Essays on the Economics of Science and Innovation

by

Carolyn Stein

A.B., Harvard University (2013)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

© Carolyn Stein, MMXXI. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Author . . . .
Department of Economics
May 14, 2021

Certified by . . . .
Heidi Williams
Charles R. Schwab Professor of Economics, Stanford University
Thesis Supervisor

Certified by . . . .
Amy Finkelstein
John & Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Certified by . . . .
Pierre Azoulay
International Programs Professor of Management
Thesis Supervisor

Accepted by . . . .
Amy Finkelstein
John & Jennie S. MacDonald Professor of Economics
Chair, Department Committee on Graduate Theses

Essays on the Economics of Science and Innovation

by

Carolyn Stein

Submitted to the Department of Economics
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis consists of three chapters on the economics of science and innovation. The first chapter studies whether the rewards for publishing first in science induce scientists to rush and produce lower-quality work; the second estimates the magnitude of these priority rewards. The third chapter studies whether male and female patent examiners treat patent applications submitted by women differently.

The first chapter, joint with Ryan Hill, investigates how competition to publish first and thereby establish priority impacts the quality of scientific research. We begin by developing a model where scientists decide whether and how long to work on a given project. When deciding how long to let their projects mature, scientists trade off the marginal benefit of higher quality research against the marginal risk of being preempted. The most important (highest potential) projects are the most competitive because they induce the most entry. Therefore, the model predicts these projects are also the most rushed and lowest quality. We test the predictions of this model in the field of structural biology using data from the Protein Data Bank (PDB), a repository for structures of large macromolecules. An important feature of the PDB is that it assigns objective measures of scientific quality to each structure. As suggested by the model, we find that structures with higher ex-ante potential generate more competition, are completed faster, and are lower quality. Consistent with the model, and with a causal interpretation of our empirical results, these relationships are mitigated when we focus on structures deposited by scientists who – by nature of their employment position – are less focused on publication and priority.

The second chapter, also joint with Ryan Hill, studies priority rewards in science. The scientific community assigns credit or "priority" to individuals who publish an important discovery first. We examine the impact of losing a priority race (colloquially known as getting "scooped") on subsequent publication and career outcomes. To do so, we take advantage of data from structural biology where the nature of the scientific process together with the Protein Data Bank — a repository of standardized research discoveries — enables us to identify priority races and their outcomes. We find that race winners receive more attention than losers, but that these contests are not winner-take-all. Scooped teams are 2.5 percent less likely to publish, are 18 percent less likely to appear in a top-10 journal, and receive 20 percent fewer citations. Getting scooped has only modest effects on academic careers. Finally, we document empirical evidence suggesting that the priority reward system reinforces inequality of attention in science.

The third chapter, joint with Jane Choi and Heidi Williams, considers the role of gender in the evaluation of patent applications submitted to the US Patent & Trademark Office (USPTO). Using the quasi-random assignment of patents to patent examiners, we document two facts. First, male examiners are more lenient overall than female examiners. Second, we find that patent examiner gender appears to have no effect on the evaluation of patent applications submitted by female inventors relative to male inventors. In other words, male examiners are *not* differentially stringent (or lenient) compared to their female counterparts when evaluating patent applications submitted by women. Our analysis is not able to assess whether the patent application evaluation system as a whole holds female inventors to a higher standard than their male counterparts. However, these results stand in contrast with evidence from other markets which has suggested that female reviewers may hold female applicants to higher standard than male reviewers.

JEL Classifications: O31, O34, I23,

Thesis Supervisor: Heidi Williams
Title: Charles R. Schwab Professor of Economics, Stanford University

Thesis Supervisor: Amy Finkelstein
Title: John & Jennie S. MacDonald Professor of Economics

Thesis Supervisor: Pierre Azoulay
Title: International Programs Professor of Management

Acknowledgments

Six years is a long time to spend working on a PhD. But looking back on it now, it truly has been a great and rewarding adventure. Many people deserve thanks for being there through the highs and inevitable lows.

First and foremost, I would like to thank my superlative dissertation committee. Heidi Williams is a brilliant economist and one of the kindest people I have ever known. She cared about me as a researcher, but more importantly, as a human being. There is no way I can repay her for the wisdom, advice, and encouragement she provided. But she is slowly creating a generation of economists eager to pay it forward, myself included. Amy Finkelstein reminded me of why I got myself into this game in the first place — it's supposed to be *fun*. Her insightful questions, enthusiasm, and humor elevate every room she walks into. Pierre Azoulay's warmth and generosity — with time, data, and espresso — propelled this dissertation forward. I think I have learned more about science from him than from anyone else (apologies to my high school science teachers). Lastly, Ryan Hill has served as a collaborator, friend, and de facto fourth advisor. His camaraderie down in the weeds of our papers made this whole dissertation vastly better (not to mention less lonely). And when it comes to the big picture — in research and in life — there is no one whose opinion I value more. I am so proud of the work we have done together.

I survived these six years thanks to my classmates turned friends. Joe Hazell carried my dead weight on his back through the macro sequence, all while politely insisting — between gasps — that I was "helping." Matt Lowe showed me the ropes early on, and I'm grateful for our continued friendship. Layne Kirshon has been there through thick and thin. I can count on him to both indulge my petty side and to help me wrestle with the problems that matter. Despite his love of jokes, he takes being a good friend seriously. Saturday spin and brunch was a wonderful ritual and the high point of many a week. Christina Patterson and Otis Reid dispensed wisdom and encouragement when it was needed most. I still look up to both of them. Maddie McKelway taught me what it means to believe deeply in one's own research. Pari Sastry is both a hilarious story-teller and an empathetic listener. Jane "Queen of Mamaleh's" Choi truly was the glue that held us all together. Finally, special thanks to Tamar Oostrom, friend and officemate extraordinaire. Thank you for sharing all the trials and tribulations that graduate school has thrown our way. You provided sympathy (never pity), occasional tough love, and a perspective different than my own that I have grown to deeply admire.

