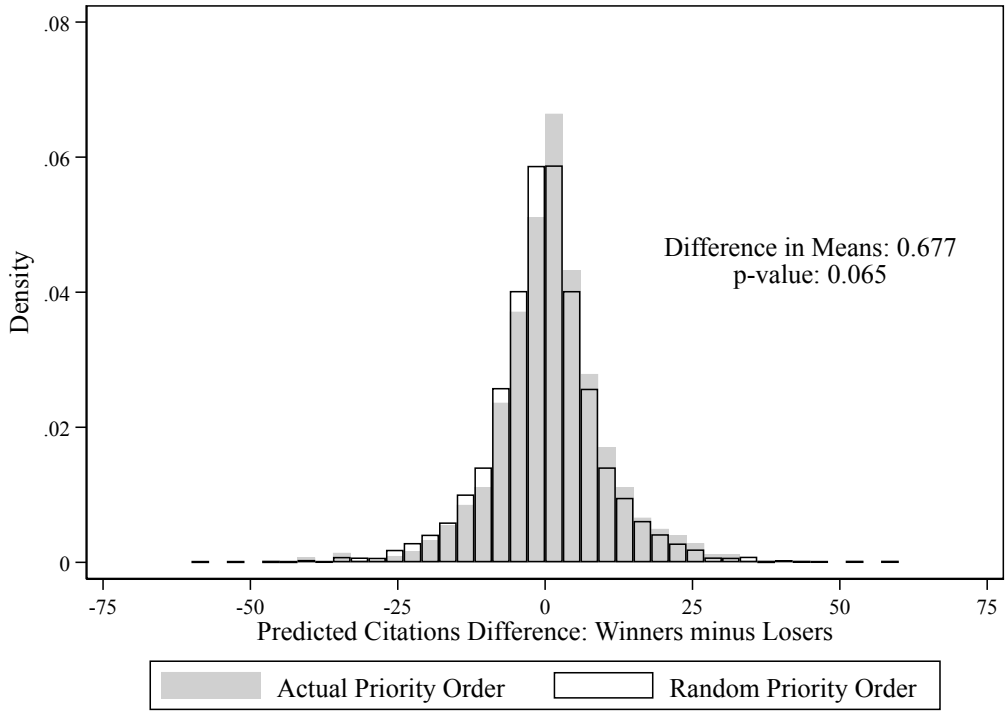
Notes: This figure shows visually the timing rule we use to define scoops. In the first example, Project A scoops Project B because both projects were deposited prior to Project A's release. These "post-deposit" scoops make up our main analysis sample of races. In the second scenario, Project A releases before Project B, but Project B had not yet deposited at the time of Project A's release. Therefore this example would be excluded from our main regression sample, but is used in our analysis of "pre-deposit" scoops in Section 5.

Figure 3: Example Priority Race — Pdx-P450cam Complex



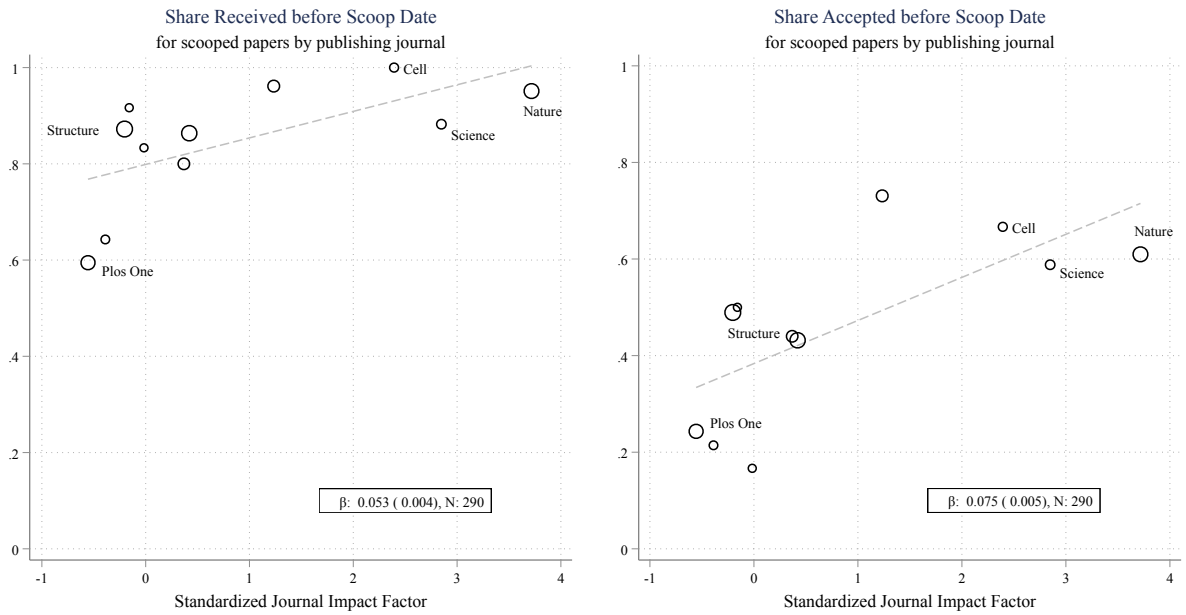
Notes: This figure presents a side-by-side comparison of the biological assembly models of the Pdx–P450cam complex protein deposited by two independent racing teams. According to the scoop definition in Section 2.4, structure deposit 4JWS scooped structure deposit 3W9C. See Table 1 for more details.

Figure 4: Histogram of Team Reputation Difference



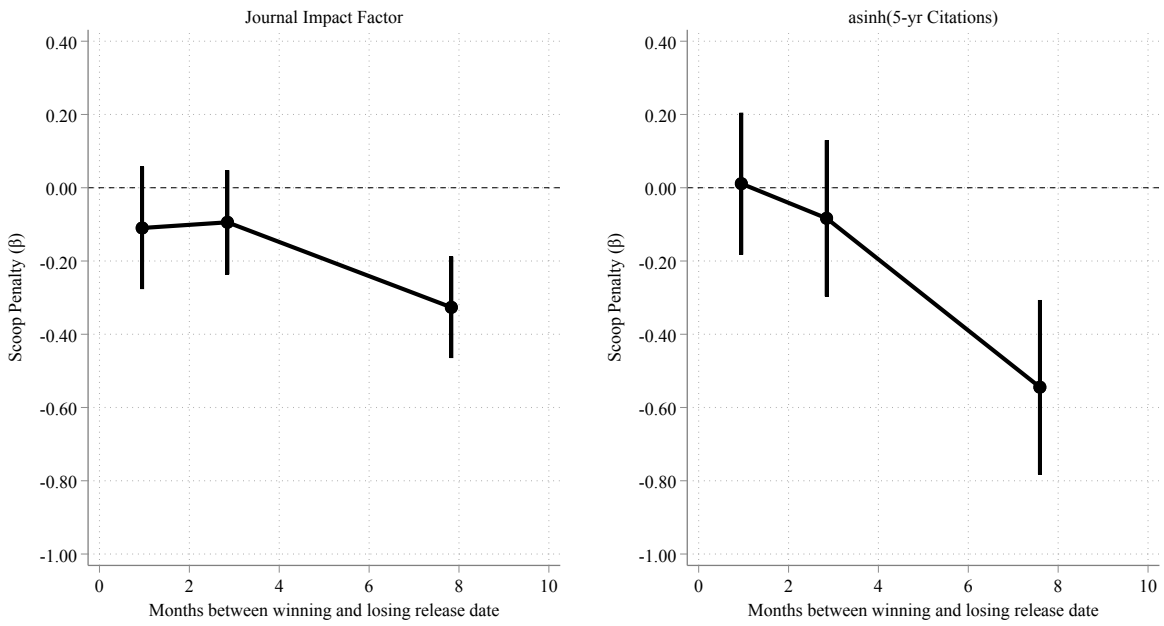
Notes: An observation in this figure is a racing pair. The blue distribution shows the actual difference in predicted citations. Bars to the right of zero represent instances when the winning team had higher predicted citations than the losing team, and bars to the left of zero represent instances when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