I thought my mediocre athletic career ended in high school, but I was thrilled to be proven wrong by the MIT Cycling Team. It was a joy to find friends who both pushed and supported me, on and off the bike. I will never forget the summer of 2019, with our mid-morning coffee rides where we laughed at the poor fools who had "real jobs." While I hope that cycling is a hobby I will take with me, I'm not sure I will ever be able to recreate that same magic.

Finally, thank you to my family for their love and encouragement. My father Jeremy was a source of great enthusiasm and healthy perspective. Thank you for reminding me that the real goal is to write something you're proud of; the rest is just gravy. My mother Anne has been my closest friend and biggest supporter for the past 30 years. She has a formidable combination of toughness and kindness that I have always tried to emulate. Despite living across the country, my sister Alison stepped up when I needed her most, and cheered me across the finish line. And my brother Jason never failed to keep things interesting.

. . . .

This thesis is dedicated to the memory of my grandfather, Eli Stein. Thank you for showing us all how it ought to be done — with joy, earnestness, and generosity. You set the bar impossibly high. We are all just reverting to the mean.

Contents

List of Figures

List of Tables

Chapter 1

Race to the Bottom: Competition and Quality in Science*

1.1 Introduction

Credit for new ideas is the primary currency of scientific careers. Credit allows scientists to build reputations, which translate to grant funding, promotion, and prizes (Tuckman and Leahey, 1975; Diamond, 1986; Stephan, 1996). As described by Merton (1957), credit comes — at least in part — from disclosing one's findings first, thereby establishing priority. It is not surprising, then, that scientists compete intensely to publish important findings first. Indeed, scientific history has been

*Contact: ryan.hill@kellogg.northwestern.edu, cstein@mit.edu. We are immensely grateful to our advisors Heidi Williams, Amy Finkelstein, and Pierre Azoulay for their enthusiasm and guidance. Stephen Burley, Scott Strobel, Aled Edwards, and Steven Cohen provided valuable insight into the field of structural biology, the Protein Data Bank, and the Structural Genomics Consortium. We thank David Autor, Jonathan Cohen, Peter Cohen, Glenn Ellison, Chishio Furukawa, Colin Gray, Sam Hanson, Ariella Kahn-Lang, Layne Kirshon, Matt Notowidigdo, Tamar Oostrom, Jonathan Roth, Adrienne Sabety, Michael Stepner, Alison Stein, Jeremy Stein, Sean Wang, Michael Wong, and participants in the MIT Labor and Public Finance lunches for their thoughtful comments and discussions. 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). All remaining errors are our own.

punctuated with cutthroat races and fierce disputes over priority (Merton, 1961; Bikard, 2020).[1] This competition and fear of pre-emption or "getting scooped" is not uniquely felt by famous scientists, but rather permeates the field. Older survey evidence from Hagstrom (1974) suggests that nearly two thirds of scientists have been scooped at least once in their careers, and a third of scientists reported being moderately to very concerned about being scooped in their current work. Newer survey evidence focusing on experimental biologists (Hong and Walsh, 2009) and structural biologists more specifically (Hill and Stein, 2020b) suggests that pre-emption remains common, and that the threat of pre-emption continues to be perceived as a serious concern.

Competition for priority has potential benefits and costs for science. Pressure to establish priority can hasten the pace of discovery and incentivize timely disclosure (Dasgupta and David, 1994). However, competition may also have a dark side. For years, scientists have voiced concerns that the pressure to publish quickly and preempt competitors may lead to "quick and dirty experiments" rather than "careful, methodical work" (Yong, 2018; Anderson et al., 2007). As early as the nineteenth century, Darwin lamented the norm of naming a species after its first discoverer, since this put "a premium on hasty and careless work" and rewarded "species-mongers" for "miserably describ[ing] a species in two or three words" (Darwin, 1887; Merton, 1957). More recently, journal editors have bemoaned what they view as increased sloppiness in science: "missing references; incorrect controls; undeclared cosmetic adjustments to figures; duplications; reserve figures and dummy text included; inaccurate and incomplete methods; and improper use of statistics" (Nature Editors, 2012). In other words, the faster pace of science has a cost: lower quality science. The goal of this paper is to consider the impact of competition on the quality of scientific work. We use data from the field of structural biology to empirically document that more competitive projects are executed with poorer quality. A variety of evidence supports a causal interpretation of competition

[1] To name but a few examples: Isaac Newton and Gottfried Leibniz famously sparred over who should get credit as the inventor of calculus. Charles Darwin was distraught upon receiving a manuscript from Alfred Wallace, which bore an uncanny resemblance to Darwin's (yet unpublished) *On the Origin of Species* (Darwin, 1887). More recently, Robert Gallo and Luc Montagnier fought bitterly and publicly over who first discovered the HIV virus. The dispute was so acrimonious (and the research topic so important) that two national governments had to step in to broker a peace (Altman, 1987).

leading researchers to rush to publication, as opposed to other omitted factors.

Economists have long studied innovation races, often in the context of patent or commercial R&D races. There is a large theoretical literature which considers the strategic interaction between two teams racing to innovate. These models have varied and often contradictory conclusions, depending on how the innovative process is modeled. For example, in models where innovation is characterized as a single, stochastic step, scientists will compete vigorously (Loury, 1979; Lee and Wilde, 1980). By contrast, if innovation is a step-by-step process, where experience matters and progress is observable, then the strategic behavior may be more nuanced (Fudenberg et al., 1983; Harris and Vickers, 1985, 1987; Aghion et al., 2001).[2] However, a common feature of these models is that innovation is binary: the team either succeeds or fails to invent. There is no notion that the invention may vary in its quality, depending on how much time or effort was spent. There are a few exceptions to this rule: Hopenhayn and Squintani (2016) and Bobtcheff et al. (2017) explicitly model the tension between letting a project mature longer (thereby improving its quality) versus patenting or publishing quickly (reducing the probability of being preempted). Tiokhin et al. (2020) develop a model of a similar spirit, where researchers choose a specific dimension of quality — the sample size. Studies with larger sample sizes take longer to complete, and so more competition leads to smaller sample sizes and less reliable science. Tiokhin and Derex (2019) test this line of thinking in a lab experiment.

Along these same lines, we develop a model of how competition spurred by priority races impacts the quality of scientific research. In our model, there is a deterministic relationship between the time a scientist spends on a project and the project's ultimate scientific quality. The scientist will choose how long to work on a given project with this relationship in mind. However, multiple scientists may be working on any given project. Therefore, there is always a latent threat of being pre-empted. The scientist who finishes and publishes the project first receives more credit and acclaim than the scientist who finishes second. This implies that a scientist deciding how long

[2] This literature has been primarily theoretical, though there are a few exceptions. Cockburn and Henderson (1994) study strategic behavior in drug development. Lerner (1997) studies strategic interaction between leaders and followers in the disk drive industry.

to work on her project must trade off the returns to continued "polishing" against the threat of potentially being scooped. As a result, the threat of competition leads to lower quality projects than if the scientist know she was working in isolation.

However, in a departure from the other models cited above, we embed this framework in a model where project entry is endogenous. This entry margin is important, because we allow for projects to vary in their ex-ante potential. To understand what we mean by "potential," consider that some projects solve long-standing open questions or have important applications for subsequent research. A scientist who completes one of these projects can expect professional acclaim, and these are the projects we consider "high-potential." Scientists observe this ex-ante project potential, and use this information to decide how much they are willing to invest in hopes of successfully starting the project. This investment decision is how we operationalize endogenous project entry. High-potential projects are more attractive, because they offer higher payoffs. As a result, researchers invest more trying to enter these projects. Therefore, the high-potential projects are more competitive, which in turn leads scientists to prematurely publish their findings. Thus, the key prediction of the model is that high-potential projects — those tackling questions that the scientific community has deemed the most important — are the projects that will also be executed with the lowest quality.

While the model provides a helpful framework, the primary contribution of this paper is to provide empirical support for the its claims. The idea that competition may lead to lower quality work is intuitive, and many scientists and journalists have speculated that this is the case (Fang and Casadevall, 2015; Vale and Hyman, 2016; Yong, 2018). However, systematically measuring the quality of scientific work is difficult. Consider the field of economics, for example — even with significant expertise, it is difficult to imagine "scoring" papers based on their quality of execution in a consistent, objective manner. Moreover, doing so at scale is infeasible.[3]

We make progress on the challenge of measuring scientific quality in the field of structural

[3] Some studies (Hengel, 2018) have used text analysis to measure a paper's readability as a proxy for paper quality, but such writing-based metrics fail to measure the underlying scientific content. Another strategy might be to use citations, but this fails to disentangle the quality of the project from the importance of the topic or the prominence of the author (Azoulay et al., 2013).

biology by using a unique data source called the Protein Data Bank (PDB). The PDB is a repository for structural coordinates of biological macromolecules (primarily proteins). The data are contributed by the worldwide research community, and then centralized and curated by the PDB. Importantly, every macromolecular structure is scored on a variety of quality metrics. At a high level, structural biologists are concerned with fitting three-dimensional structure models to experimental data, and so these quality metrics are measures of goodness of fit. They allow us to compare quality across different projects in an objective, science-based manner. To give an example of one of our quality metrics, consider refinement resolution, which measures the distance between crystal lattice planes. Nothing about this measure is subjective, nor can it be manipulated by the researcher.[4] Figure 1.1 shows the same protein structure solved at different refinement resolutions, to illustrate what a higher quality protein structure looks like.

The rich data in the PDB also allow us to construct additional variables necessary to test our model. The PDB groups identical proteins together into "similarity clusters" — proteins within the same cluster are identical or near-identical. By counting the number of deposits in a similarity cluster within a window of time after the first deposit, we can proxy for the competition researchers solving that structure likely faced. If we see multiple deposits of the same structure uploaded to the PDB in short succession, then researchers were likely engaged in a competitive race to deposit and publish first. Moreover, the PDB includes detailed timelines for most structures. In particular, they note the collection date (the date the researcher collected her experimental data) and the deposition date (roughly the date the researcher finished her manuscript). The difference in these two dates approximates the maturation period in the model.

The PDB has no obvious analog to project importance or potential, which is a pivotal variable in our model. Therefore, we use the rich meta-data in the PDB to construct our own measure. Rather than use ex-post citations from the linked publications as our measure of ex-ante potential (which might conflate potential with the ex-post quality of the work), we leverage the extensive structure-level covariates in the PDB to instead predict citations. These covariates include detailed

[4] Though of course researchers can "target" certain quality measures, in an attempt to reach a certain threshold.

characteristics of the protein known to the scientist before she begins working on the structure, such as the protein type, the protein's organism, the gene-protein linkage, and the prior number of papers written about the protein. Because the number of covariates is large relative to the number of observations, overfitting is a concern. To avoid this, we implement Least Absolute Shrinkage and Selection Operator (LASSO) to select our covariates, and then impute the predicted values.

We use our computed values of potential to test the key predictions of the model. Comparing structures in the 90^{th} versus 10^{th} percentile of the potential distribution, we find that high-potential projects induce meaningfully more competition, with about 30 percent more deposits in their similarity cluster. This suggests that researchers are behaving rationally by pursuing the most important (and highest citation-generating) structures. We then look at how potential impacts maturation and quality. We find that high-potential structures are completed about two months faster, and have quality measures that are about 0.7 standard deviations lower than low-potential structures. These results echo recent findings by a pair of structural biologists (Brown and Ramaswamy, 2007), who show that structures published in top general interest journals tend to be of lower quality than structures published in less prominent field journals.

However, a concern when interpreting these results is that competition and potential might be correlated with omitted factors that are also correlated with quality. In particular, we are concerned about complexity as an omitted variable — if competitive or high-potential structures are also more difficult to solve, our results may be biased. We take several approaches to address this concern. First, we investigate how long scientists spend working on their projects. If competitive and high-potential projects are more complex, we would expect researchers to spend *longer* on these projects in the absence of competition. However, we find the exact opposite: researchers spend *less* time on more competitive and higher potential projects. This suggests that complexity alone cannot explain our results, and that racing concerns must be at play. We also attempt to control for complexity directly. This has a minimal effect on the magnitude of our estimates.