Figure 5: Journal Placement and Timing of Scoops



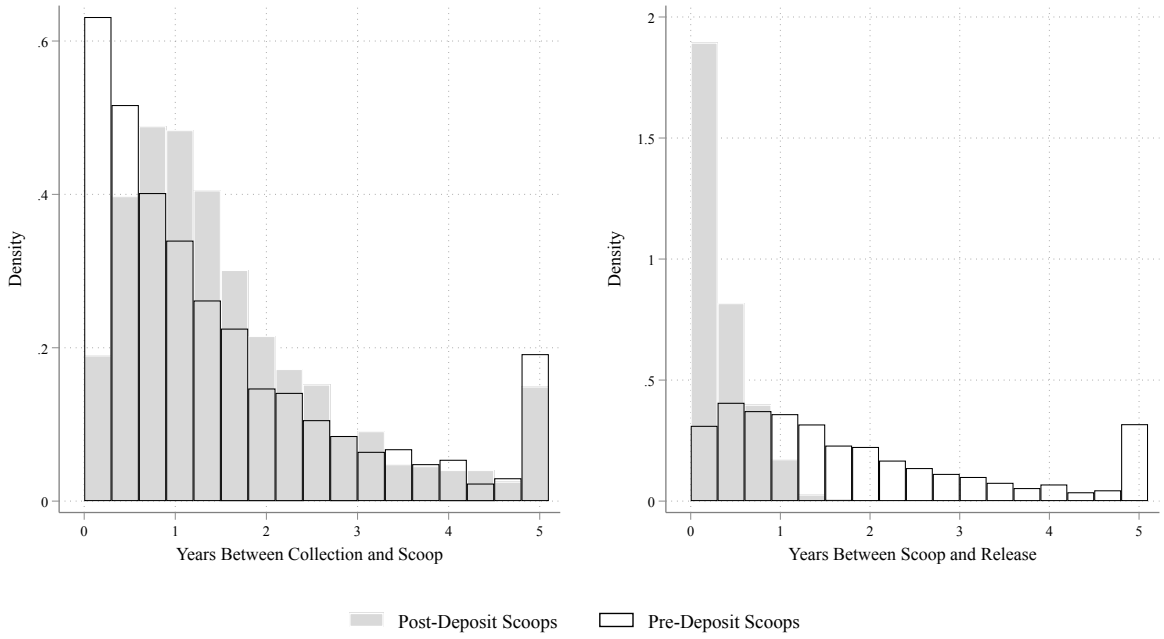
Notes: The figure reports the share of scooped papers that were received and accepted before the scoop date at different journals. Each circle represents one of the eleven largest journals that we collected supplemental data on the editorial timeline. Journals are arranged along the x-axis by their standardized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

Figure 6: JIF and Citation Penalty by Scooped Project Release Delay



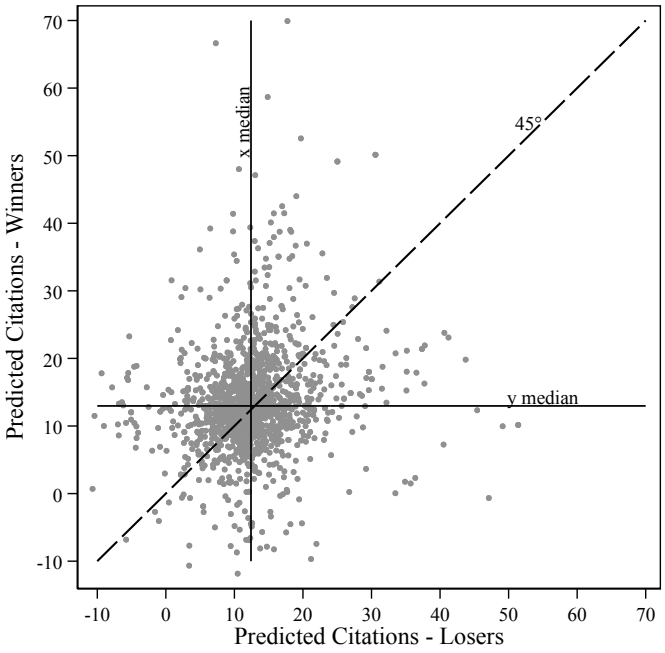
Notes: The sample of races is divided into three terciles along the distribution of time between winning and losing release date. Races are positioned along the x-axis at the average scoop release delay within each group. Projects released in close proximity are to the left, and those with a long delay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

Figure 7: Gaps Between Collection, Scoop, and Release for Pre- and Post-Deposit Scoops



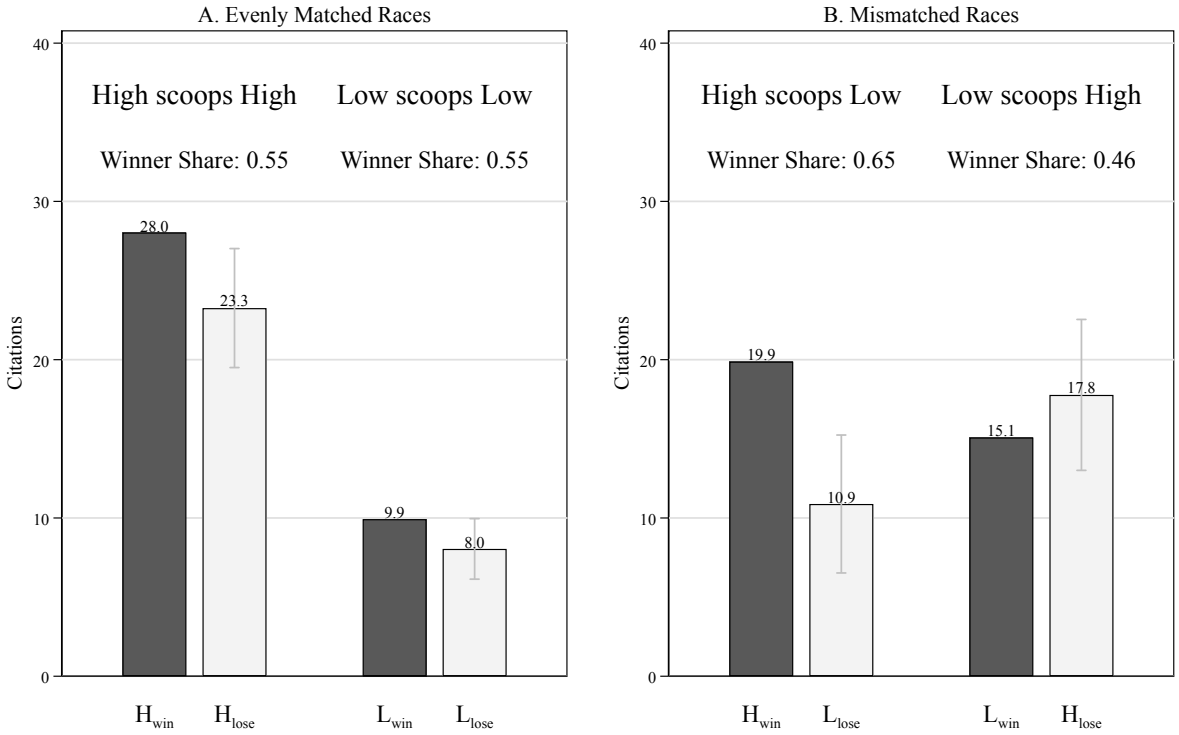
Notes: The figure shows the amount of time that passes between the collection date and the scoop (left panel) and between the scoop date and release date (right panel) for pre-deposit and post-deposit scoops. Histogram is top-coded at five years

Figure 8: Scatter Plot of Team Reputation Difference



Notes: An observation in this figure is a racing pair. The y-axis shows the predicted citations for the winning team, and the x-axis shows the predicted citations for the losing team. Perfectly matched teams would lie on the 45-degree line. If the winning team has higher predicted citations than the losing team, the dot will lie above the 45-degree line. If the winning team has lower predicted citations than the losing team, the dot will lie below the 45-degree line.