To further probe this concern, we leverage another source of variation – namely, whether the protein was deposited by a structural genomics group. The majority of PDB structures are deposited by university- or industry-based scientists, both of which face the types of incentives we have described to publish early and obtain priority. In contrast, structural genomics (SG) researchers are federally-funded scientists with a mission to deposit a variety of structures, with the goal of obtaining better coverage of the protein-folding space and make future structure discovery easier. Qualitative evidence suggests these groups are less focused on publication and priority, which is consistent with the fact that only about 20 percent of SG structures ever appear in journal publications, compared to over 80 percent of non-SG structures.

Because the SG groups are less motivated by competition, we can contrast the relationships between potential and quality for SG structures versus non-SG structures. If complexity is correlated with potential, then this should be the case for both the SG and non-SG structures. Intuitively, by comparing the slopes across both groups, we thus "net out" the potential omitted variables bias. Consistent with competition acting as the causal channel, we find more negative relationships potential and quality among non-SG (i.e., more competitive) structures.

The fact that the most scientifically important structures are also the lowest quality intuitively seems suboptimal from a social welfare perspective. If project potential and project quality are complements (as we assume in the model), then a lack of quality among high-potential projects is particularly costly from a welfare perspective. Indeed, relative to a first-best scenario in which a social planner could dictate both investment and maturation to each researcher, the negative relationship between potential and quality does imply a welfare loss.

However, the monitoring and coordination costs make this type of scheme unrealistic from a policy perspective. Instead, we consider a different policy lever: allowing the social planner to dictate the division of credit between the first- and second-place teams. We consider this policy response in part because some journals have recently enacted "scoop protection" policies[5] explicitly aimed at increasing the share of credit awarded to teams who lose priority races. We then ask: with this single policy lever, can the social planner jointly achieve the optimal level of investment *and* maturation? Our model suggests no. While making priority rewards more equal

[5] These policies ask reviewers to treat recently scooped papers as if they are novel contributions; see Section 1.5.2 for more detail and examples.

does increase maturation periods toward the socially optimal level, it simultaneously may reduce investment levels. If the social planner values the project more than the individual researcher (consistent with the notion of research generating positive spillovers), then this reduced investment may be costly from a social welfare perspective. The optimal choice of how to allocate credit depends on the balance of these two forces, but ultimately may lead to a credit split that is lopsided. This in turn will lead to the observed negative relationship between potential and quality. Therefore, while this negative relationship tells us we are not at an unconstrained first-best, it cannot rule out that we are at a constrained second-best.

The remainder of this paper proceeds as follows. Section 2 presents the model. Section 3 describes our setting and data. Section 4 tests the predictions of the model, and Section 5 considers the welfare and policy implications. Section 6 concludes.

1.2 A Model of Competition and Quality in Scientific Research

The idea that competition for priority drives researchers to rush and cut corners in their work is intuitive. Our goal in this section is to develop a model that formalizes this intuition, and that generates additional testable predictions. Scientists in our model are rational agents, seeking to maximize the total credit or recognition they receive for their work. This is consistent with views put forth by Merton (1957) and Stephan (2012), though it stands in contrast with the idea that scientists are purely motivated by the intrinsic satisfaction derived from "puzzle-solving" (Hagstrom, 1965).

The model has two stages. In the first stage, a scientist decides how much effort to invest in starting the project. More investment at this stage translates to a higher probability of successfully starting the project. We call this the entry decision. When making this decision, a scientist will take into account each project's potential payoffs, and weigh these against the costs of investing. In the second stage, the scientist then decides how long to let the project mature. The choice of project maturation involves a tradeoff between higher project quality and an increasing probability of getting scooped.

We begin by solving the second-stage problem. In equilibrium, the researcher will know the probability that her competitor has entered the race, and she will have some prior on whether she is ahead of or behind her competitor. She will use these pieces of information to trade off marginal quality gains against the threat of pre-emption. The threat of competition will drive her to complete her work more quickly than if there were no competition (or if she were naïve to this threat). This provides us with our intuitive result that competition leads to lower scientific quality.

In the first stage, the researcher decides how much to invest in an effort to start the project, taking second-stage decisions as given. Projects have heterogenous payoffs, with important projects yielding more recognition than incremental projects. Scientists factor these payoffs into their investment decision. Therefore, the model generates predictions about *which* projects are the most competitive (i.e., induce the most entry) and thus the lowest quality. Because the highest expected payoff (i.e., the most important or "highest potential") projects offer the largest rewards, it is exactly these projects that our model predicts will have the most entry, competition, and rushing. This leads to the key insight from our model: the most ex-ante important projects are executed with the lowest quality ex-post. In the following sections, we formalize the intuition laid out above.

1.2.1 Preliminaries

Players. There are two symmetric scientists, i and j. Throughout, i will index an arbitrary scientist and j will index her competitor. Both scientists are working on the same project and only receive credit for their work once they have disclosed their findings through publication.

Timing, Investment, and Maturation. Time is continuous and indexed by t. From the perspective of each scientist, the model consists of two stages. In the first stage, scientist i has an idea. We denote the moment the idea arrives as the start time, or t_i^S. However, the scientist must pay an upfront cost in order to pursue the idea. At t_i^S, scientist i must decide how much to invest in starting the project. If she invests I_i, she has probability $g\left(I_i\right) \in [0,1]$ of successfully starting the project, where $g(\cdot)$ is an increasing, concave function and the Inada conditions hold. These assumptions reflect that more investment results in a higher probability of successfully entering a project, but that the returns are diminishing. I could be resources spent writing a grant proposal or trying to generate preliminary results. In our setting, a natural interpretation is that I represents the time and resources spent trying to grow a protein crystal.[6]

The second stage occurs if the scientist successfully starts the project. Then, she must decide how long to work on the project before publicly disclosing her findings. Let m_i denote the time she spends on the project, or the "maturation period." The project is then complete at $t_i^F = t_i^S + m_i$.

Payoffs and Credit Sharing. Projects vary in their ex-ante potential, which we denote P. For example, an unsolved protein structure may be relevant for drug development, and therefore a successful structure determination would be published in a top journal and be highly cited. We call this a "high-potential" protein or project.

Projects also vary in their ex-post quality, depending on how well they are executed. Quality is a deterministic function of the maturation period, which we denote $Q(m)$. Q is an increasing, concave function and the Inada conditions hold. Without loss of generality, we impose that $\lim_{m\to\infty} Q(m) = 1$. This facilitates the interpretation of quality as the share of the project's total potential that the researcher achieved. Then the total value of the project is the product of potential and quality.

The first team to finish a project receives a larger professional benefit (through publication, recognition, and citations) than the second team. To operationalize this idea as generally as possible, we say that the first team receives a reward equal to $\overline{\theta}$ times the project's value (through publication, recognition, and citations). The second team receives a smaller benefit, equal to $\underline{\theta}$ times the project's value. If r denotes the discount rate, then the present-discounted value of the project to the first-place finisher is given by:

$$\overline{\theta} e^{-rm} P Q(m). \tag{1.1}$$

[6] Indeed, the laborious process of growing protein crystals is almost universally a prerequisite for receiving a grant; the NIH typically to takes a "no crystals, no grant" stance on funding projects in structural biology (Lattman, 1996).

Similarly, the present-discounted value of the project to the second-place finisher is given by:

$$\underline{\theta} e^{-rm} PQ(m). \tag{1.2}$$

We make no restrictions on these weights, other than to specify that they are both positive and $\overline{\theta} \geq \underline{\theta}$. Importantly, we do not assume that the race is winner-take-all (i.e., $\underline{\theta} = 0$), as is common in the theoretical patent and priority race literature (for example, Loury (1979); Fudenberg et al. (1983); Bobtcheff et al. (2017)). Rather, consistent with empirical work on priority races (Hill and Stein, 2020b) and anecdotal evidence (Ramakrishnan, 2018), we allow for the second-place team to share some of the credit.

Information Structure. The competing scientists have limited information about their competitor's progress in the race. Scientist i does not observe I_j, and so she doesn't know the probability her opponent enters, although she will have correct beliefs about this probability in equilibrium. In addition, she does not know her competitor's start time t_j^S. All she knows is that it is uniformly distributed around her own start time. In other words, she believes that $t_j^S \sim \text{Unif}\left[t_i^S - \Delta, t_i^S + \Delta\right]$ for some $\Delta > 0$. Figure 1.2 summarizes the model setup.

1.2.2 Maturation

We begin by solving the second stage problem of the optimal maturation delay, taking the first stage investment as given. In other words, we explore what the scientist does once she has successfully entered the project, and all her investment costs are already sunk. Our setup is similar to the approach of Bobtcheff et al. (2017), but an important distinction is that we only allow the project's value to depend on the maturation time m, and not on calendar time t. This simplifies the second stage problem, and allows us to embed the solution into the first stage investment decision in a more tractable way.

The No Competition Benchmark

We start by solving for the optimal maturation period of a scientist who knows that she is not competing for priority. Alternatively, we could consider this the behavior of a naive scientist, who does not recognize the risk of being scooped. This will serve as a useful benchmark once we re-introduce the possibility of competition.

Without competition, the scientist simply trades off the marginal benefit of further maturation against the marginal cost of time discounting. The optimal maturation delay $m_i^{NC^*}$ is given by

$$m_i^{NC^*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ(m_i) \right\}. \tag{1.3}$$

Taking the first-order condition and re-arranging (dropping the *i* subscripts for convenience) yields

$$\frac{Q'\left(m^{NC^*}\right)}{Q\left(m^{NC^*}\right)} = r. \tag{1.4}$$

In other words, the scientist will stop work on the project and publish the paper when the rate of improvement equals the discount rate.

Adding Competition

We continue to study the problem of the scientist who has already entered the project and already sunk the investment cost. However, now we allow for the possibility of a competitor. We call the solution to this problem the optimal maturation period with competition, and denote it $m_i^{C^*}$. Scientist i believes that her competitor has also entered the project with some probability $g(I_j^{C^*})$, where $I_j^{C^*}$ is j's equilibrium first-stage investment. However, because investment is sunk in the first stage, we can treat $g(I_j^{C^*})$ as a parameter (simply g) in this part of the model to simplify the notation.

While scientist i knows the probability that j entered the project, she does not know her potential competitor's start time, t_j^S. As described above, her prior is that t_j^S is uniformly distributed around her own start time. Let $\pi(m_i, m_j)$ denote the probability that scientist i wins the race, conditional on successfully entering. This can be written as

$$\pi(m_i, m_j) = (1-g) + gPr(t_i^F < t_j^F) = (1-g) + gPr(t_i^S + m_i < t_j^S + m_j). \tag{1.5}$$

The first term represents the probability that j fails to enter (and so i wins for sure), and the second term is the probability that j enters, but i finishes first. The optimal maturation period is given by

$$m_i^{C^*} \in \underset{m_i}{\arg\max} \left\{ e^{-rm_i} PQ(m_i) \left[\pi(m_i, m_j)\overline{\theta} + (1 - \pi(m_i, m_j))\,\underline{\theta} \right] \right\}. \tag{1.6}$$

The term outside the square brackets represents the full present discounted value of the project. The terms inside the brackets denote i's expected share of the credit, conditional on i successfully starting the project. The product of these two terms is scientist i's expected payoff conditional on successfully starting the project. Taking the first-order condition of Equation 1.6 implicitly defines scientist i's best-response function, which depends on m_j and other parameters:

$$\frac{Q'\left(m_i^{C^*}\right)}{Q\left(m_i^{C^*}\right)} = r + \frac{1}{\Delta\left(\frac{2\overline{\theta} - g(\overline{\theta} - \underline{\theta})}{g(\overline{\theta} - \underline{\theta})}\right) + m_j - m_i^{C^*}}. \tag{1.7}$$

If we look for a symmetric equilibrium, this yields Proposition 1 below.