Figure 9: Priority Effect by Reputation Match-up



Notes: We divide the sample of races from Figure 8 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 — Pdx-P450cam Complex

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

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

Table 2: Summary Statistics for Structure-Level Data

Variable	Racing (1)	Not racing (2)	Difference (race - not race) (3)	Std. error of difference (4)
Panel A. Team characteristics				
Number of authors	7.120	7.454	-0.333	(0.079) ***
Affiliation in North America	0.291	0.351	-0.060	(0.008) ***
Affiliation in Europe	0.152	0.158	-0.006	(0.006)
Affiliation in Asia	0.191	0.134	0.056	(0.007) ***
Top 50 university	0.250	0.240	0.010	(0.008)
Rank 51-200 university	0.238	0.261	-0.023	(0.008) ***
Other affiliation	0.512	0.499	0.013	(0.009)
Industry or non-profit affiliation	0.152	0.171	-0.018	(0.006) ***
First author experience (years)	5.444	5.983	-0.538	(0.109) ***
Last author experience (years)	7.418	7.806	-0.387	(0.120) ***
Panel B. Project outcomes				
Published	0.866	0.752	0.114	(0.006) ***
Standardized impact factor	0.113	-0.045	0.158	(0.021) ***
Top ten journal	0.356	0.283	0.073	(0.010) ***
Five-year citation counts	26.178	17.245	8.933	(0.736) ***
Top 10% in five-year citations	0.148	0.083	0.065	(0.007) ***
Observations	3,279	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	7.183	7.056	0.127	(0.205)	
Affiliation in North American	0.351	0.262	0.320	-0.057	(0.022)	***
Affiliation in Europe	0.158	0.134	0.170	-0.036	(0.018)	**
Affiliation in Asia	0.134	0.224	0.156	0.067	(0.018)	***
Top 50 university	0.240	0.223	0.278	-0.055	(0.021)	***
Rank 51-200 university	0.261	0.248	0.228	0.020	(0.020)	
Other affiliation	0.499	0.529	0.494	0.035	(0.023)	
Industry or non-profit affiliation	0.171	0.153	0.152	0.001	(0.018)	
First author experience (years)	5.983	5.744	5.134	0.611	(0.279)	**
Last author experience (years)	7.806	7.521	7.311	0.210	(0.313)	
Panel B. First author productivity (prior five years)						
Deposits	12.361	3.791	5.504	-1.714	(0.687)	**
Publications	2.893	2.591	3.139	-0.548	(0.464)	
Top-10 publications	0.656	0.709	0.670	0.038	(0.065)	
Top-5 publications	0.222	0.262	0.239	0.023	(0.032)	
Panel C. Last author productivity (prior five years)						
Deposits	44.269	30.937	28.991	1.946	(4.327)	
Publications	9.905	12.513	13.398	-0.884	(2.241)	
Top-10 publications	4.027	4.669	4.622	0.047	(0.511)	
Top-5 publications	1.421	1.653	1.799	-0.146	(0.190)	
Panel D. Project quality metrics (lower is better)						
Resolution (Å)	2.244	2.328	2.315	0.013	(0.062)	
R-free goodness-of-fit	0.236	0.245	0.243	0.002	(0.002)	
Observations	64,018	1,668	1,611	F-stat:	4.019	***

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	-0.025	-0.192***	-0.065***	-0.245***	-0.037***
	(0.015)	(0.044)	(0.020)	(0.071)	(0.014)
Panel B. Base controls					
Scooped	-0.026**	-0.182***	-0.063***	-0.216***	-0.028**
	(0.013)	(0.045)	(0.021)	(0.063)	(0.014)
Panel C. PDS-Lasso selected controls					
Scooped	-0.026***	-0.186***	-0.062***	-0.208***	-0.036***
	(0.010)	(0.032)	(0.015)	(0.045)	(0.010)
Winner Y mean	0.879	-0.027	0.320	28.830	0.149
Observations	3,279	3,279	3,279	2,514	2,514

Notes: This table presents regression estimates of the scoop penalty, following equation 2 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 Five-Year Productivity

			Total count within five years after race				
	Any PubMed within five years	Any PDB within five years	PubMed publications	PDB publications	Top-ten publications	Citation-weighted publications	Top-10% cited publications
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. All scientists							
Scooped	-0.018***	-0.042***	-1.129	-0.072	-0.139	-0.200***	-0.589***
	(0.006)	(0.010)	(1.046)	(0.221)	(0.101)	(0.045)	(0.202)
Winner Y mean	0.841	0.702	45.869	7.154	3.610	497.203	7.741
Observations	8,624	8,624	8,624	8,624	8,624	6,484	6,484
Panel B. Novices							
Scooped	-0.058***	-0.040**	-0.019	0.003	0.106	-0.335***	-0.181
	(0.018)	(0.019)	(0.276)	(0.168)	(0.068)	(0.102)	(0.118)
Winner Y mean	0.469	0.356	4.243	1.890	0.616	75.691	1.165
Observations	2,033	2,033	2,033	2,033	2,033	1,529	1,529
Panel C. Veterans							
Scooped	-0.006*	-0.040***	-1.179	-0.163	-0.235	-0.160***	-0.821***
	(0.003)	(0.012)	(1.553)	(0.309)	(0.145)	(0.046)	(0.286)
Winner Y mean	0.990	0.839	61.681	9.261	4.787	667.421	10.388
Observations	5,821	5,821	5,821	5,821	5,821	4,378	4,378

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with seven years or less 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 6: Decomposing Citation and Journal Effect

	Five-year citations			
Dependent variable	(1)	(2)	(3)	(4)
Scooped	-0.155***	-0.111***	-0.102***	-0.044*
	(0.032)	(0.029)	(0.028)	(0.027)
Journal controls	None	Linear JIF	Cubic JIF	Journal FE
Winner Y mean	34.7	34.7	34.7	34.7
Observations	1,891	1,891	1,891	1,891

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 4, 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 7: Strategic Responses to Pre- and Post-deposit Scoops

Dependent variable	Maturation (1)	Number of Proteins in Paper (2)	Multiple Proteins in Paper (3)	"Structure" in Paper Title (4)	"Function", "Mechanism" or "Analysis" in Title (5)	Uses Molecular Replacement (6)
Constant	1.173***	1.395***	0.437***	0.876***	0.156***	0.614***
	(0.068)	(0.088)	(0.013)	(0.016)	(0.010)	(0.010)
Scooped	0.025	-0.041	-0.005	-0.017	0.001	0.022
	(0.062)	(0.072)	(0.021)	(0.016)	(0.017)	(0.017)
Pre-deposit x Scooped	1.362***	0.201**	0.061**	-0.062***	0.042**	0.080***
	(0.095)	(0.090)	(0.027)	(0.021)	(0.021)	(0.021)
Observations	5,398	5,398	5,398	5,398	5,398	5,398

Notes: This table presents regression estimates of strategic reponse outcomes on a scooped indicator and an interaction between a pre-deposit indicator and the scooped indicator. Pre-deposit scoops are those where the scooped team had collected data but not yet deposited at the time of the first paper release. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. All regressions include controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level.
**p<0.1, **p<0.05, ***p<0.01.*

Table 8: Survey Benchmark of Scoop Penalty

	PDB estimate		Survey estimate			
	Full sample (1)	Comparable subsample (2)	All respondents (3)	Below-median reputation (4)	Above-median reputation (5)	Column (4) - (5) difference (6)
Prob (Scoop)	0.029	0.081	0.266	0.267	0.266	0.001
						(0.016)
Prob (Publication)	0.854	0.976	0.665	0.636	0.694	-0.059***
						(0.022)
Journal impact factor penalty	-0.186	-1.208	-2.918	-2.937	-2.900	-0.036
						(0.084)
Citation penalty	-0.208	-0.135	-0.594	-0.614	-0.575	-0.040*
						(0.024)

Notes: This table reports results from the PDB (columns (1) and (2)) and from a survey of structural biologists (columns (3) to (6)). In column (1) we report the mean values of a scoop indicator in the full sample (N = 67,933). The remaining estimates are estimated identically to Table 4, Panel C. In column (2), we repeat this procedure in a subsample of the PDB that is comparable to the survey text. Specifically, we restrict to PDB races where one racer published in Science, Nature, or Cell, and losing team was scooped early in the process (quarter of sample with the shortest time between loser deposit and winner release; N = 3,152). In column (3), we report the results of a survey of structural biologists (N = 822 respondents). The survey asked respondents to estimate the probability and consequences of getting scooped on a hypothetical project. See Appendix D for full survey text. Note that not all respondents answered all questions, so the sample size varies across rows (N = 768, 672, 597, and 675 respectively). 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.[31]

We access the data directly from the RCSB Custom Report Web Service.[32] The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

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

Additional data on cluster entities was accessed through a separate raw file archive at RCSB[33] 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.[34] 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. The JIF data was only avialable between 1997 and 2017, so we imputed impact factors in years before or after that window with the 1997 or 2017 impact factors respectively. We standardize 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,[35] 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.