Proposition 1. *Assume that first stage equilibrium investment is equal for both researchers, i.e., $I_i^{C^*} = I_j^{C^*} = I^{C^*}$. Further assume that Δ is sufficiently large. Then in the second stage, there is a unique symmetric pure strategy Nash equilibrium where $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ and m^{C^*} is implicitly defined by*

$$\frac{Q'\left(m^{C^*}\right)}{Q\left(m^{C^*}\right)} = r + \frac{g(I^{C^*})(\overline{\theta} - \underline{\theta})}{\Delta\left(2\overline{\theta} - g(I^{C^*})(\overline{\theta} - \underline{\theta})\right)}. \tag{1.8}$$

Proof. See Appendix A.1.1. □

Because $Q(m)$ is increasing and concave, we know Q'/Q is a decreasing function. Therefore, by comparing Equations 1.4 and 1.8, we can see that $m^{NC} > m^C$. In other words, competition leads to shorter maturation periods. This shortening is exacerbated when the difference between $\overline{\theta}$

and $\underline{\theta}$ is large (priority rewards are more lopsided), Δ is small (competitors start the projects close together, and so the "flow risk" of getting scooped is high), or when g is close to one (the entry of a competitor is likely). On the other hand, if $\overline{\theta} = \underline{\theta}$ (first and second place share the rewards evenly), $\Delta \to \infty$ (competition is very diffuse, so the "flow risk" of getting scooped is low), or $g = 0$ (the competitor doesn't enter), then we recover the no competition benchmark.

1.2.3 Investment

In the first stage, scientist i decides how much she would like to invest in hopes of starting the project. Let I_i denote this investment, and let $g\left(I_i\right)$ be the probability she successfully enters the project, where g is an increasing, concave function. With probability $1 - g\left(I_i\right)$ she fails to enter the project, and her payoff is zero. With probability $g\left(I_i\right)$ she successfully enters the project, and begins work at t_i^S. Once she enters, there are two ways she can win the priority race: first, if her competitor fails to enter, she wins for sure. Second, if her competitor enters but she finishes first, she also wins. In either case, she gets a payoff of $\overline{\theta} PQ\left(m_i^C\right)$. On the other hand, if her competitor enters and she loses, her payoff is $\underline{\theta} PQ\left(m_i^C\right)$. Putting these pieces together (noting that in equilibrium, if both i and j enter, they are equally likely to win) and re-arranging, the optimal level of investment is

$$I_i^{C^*} \in \arg\max_{I_i} \left\{ g\left(I_i\right) e^{-rm_i^{C^*}} PQ\left(m_i^{C^*}\right) \left[\overline{\theta} - \frac{1}{2} g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right] - I_i \right\}. \tag{1.9}$$

Taking the first-order condition of Equation 1.9 implicitly defines scientist i's best-response function, which depends on I_j, $m_i^{C^*}$, and other parameters:

$$g'(I_i^{C^*}) = \frac{1}{e^{-rm_i^{C^*}} PQ\left(m_i^{C^*}\right)\left[\overline{\theta} - \frac{1}{2} g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{1.10}$$

If we look for a symmetric equilibrium, this yields Proposition 2 below.

Proposition 2. *Assume that researchers are playing a symmetric pure strategy Nash equilibrium when selecting m in the second stage. Then, in the first stage, there is a unique symmetric pure* *strategy Nash equilibrium where $I_i^C = I_j^C = I^C$ and I_i^C is implicitly defined by*

$$g'(I^{C^*}) = \frac{1}{e^{-rm^{C^*}} PQ\left(m^{C^*}\right)\left[\overline{\theta} - \frac{1}{2} g\left(I^{C^*}\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{1.11}$$

Together with Proposition 1, this shows that there is a unique symmetric pure strategy Nash equilibrium for both investment and maturation.

Proof. See Appendix A.1.1. □

Equations 1.11 and 1.8 together define the optimal investment level and maturation period for scientists when entry into projects is endogenous. This allows us to prove three key results.

Proposition 3. *Consider an exogenous increase in the probability of project entry, g. This corresponds to an increase in competition, because it makes racing more likely. When projects become more competitive, the maturation period becomes shorter and projects become lower quality. In other words, $\frac{dm^{C^*}}{dg} < 0$ and $\frac{dQ(m^{C^*})}{dg} < 0$.*

Proof. See Appendix A.1.1. Scientist i selects m_i^C by considering the probability that her competitor enters $g(I_j)$. If this probability goes up, she will choose a shorter maturation period which results in lower quality. □

Proposition 4. *Higher potential projects generate more investment and are therefore more competitive. In other words, $\frac{dI^{C^*}}{dP} > 0$ and $\frac{dg(I^{C^*})}{dP} > 0$.*

Proof. See Appendix A.1.1. Scientist i will invest more to enter a high-potential project. Her competitor will do the same. In equilibrium, high-potential projects are more likely to result in priority races. □

Proposition 5. *Higher potential projects are completed more quickly, and are therefore of lower quality. In other words, $\frac{dm^{C^*}}{dP} < 0$ and $\frac{dQ(m^{C^*})}{dP} < 0$.*

Proof. This comes immediately from Propositions 3 and 4, by applying the chain rule. □

1.3 Structural Biology and the Protein Data Bank

This section provides some scientific background on structural biology and describes our data. We take particular care to explain how we map key variables from our model into measurable objects in our data. Our empirical work focuses on structural biology precisely because there is such a clean link between our theoretical model and our empirical setting. Section 1.3.1 provides an overview of the field of structural biology, while sections 1.3.2 and 1.3.3 describe our datasets. Section 1.3.4 describes how we construct our primary analysis sample and provides summary statistics. Appendix A.2 provides additional detail on our data sources and construction.

1.3.1 Structural Biology

Structural biology is the study of the three-dimensional structure of biological macromolecules, including deoxyribonucleic acid (DNA), ribonucleic acids (RNA), and most commonly, proteins. Understanding how macromolecules perform their functions inside of cells is one of the key themes in molecular biology. Structural biologists shed light on these questions by determining the three-dimensional arrangement of a protein's atoms.