[31] http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction
[32] https://www.rcsb.org/pdb/results/reportField.do
[33] ftp://resources.rcsb.org/sequence/clusters/ clusters50.txt and clusters70.txt
[34] Both data sources were owned by Thompson Reuters at the time of access, but have since been sold to Clarivate Analytics.
[35] Because structural biology falls squarely within the life sciences, restricting to citations with PubMed IDs is does not have a large effect on citation counts.

A.3 Altmetric.com Data

We use data from Altmetric.com to measure alternative forms of attention for academic research.[36] 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 later. 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. 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. This subsample covers 19 percent of our primary regression sample.

We use these data to look at the correspondence between the journal publication date and the release date. Appendix Figure A4 reports the correspondence between the PDB release date and the publication date for the 616 articles in the racing sample for which they are available. This correspondence is not exact for a few reasons. First, according to PDB policy, scientists are allowed to release their findings immediately after deposit, which could potentially come before the publication date. In typical practice, the scientists prefer to wait until publication so that other scientists cannot use the information for follow-on work until after publication. In fact, scientists prefer to wait for release as long as possible to maintain a competitive advantage, which was the motivation behind the 1998 policy change to align release and publication (Campbell 1998). Another reason that release may come earlier than publication is because of the policy that all data is released after one year. If a team takes more than one year to publish results after the deposit, they would be forced to release at the one year point even if they eventually publish. Release sometimes happens after publication, but these cases should be rare and only be delayed for a few weeks. Any longer delays for release is either due to data errors or non-compliance with PDB policies.

Overall, 49 percent of the release dates are within two weeks of publication. This may lead to concerns about potential measurement error in the definition of the priority ordering. Throughout the paper, we always define the order of PDB release as the rule for being scooped. The community tracks public PDB releases carefully, so we believe this is a valid definition of priority. Publication dates are also complicated in recent years by the practice of online publication, which sometimes comes weeks before the print edition is published. But even if we prefer to consider only the publications as a claim to priority, our release date definition appears to usually correspond to the publication date ordering. In the 99 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 83 percent of the time. To the degree that this is

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

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 44,141 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.[37] This website provides rankings for 500 top academic programs based on surveys of academics and employers as well as citations per paper and h-index of the scientists affiliated with each department.

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

At various points in our analysis, we construct panel data of individual scientist and team productivity. First, we use measures of past PDB and PubMed productivity as control variables (Tables 3 and 4) and to predict citations as a measure of team reputation (Figures 8 and 9). Second, we use a panel of publications to construct long-run outcomes in the years following a scoop event (Table 5). The PDB does not explicitly link authors between deposits, and neither PubMed nor our version of Web of Science have author identifiers across publications. A further challenge is that many PDB deposits are not linked to a publication, so constructing control variables of past productivity is difficult using only publication data. We therefore use two separate approaches for constructing author-level panel variables: 1) Link PDB deposits by simple author name matching for control variables, 2) Use name disambiguation clustering from the Author-ity project (Torvik et al. 2005; Torvik and Smalheiser 2009) to count future publications and citations for long-run outcomes.

Simple Author Name Matching in PDB

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

The name disambiguation procedure using only last names and initials is more reliable in a smaller subset of academic papers. We therefore choose to focus the panel only on PubMed papers that are linked

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

to the PDB instead of trying to use the full PubMed archive, which covers all of the medical and life science literature. This choice improves the reliability of our name-matching, but offers less information about academic productivity. Since we can use PDB name matching for unpublished deposits, we use this approach for constructing control variables for our main analysis.

Scientists usually identify themselves on publications with a consistent last name, but are sometimes inconsistent with their use of first and last initials, or first names and nicknames.[38] 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. We include a dummy control variable throughout the analysis that indicates the common names to help account for the possibility that name-matching errors are correlated with treatment.

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

Author-ity Name Disambiguation

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

[38] 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.

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

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

1. Names that do not match to any Author-ity cluster (12 percent of racing sample authors). These are individuals who deposit at least once in the PDB, but never publish a paper (e.g. a graduate student that does not pursue academia).

2. Names that have PubMed IDs that match to one and only Author-ity person ID (60 percent of racing sample authors). We take this exclusive matching as evidence that all instances of the name in the PDB is a single person that is represented by the matched Author-ity person ID.

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

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

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

B Protein Similarity and Race Definition

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

B.1 Sequence Similarity Algorithm

Each protein in the PDB is a chain composed of the 22 different types of proteinogenic amino acids in some combination. The order of these molecules in the chain defines the type of protein, and we use this code to compare the similarity of the proteins that scientists are working on. The PDB provides a clustering algorithm called the Basic Local Alignment Search Tool or BLAST (Altschul et al. 1990) which creates groupings of structure deposits that have identical or similar amino acid chains. The clusters can be defined at different thresholds of similarity, including 100 percent, 90 percent, 70 percent, and 50 percent. One possible approach to defining races would be to only focus on competing projects that determine the structure of proteins that are 100 percent similar. But in many cases, two proteins that are 90 percent similar or lower have many of

the same defining features and functions within the same organism or across different species. Therefore, many interesting priority races are between teams working on very similar if not identical proteins. Following the similarity threshold chosen by (Brown and Ramaswamy 2007), we define racing for proteins all the way down to 50 percent similarity. We include races with a broad threshold in part to increase the sample size for our regressions, but also to include races over discoveries that were exceedingly different from any past structure discoveries. Other recent economics papers that study protein clusters also use similar cutoffs (Kim, 2023; Zhuo, 2022). We further validate our choice of similarity by comparing pairs of papers in each similarity category. Figure A1 calculates the share of scooped papers that cite the winning paper and plots it separately by sequence similarity. These are constructed as mutually exclusive groups with structures placed in the highest similarity cluster they appear in together. 70 percent, 90 percent, and 100 percent show almost identical rates, and the 50 percent similar pairs have only a slightly lower rate. Figure A2 shows the similarity between the winning and losing paper titles calculated with a character-replace string similarity metric. Here we see that the titles are equally similar between 50 percent, 70 percent, 90 percent, and 100 percent similarity papers.

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 (79 percent of structures in the PDB), a new structure discovery might directly scoop another team working on the same entity. But in some cases, a team working on a single entity might scoop a team that is working on a complex protein with multiple entities, only one of which was being worked on by both teams. These deposits will still be linked by the algorithm, but the interpretation of the scooping event is less obvious. We consider these cases to be "partial scoops" where some part of the scientific discovery was overshadowed by the winning team. Since outcomes are defined at the protein and paper level, including these partial scoops will potentially understate the effect of an average "full scoop." In our final regression sample, 68 percent of races are composed only of single-entity structures, 16 percent are exclusively multi-entity structures, and 16 percent are a mix of single- and multi-entity structures. 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.[39]

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. In Appendix Table A5, we include robustness results of our main effects for races defined at the 100 percent level, and show that the results are comparable.