Proteins are composed of building blocks called amino acids. These amino acids are arranged into a single chain, which folds up onto itself, creating a three-dimensional structure. While the shape of these proteins is of great interest to researchers, the proteins themselves are too small to observe directly under a microscope.[7] Therefore, structural biologists use experimental data to propose three-dimensional models of the protein shape to better understand biological function.

Structural biology has several unique features that make it amenable for our purposes (see Section 1.3.1 below), but it is also an important field of science. Proteins contribute to nearly every process inside the body, and understanding the shape and structure of proteins is critical to understanding how they function. Moreover, many heritable diseases — such as sickle-cell anemia,

[7] Recent developments in the field of cryo-electron microscopy now allow scientists to observe larger structures directly (Bai et al., 2015). However, despite the recent growth in this technique, fewer than five percent of PDB structures deposited since 2015 have used this method.

Alzheimer's disease, and Huntington's disease — are the direct result of protein mis-folding. Protein structures also play a critical role in drug development and vaccine design (Westbrook and Burley, 2018). Protease inhibitors, a type of antiretroviral drug used to treat HIV, are one important example of successful structure-based drug design (Wlodawer and Vondrasek, 1998). The rapid discovery and deposition of the SARS-CoV-2 spike protein structure has proven to be a key input in the ongoing development of COVID-19 vaccines and therapeutics (Wrapp et al., 2020). Over a dozen Nobel prizes have been awarded for advances in the field (Martz et al., 2019).

Why Structural Biology?

Our empirical work focuses on the field of structural biology for several reasons. First, and most importantly, structural biology has unique measures of objective project quality. Scientists in this field work to solve the three-dimensional structure of known proteins, and there are several measures of how precise and correct their solutions are. We will discuss these measures in the subsequent sections, but we want to highlight the importance of this feature: it is difficult to imagine how one might objectively rank the quality (distinct from the importance or relevance) of papers in other fields, such as economics or mathematics. Our empirical work hinges on the fact that structural biologists have developed unbiased, science-based measures of structure quality.

Second, we can measure competition and racing behavior using biological similarity measures and project timelines. By comparing the amino acid sequences of different proteins, we can detect when two proteins are similar or identical to one another. This allow us to find projects that focus on similar proteins, while the timeline data allows us to determine if researchers were working on these projects contemporaneously. Together, this allows us to determine which structures faced heavy competition while the scientists were doing their research.

Third, the PDB contains rich descriptive data on each protein structure. For each structure, we observe covariates like the detailed protein classification, the taxonomy / organism, and the associated gene. Together, these characteristics allow us to develop measures of the protein's importance, based purely on ex-ante characteristics — a topic we discuss in more detail in Section

1.4.1.

Solving Protein Structures Using X-Ray Crystallography

How do scientists solve protein structures? Understanding this process is important for interpreting the various quality measures used in our analysis. We focus on proteins solved using a technique called x-ray crystallography. The vast majority (89 percent) of structures are solved using this method.

X-ray crystallography broadly consists of three steps (see Figure 1.3). Individual proteins are too small to analyze or observe directly. Therefore, as a first step, the scientist must distill a concentrated solution of the protein into orderly crystals. Growing these crystals is a slow and difficult process, often described as “more art than science” (Rhodes, 2006) or at times simply “dumb luck” (Cudney, 1999). Success typically comes from trial and error, and a healthy dose of patience.[8]

Next, the scientist will bring her crystals to a synchrotron facility and subject the crystals to x-ray beams. The crystal’s atom planes will diffract the x-rays, leading to a pattern of spots called a “diffraction pattern.” Better (i.e., larger and more uniform) crystals yield superior diffraction patterns and improved resolution. If the scientist is willing to spend more time improving her crystals — by repeatedly tweaking the temperature or pH conditions, for example — she may be rewarded with better experimental data.

Finally, the scientist will use these diffraction patterns to first build an electron density map, and then an initial atomic model. Building the atomic model is an iterative process: the scientist will compare simulated diffraction data from her model to her actual experimental data and adjust

[8] As Cudney colorfully explains: “How many times have you purposely designed a crystallization experiment and had it work the first time? Liar. Like you really sit down and say ‘I am going to use pH 6 buffer because the p1 of my protein is just above 6 and I will use isopropanol to manipulate the dielectric constant of the bulk solvent, and add a little BOG to mask the hydrophoic interactions between sample molecules, and a little glycerol to help stabilize the sample, and [a] pinch of trimethylamine hydrochloride to perturb water structure, and finally add some tartate to stabilize the salt bridges in my sample.’ Right...Finding the best crystallization conditions is a lot like looking for your car keys; they’re always the last place you look” (Cudney, 1999).

the model until she is satisfied with the goodness of fit. This process is known as “refinement,” and depending on the complexity of the structure can take an experienced crystallographer anywhere from hours to weeks to complete. Refinement can be a “tedious” process (Strasser, 2019), and involves “scrupulous commitment to the iterative improvement and interpretation of the electron density maps” (Minor et al., 2016). Refinement is a back-and-forth process of trying to better fit the proposed structural model to the experimental data, and the scientist has some discretion in when she decides the final model is “good enough” (Brown and Ramaswamy, 2007). More time and effort spent in this phase can translate to better-quality models.

1.3.2 The Protein Data Bank

Our primary data source is the Protein Data Bank (PDB). The PDB is a worldwide repository of biological macromolecules, 95 percent of which are proteins.[9] It was established in 1971 with just seven entries, and today contains upwards of 150,000 structures. Since the late 1990s, the vast majority of journals and funding agencies have required that scientists deposit their findings in the PDB (Barinaga, 1989; Berman et al., 2000, 2016; Strasser, 2019). Therefore, the PDB represents a near-universe of macromolecule structure discoveries. For more detail on both the history and mechanics of depositing in the PDB, see Berman et al. (2000, 2016). Below, we describe the data collected by the PDB. The primary unit of observation in the PDB is a structure, representing a single protein. Most variables in our data are indexed at the structure level.[10]

Measuring Quality

The PDB provides a myriad of measures intended to assess quality. These quality measures were developed by the X-Ray Validation Task of the PDB in 2008, in an effort to increase the overall social value of the PDB (Read et al., 2011). Validation serves two purposes: it can detect large

[9] Because the vast majority of structures deposited to the PDB are proteins, we will use the terms “structure” and “protein” interchangeably throughout this paper.