[39] 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 Theoretical Appendix

C.1 A Model of Strategic Responses to Getting Scooped Before Project Completion

C.1.1 Setup

We start by considering the optimization problem of a scientist at the outset of a race with no information of her competitor's progress relative to her own. First, she chooses a maturation period m, which is the time she will spend on the project from start to finish. Higher m increases the value $V(m)$ of the project but also increases the cost $c \cdot m$ of the project, where $V(0) = 0$, $V'(m) > 0$, and $V''(m) < 0$. Given the utility function

$$u(m) = V(m) - c \cdot m$$

the scientist will select some m^* that maximizes her utility.[40] This m^* is defined by the first-order condition:

$$V'(m) = c.$$

As structural biology is a secretive field, we assume she has no knowledge about any potential competitors or their progress, and thus she will commit to this m^* unless additional information is revealed.

Now suppose there is a penalty for being scooped $\theta \in [0, 1]$ such that the losing team gets $\theta V(m)$ rather than $V(m)$ if the other team finishes first. Due to the attribution frictions raised by Dasgupta and David (1994) and the empirical evidence we show in Figure 6, we let the scoop penalty vary with the gap between when the two papers are released. In particular, let this gap be denoted $g(m)$ and let the scoop penalty be decreasing and convex in g. In other words, $\theta(0) = 1$, $\theta'(g) < 0$, and $\theta''(g) > 0$

C.1.2 Re-optimization.

If the scientist learns that she has been scooped before completing the project, new information is revealed and she has a chance to re-optimize. In other words, if she learns that she has been scooped at time $m_1 < m^*$, she now maximizes

$$u(\tilde{m}) = \theta(g(\tilde{m}))V(\tilde{m}) - \tilde{c} \cdot (\tilde{m} - m_1) \quad \text{subject to } \tilde{m} \geq m_1.$$

We let $\tilde{m}$ represent her new choice of m, after this revelation of information. $\theta(g(\tilde{m}))V(\tilde{m})$ is the value of the project despite getting scooped, and $\tilde{c} \cdot (\tilde{m} - m_1)$ is the remaining costs left to pay. In this case, $g(\tilde{m}) = \tilde{m} - m_1$. We assume that $\tilde{c} \leq c$, capturing the fact that the release of the first project may make some aspects of the project easier, due to informational spillovers. The solution to this optimization problem is $\tilde{m}^*$, which is implicitly defined by the first-order condition:

$$\theta(g(\tilde{m}^*))V'(\tilde{m}^*) = -\theta'(g(\tilde{m}^*))V(\tilde{m}^*) + \tilde{c}.$$

The left side of this equation represents the marginal benefit of increasing m: the marginal benefit of adding value. The right hand side represents the marginal costs of increasing m: the marginal decay in credit plus the marginal costs of spending additional time on the project. The behavior of both sides of the equation

[40] Note that to keep things simple, the researcher does not consider the probability of getting scooped in her selection of m. We could modify the problem to make $u(m)$ depend on some expectation of the credit she gets for $V(m)$ (similar to Hill and Stein (2024)). All we require here is a non-zero solution to the maximization problem.

depends on the specific functional forms of $\theta(\cdot)$ and $V(\cdot)$. However, we would like to show that it is possible for either $\tilde{m}^* < m^*$ or $\tilde{m}^* > m^*$. We can simply show that this is the true via specific examples.

Let $V(m) = \ln(1+m)$, and let $c = 0.5$. In this case, before knowledge of the scoop, the first-order condition yields:

$$V'(m^*) = c \implies \frac{1}{1+m^*} = 0.5 \implies m^* = 1.$$

Now, suppose that $\theta(g) = e^{-0.1g}$. To keep things simple, further assume that $m_1 = 0$, so that $\theta(g) = \theta(\tilde{m}) = e^{-0.1\tilde{m}}$. Start by letting $\tilde{c} = 0.5 = c$. The first-order condition for this problem yields:

$$\begin{aligned}
\theta(\tilde{m}^*)V'(\tilde{m}^*) &= -\theta'(\tilde{m}^*)V(\tilde{m}^*) + \tilde{c} \\
e^{-0.1\tilde{m}^*}\left(\frac{1}{1+\tilde{m}^*}\right) &= 0.1e^{-0.1\tilde{m}^*}\ln(1+\tilde{m}^*) + 0.5 \\
\tilde{m}^* &\approx 0.70 < m^*
\end{aligned}$$

In this case, upon learning she has been scooped, the researcher will publish earlier than she originally planned. However, if instead $\tilde{c} = 0.25 < c$ due to informational spillovers from the first project, then solving the first-order condition yields $\tilde{m}^* \approx 1.58 > m^*$. In this case, upon learning she has been scooped, the researcher decides to slow down.

In general, we have three possible cases:

1. Case 1 ("hurry up and finish"): In this case, the researcher chooses $\tilde{m}^* < m^*$, because the decaying credit plus the continuation costs outweigh increasing the value of the project.
2. Case 2 ("delay and expand"): In this case, the researcher chooses $\tilde{m}^* > m^*$, because the increasing value of the project outweighs the decaying credit plus continuation costs.
3. Case 3 ("abandon"): Of course, it is possible that even after selecting a new $\tilde{m}^*$, the benefits once the researcher knows they will be scooped are not enough to outweigh the costs of completing the project. These projects will therefore go unfinished.

C.2 A Model of Academic Attention

C.2.1 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 (2020) and may arise from inattention or uncertainty about the importance of the contribution. The signal, s, is a function of the paper's true underlying quality (q) as well as a noise term, u:

$$s = q + u$$

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

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

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

signal is informative relative to the noise term.

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 $\hat{q}_1 + f < \hat{q}_2$.

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:

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

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 of Proposition 1.

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

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

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

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 of Proposition 2.

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.

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:

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

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

cited is:

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

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

D 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

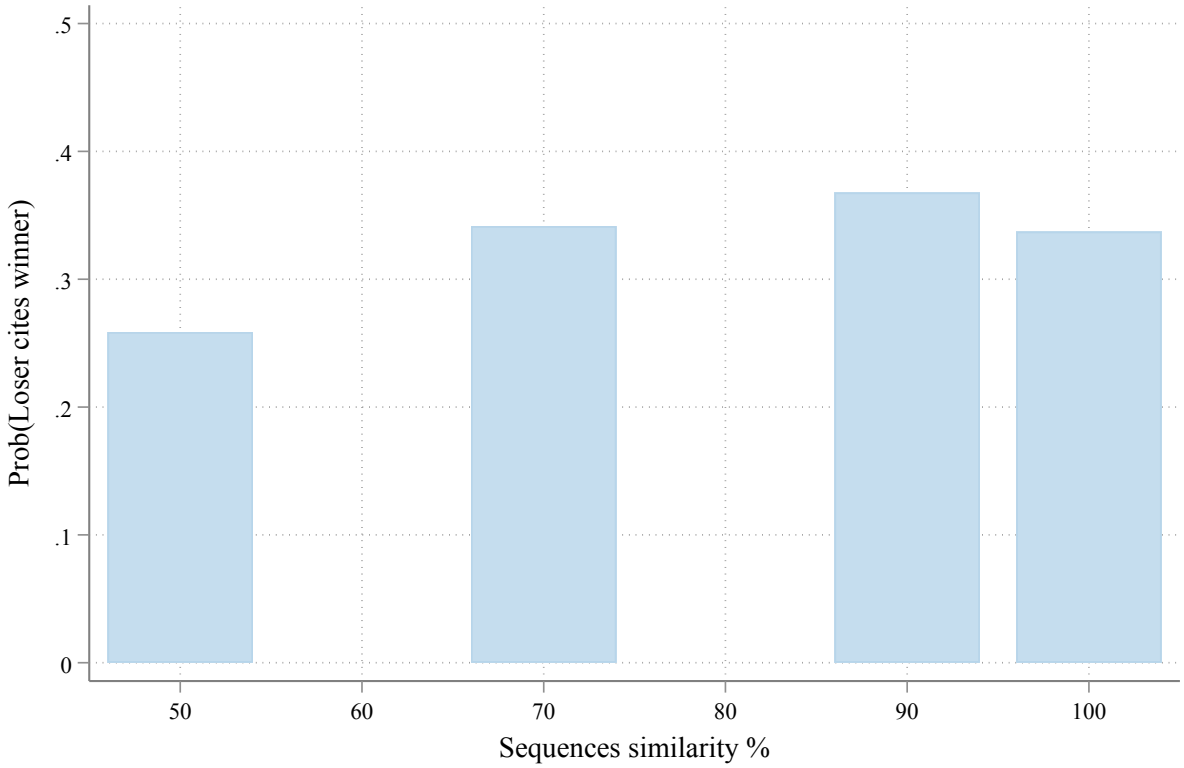
Probability of Publication

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?

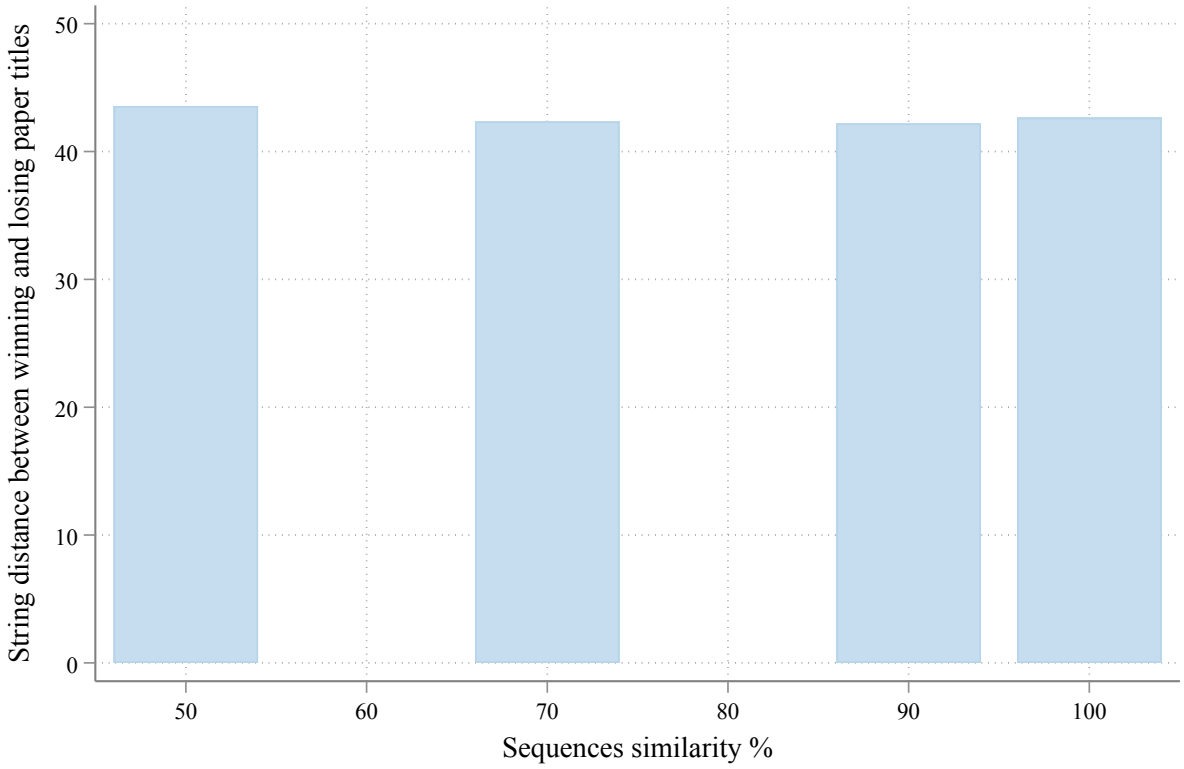
E Appendix Figures and Tables

Figure A1: Probability of Loser Citing Winner by Sequence Similarity



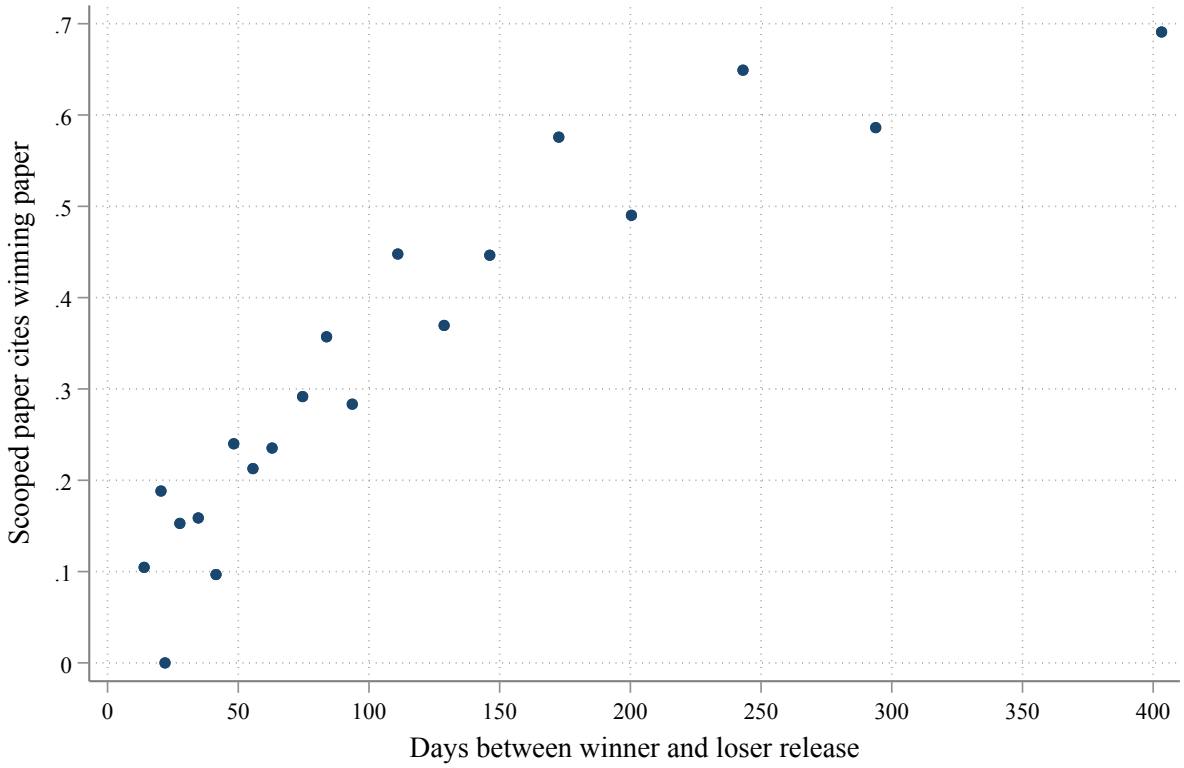
Notes: This figure shows the probability that the losing team cites the winning team at increasing levels of sequence similarity. Similarity groups are mutually exclusive so that races are placed in the highest similarity cluster in which they appear together.

Figure A2: Title Similarity Between Winning and Losing Paper by Sequence Similarity



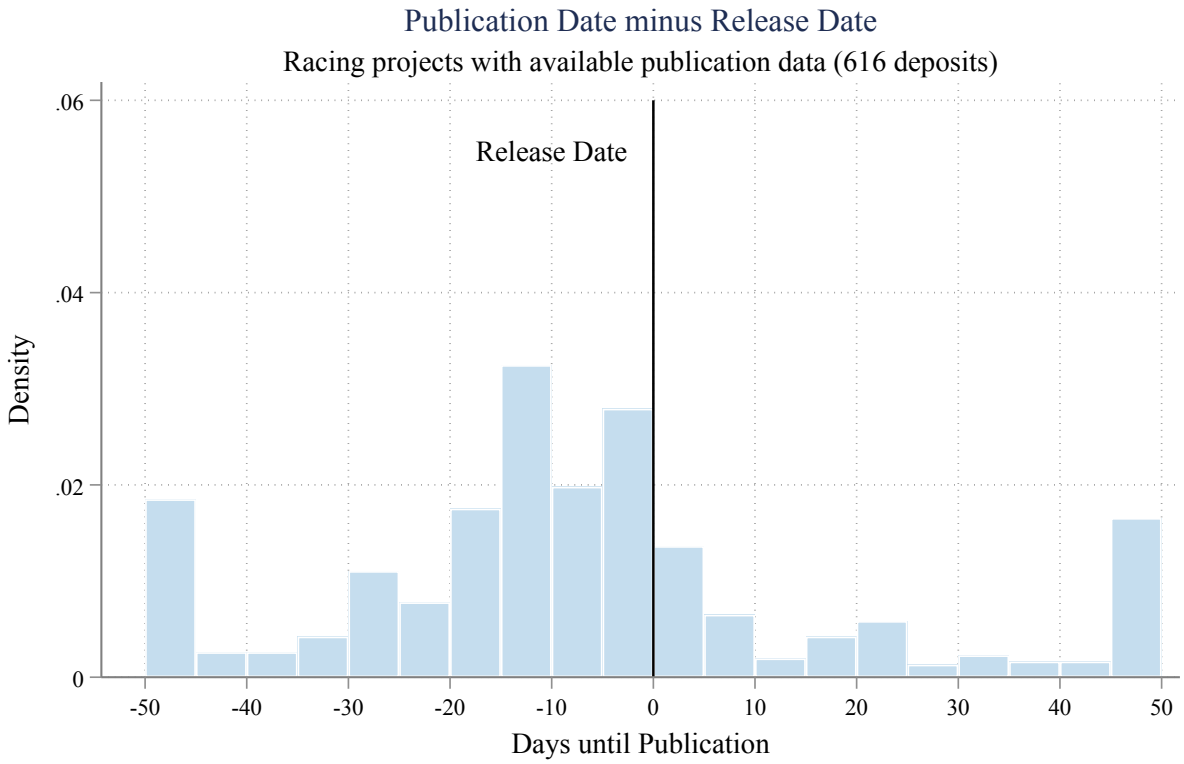
Notes: This figure shows the character replacement string similarity of the titles of the winning and losing papers at increasing levels of sequence similarity. Similarity groups are mutually exclusive so that races are placed in the highest similarity cluster in which they appear together.

Figure A3: Probability that Scooped Paper Cites Winning Paper by Release Date Gap



Notes: This binned scatterplot shows the probability that the scooped paper cited the winning paper by the number of days between the release dates of the winning and losing projects. Sample is 1,149 races where both teams published papers with a PubMed ID.

Figure A4: 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	1.72
Affiliation in Asia	-3.53
Non-academic affiliation	1.73
First author experience (years)	-0.20
First author top-5 publications, 5 prior years	2.45
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	0.21
Last author top-5 publications, 5 prior years	2.14
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	3.48
71-80	-0.17
81-90	-1.03
101-110	-2.39
111-120	5.03
151-160	-2.67
171-180	-2.08
181-190	-0.40
211-220	-5.18
231-240	-1.43
271-280	-4.11
291-300	-2.85
401-410	-2.70
Constant	10.28
R-squared	0.102
N	58,758

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

Table A2: Effect of Getting Scooped on Project Outcomes - Oster (2019) Robustness Check

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, no FE					
Scooped	-0.025**	-0.191***	-0.064***	-0.239***	-0.035***
	(0.011)	(0.031)	(0.014)	(0.051)	(0.010)
	[0.001]	[0.008]	[0.005]	[0.006]	[0.003]
Panel B. Base controls, protein FE					
Scooped	-0.026**	-0.182***	-0.063***	-0.216***	-0.028**
	(0.013)	(0.045)	(0.021)	(0.063)	(0.014)
	[0.704]	[0.676]	[0.607]	[0.767]	[0.725]
Oster (2019) Bias-adjusted β	-0.027	-0.177	-0.061	-0.209	-0.025
Selection ratio (δ) needed for $\beta = 0$	60.0	15.2	14.0	16.5	7.9

Notes: This table presents regression estimates of the scoop penalty following equation 2 in the text (see Table 4). Panel A reports coefficients from a simple bivariate regression with no controls or protein fixed effects with standard errors in parentheses and R^2 in brackets. Panel B includes all base controls and protein fixed effects, comparable to panel B in Table 4. The Oster (2019) bias adjusted coefficient assumes a maximum R^2 =1 and δ=1, meaning we assume that treatment is selected equally on observables and unobservables. The selection ratio (δ) needed for $\beta = 0$ shows that treatment would need to be 7 times more selected on unobservables than observables for the coefficient to equal zero.
**p<0.1, **p<0.05, ***p<0.01.*

Table A3: Effect of Getting Scooped on Project Outcomes - Alternative Hit Rate Metrics

Dependent variable	Top-1% three year citations (1)	Top-5% three year citations (2)	Top-10% three year citations (3)	Top-1% five year citations (4)	Top-5% five year citations (5)	Top-10% five year citations (6)	Top-1% ten year citations (7)	Top-5% ten year citations (8)	Top-10% ten year citations (9)
Panel A. No controls									
Scooped	-0.007	-0.023**	-0.033**	-0.006	-0.017*	-0.037***	-0.011*	-0.031**	-0.049***
	(0.005)	(0.009)	(0.013)	(0.005)	(0.010)	(0.014)	(0.006)	(0.014)	(0.017)
Panel B. Base controls									
Scooped	-0.007*	-0.020**	-0.027**	-0.005	-0.013	-0.028**	-0.009	-0.027*	-0.044***
	(0.004)	(0.010)	(0.013)	(0.005)	(0.010)	(0.014)	(0.006)	(0.014)	(0.017)
Panel C. PDS-Lasso selected controls									
Scooped	-0.007**	-0.022***	-0.031***	-0.005	-0.015**	-0.036***	-0.010**	-0.030***	-0.046***
	(0.003)	(0.007)	(0.010)	(0.003)	(0.007)	(0.010)	(0.004)	(0.010)	(0.012)
Winner Y mean	0.012	0.076	0.148	0.011	0.068	0.149	0.012	0.081	0.153
Observations	2,931	2,931	2,931	2,514	2,514	2,514	1,515	1,515	1,515

Notes: This table presents regression estimates of the scoop penalty with alternative hit rate measures, following equation 2 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.
**p<0.1, **p<0.05, ***p<0.01.*

Table A4: Effect of Getting Scooped on Project Outcomes - No Protein (i.e., Race) Fixed Effects

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	-0.025**	-0.191***	-0.064***	-0.239***	-0.035***
	(0.011)	(0.031)	(0.014)	(0.051)	(0.010)
Panel B. Base controls					
Scooped	-0.022**	-0.154***	-0.051***	-0.180***	-0.025**
	(0.009)	(0.032)	(0.014)	(0.046)	(0.010)
Panel C. PDS-Lasso selected controls					
Scooped	-0.021**	-0.146***	-0.058***	-0.169***	-0.024**
	(0.010)	(0.031)	(0.015)	(0.046)	(0.010)
Winner Y mean	0.879	-0.027	0.320	28.830	0.149
Observations	3,279	3,279	3,279	2,514	2,514

Notes: This table presents regression estimates of the scoop penalty, following equation 2 in the text, but excluding 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 A5: Effect of Getting Scooped on Project Outcomes - 100 Percent Sequence Similarity

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls					
Scooped	-0.023	-0.174**	-0.051	-0.272**	-0.047**
	(0.025)	(0.070)	(0.032)	(0.112)	(0.021)
Panel B. Base controls					
Scooped	-0.035	-0.156**	-0.044	-0.288***	-0.032
	(0.022)	(0.074)	(0.034)	(0.109)	(0.020)
Panel C. PDS-Lasso selected controls					
Scooped	-0.027	-0.172***	-0.049**	-0.253***	-0.047***
	(0.018)	(0.052)	(0.023)	(0.080)	(0.015)
Winner Y mean	0.882	-0.078	0.289	27.956	0.138
Observations	1,178	1,178	1,178	891	891

Notes: This table presents regression estimates of the scoop penalty comparable to Table 4 in the main text. This version restricts to protein clusters in which the BLAST algorithm classifies the protein sequences as being 100% similar. This sub-sample therefore offers the narrowest definition of a scoop where the racing projects are scientifically identical. See Table 4 notes for regression details.
**p<0.1, **p<0.05, ***p<0.01.*

Table A6: 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	-0.468***	-0.108**	-0.038**	-0.009	-0.112	-0.246***
	(0.151)	(0.042)	(0.018)	(0.028)	(0.078)	(0.095)
Panel B. Base controls						
Scooped	-0.462***	-0.092**	-0.031	-0.003	-0.094	-0.216**
	(0.146)	(0.043)	(0.020)	(0.031)	(0.075)	(0.091)
Panel C. PDS-Lasso selected controls						
Scooped	-0.427***	-0.103***	-0.036***	-0.010	-0.095*	-0.228***
	(0.103)	(0.031)	(0.014)	(0.021)	(0.054)	(0.066)
Winner Y mean	43.025	0.650	0.105	0.262	3.974	9.201
Observations	1,321	1,321	1,321	1,321	1,321	1,321

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 A7: Effect of Getting Scooped on Three-Year Productivity

			Total count three years after race				
Dependent variable	Any PubMed 3 years later (1)	Any PDB 3 years later (2)	PubMed Publications (3)	PDB Publications (4)	Top-ten publications (5)	Citation-weighted publications (6)	Top-10% cited publications (7)
Panel A. All scientists							
Scooped	-0.012**	-0.039***	-0.340	-0.097	-0.037	-0.188***	-0.334**
	(0.006)	(0.011)	(0.520)	(0.117)	(0.061)	(0.042)	(0.133)
Winner Y mean	0.824	0.646	27.145	4.305	2.184	297.661	4.655
Observations	10,033	10,033	10,033	10,033	10,033	7,660	7,660
Panel B. Novices							
Scooped	-0.034**	-0.019	-0.044	-0.091	0.074*	-0.271***	-0.069
	(0.016)	(0.018)	(0.142)	(0.097)	(0.040)	(0.085)	(0.065)
Winner Y mean	0.428	0.309	2.307	1.097	0.334	44.063	0.680
Observations	2,369	2,369	2,369	2,369	2,369	1,806	1,806
Panel C. Veterans							
Scooped	-0.006	-0.037***	-0.184	-0.060	-0.060	-0.171***	-0.472**
	(0.004)	(0.013)	(0.794)	(0.163)	(0.089)	(0.047)	(0.191)
Winner Y mean	0.983	0.781	36.687	5.595	2.917	400.753	6.263
Observations	6,729	6,729	6,729	6,729	6,729	5,167	5,167

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the three years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with seven years or less 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 A8: Effect of Getting Scooped on Ten-Year Productivity

			Total count ten years after race				
Dependent variable	Any PubMed 10 years later (1)	Any PDB 10 years later (2)	PubMed Publications (3)	PDB Publications (4)	Top-ten publications (5)	Citation-weighted publications (6)	Top-10% cited publications (7)
Panel A. All scientists							
Scooped	-0.012*	-0.034***	-2.786	0.042	-0.269	-0.026	-1.008
	(0.007)	(0.013)	(2.786)	(0.526)	(0.232)	(0.071)	(0.635)
Winner Y mean	0.857	0.739	91.647	13.965	7.090	928.013	14.076
Observations	5,351	5,351	5,351	5,351	5,351	3,114	3,114
Panel B. Novices							
Scooped	-0.044*	-0.056**	0.192	0.276	0.212	-0.168	0.410
	(0.022)	(0.026)	(0.803)	(0.470)	(0.183)	(0.148)	(0.338)
Winner Y mean	0.513	0.417	9.900	3.739	1.301	122.905	1.792
Observations	1,258	1,258	1,258	1,258	1,258	743	743
Panel C. Veterans							
Scooped	-0.002	-0.026*	-5.335	-0.873	-0.749**	-0.121*	-1.988**
	(0.002)	(0.013)	(4.058)	(0.710)	(0.333)	(0.065)	(0.875)
Winner Y mean	0.995	0.869	124.346	18.108	9.418	1243.954	18.891
Observations	3,607	3,607	3,607	3,607	3,607	2,079	2,079

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the ten years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with seven years or less 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 A9: Effect of Getting Scooped on Five-Year Productivity, Author Position Subsamples

			Total count within five years after race				
	Any PubMed within five years	Any PDB within five years	PubMed publications	PDB publications	Top-ten publications	Citation-weighted publications	Top-10% cited publications
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. First Authors							
Scooped	-0.031*	-0.037	1.982	-0.025	0.018	-0.133	0.691
	(0.017)	(0.023)	(2.169)	(0.283)	(0.135)	(0.108)	(0.657)
Winner Y mean	0.821	0.692	31.576	4.191	2.045	278.251	4.296
Observations	1,166	1,166	1,166	1,166	1,166	890	890
Panel B. Middle Authors							
Scooped	-0.020**	-0.047***	-1.828	-0.279	0.013	-0.237***	-0.613**
	(0.010)	(0.016)	(1.428)	(0.237)	(0.129)	(0.071)	(0.298)
Winner Y mean	0.828	0.658	42.433	5.476	3.020	481.690	7.378
Observations	4,833	4,833	4,833	4,833	4,833	3,624	3,624
Panel C. Last Authors							
Scooped	-0.012	-0.044**	-2.515	-0.721	-0.775**	-0.311***	-1.099**
	(0.009)	(0.019)	(2.330)	(0.641)	(0.310)	(0.093)	(0.481)
Winner Y mean	0.901	0.843	61.543	14.557	6.629	669.102	10.964
Observations	1,190	1,190	1,190	1,190	1,190	900	900

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for the first scientist listed on the structure deposit, Panel B restricts to middle authors, and Panel C restricts to last authors. We use the the author list and ordering on the structure deposit because it is available for all teams regardless of publication status. It is usually the same as the resulting paper author list and ordering but with occasional differences. 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 A10: Effect of Getting Scooped Prior to Deposit

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	0.023*	-0.156***	-0.072***	-0.136**	-0.038***
	(0.014)	(0.038)	(0.016)	(0.067)	(0.014)
Panel B. Base controls					
Scooped	-0.025**	-0.225***	-0.093***	-0.310***	-0.046***
	(0.011)	(0.040)	(0.016)	(0.061)	(0.015)
Panel C. PDS-Lasso selected controls					
Scooped	-0.020**	-0.216***	-0.087***	-0.284***	-0.042***
	(0.008)	(0.029)	(0.012)	(0.044)	(0.010)
Winner Y mean	0.842	-0.116	0.278	29.167	0.152
Observations	4,830	4,830	4,830	3,238	3,238

Notes: This table presents regression estimates of the scoop penalty restricting to scoops that occured prior to deposit. 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 A11: Structure Quality Balance in High- and Low-Reputation Match-ups

Matchup subsample	Loser structure quality (1)	Winner structure quality (2)	Difference: (lose - win) (3)	Std. error of difference (4)	Observations (5)
Panel A. Resolution (Å)					
High scoops High	2.566	2.496	0.070	(0.202)	724
Low scoops Low	2.362	2.258	0.104	(0.123)	491
High scoops Low	2.193	2.183	0.009	(0.058)	520
Low scoops High	2.148	2.155	-0.007	(0.050)	697
Panel B. R-free goodness-of-fit					
High scoops High	0.256	0.250	0.006	(0.004) *	701
Low scoops Low	0.246	0.243	0.003	(0.004)	486
High scoops Low	0.242	0.245	-0.003	(0.004)	512
Low scoops High	0.240	0.238	0.002	(0.004)	695

Notes: This table compares structure quality metrics of winning and losing projects in subsamples of races divided by team reputation as measured by predicted citations. Lower values of resolution and r-free represent better quality. Observations are at the structure level. Column 1 shows the means of the losing projects in the racing sample, and column 2 shows the means of the winning projects in the racing sample. Column 3 shows the difference between the losing and winning projects, and column 4 shows the heteroskedasticity-robust standard error of the difference.
**p<0.1, **p<0.05, ***p<0.01.*