[10] Some structures are composed of multiple “entities,” and some variables are indexed at the entity level. We discuss this in more detail in Appendix A.2.

structure errors, thereby increasing overall user confidence, and it makes the PDB more useful and accessible for scientists who do not possess the specialized knowledge to critically evaluate structure quality. Below, we describe the three measures that we use in our empirical analysis. We selected these three because they are scientifically distinct and have good coverage in our data. We also combine these three measures into a single quality index, described below. Together, these measures map exactly to Q in our model. Importantly, they score a project on its quality of execution, rather than on its importance or relevance.

An important feature of these measures is that they are all either calculated or independently validated by the PDB, leaving no scope for misreporting or manipulation by authors. Since 2013, the PDB has required that x-ray structures undergo automatic validation reports prior to deposition. These reports take the researcher's proposed model and experimental data as inputs, and use a suite of software programs to produce and validate various quality measures. In 2014, the PDB ran the same validation reports retrospectively on all structures that were already in the PDB (Worldwide Protein Data Bank, 2013), so we have full historical coverage for these quality measures. Appendix Figure A6 provides a snapshot from one of these reports.

Refinement resolution. Refinement resolution measures the smallest distance between crystal lattice planes that can be detected in the diffraction pattern. It is somewhat analogous to resolution in a photograph. Resolution is measured in angstroms (Å), which is a unit of length equal to 10^{-10} meters. Smaller resolution values are better, because they imply that the diffraction data is more detailed. This in turn allows for better electron density maps, as shown in Figure 1.1. At resolutions less than 1.5Å, individual atoms can be resolved and structures have almost no errors. At resolutions greater than 4Å, individual atomic coordinates are meaningless and only secondary structures can be determined. As described in Section 1.3.2, scientists can improve resolution by spending time improving the quality of the protein crystals and by fine-tuning the experimental conditions during x-ray exposure. In our main analysis, we will standardize refinement resolution so that the units are in standard deviations and higher values represent better quality.

R-free. The R-free is one of several residual factors (i.e., R-factors) reported by the PDB. In general, R-factors are a measure of agreement between a scientist's structure model and experimental data. Similar to resolution, lower values are better. An R-factor of zero means that the model fits the experimental data perfectly; a random arrangement of atoms would give an R-factor of about 0.63. Two R-factors are worth discussing in more detail: R-work and R-free. When fitting a model, the scientist will set aside about ten percent of the data for cross-validation. R-work measures the goodness of fit in the non-cross-validation sample. R-free measures the goodness of fit in the cross-validation sample. R-free is our preferred R-factor, because it is less likely to suffer from overfitting (Goodsell, 2019a; Brünger, 1992). Most crystallographers agree it is the most accurate measure of model fit (Read et al., 2011).

While an R-free of zero is the theoretical best that the scientist could attain, in reality R-free is constrained by the resolution. Structures with worse (i.e., higher) resolution have worse (i.e., higher) R-free values. As a rule of thumb, models with a resolution of 2Å or better should have an R-free of $(\text{resolution}/10+0.05)$ or better. In other words, if the resolution is 2Å, the R-free should not exceed 0.25 (Martz and Hodis, 2013). A researcher who spends more time refining her model can attain better R-free values. In our main analysis, we will standardize R-free so that the units are in standard deviations and higher values represent better quality.

Ramachandran outliers. Ramachandran outliers are one form of outliers calculated by the PDB. Protein chains tend to bond in certain ways (at specified angles, with atoms at specified distances, etc.). Violations of these "rules" may be features of the protein, but typically they represent errors in the model. At a high level, most outlier measures calculate the percent of amino acids that are conformationally unrealistic. Ramachandran outliers (Ramachandran et al., 1963) focus on the angles of the protein's amino acid backbone, and flag instances where the bond angles are too small or large. Again, in our main analysis, we will standardize Ramachandran outliers so that the units are in standard deviations and higher values represent better quality.

Quality index. Finally, we combine the three measures above into a single quality index. All three measures are correlated, with correlation coefficients in the 0.4 to 0.6 range (see Appendix Table A1). We create the index by adding all three standardized quality measures and then standardizing the sum.

Measuring Maturation

We refer to the time the scientist spends working on a protein structure as the "maturation" period, corresponding to *m* in our model. We are interested in whether competition reduces structure quality via rushing, i.e., shortening the maturation period. In most scientific fields, it would be impossible to measure the time researchers spend on each project, but the PDB metadata provides unique insight about project timelines.

For most structures, the PDB collects two key dates which allow us to infer the maturation period: the collection date and the deposition date. The collection date is self-reported and date corresponds to the date that the scientist subjected her crystal to x-rays and collected her experimental data. The deposition date corresponds to the date that the scientist deposited (i.e., uploaded) her structure to the PDB. Because journals require evidence of deposition before publishing articles, the deposition date corresponds roughly to when the scientist submitted her paper for peer review.[11] The timespan between these two dates represents the time it takes the scientist to go from the raw diffraction data to a completed draft (the "diffraction pattern" stage to the "completed structure" stage in Figure 1.3). In other words, it is the time spent determining the protein's structure, refining the structure, and writing the paper. However, note that this maturation period only includes time spent working on the structure once the protein was successfully crystallized and taken to a synchrotron. Anecdotally, crystallizing the protein (the first step in Figure 1.3) can be the most time-consuming step. Because we do not observe the date the scientist began attempting

[11] Rules governing when a researcher must deposit her structure to the PDB have changed over time. However, following an advocacy campaign by the PDB in 1998, the NIH as well as *Nature* and *Science* began requiring that authors deposit their structures prior to publication (Campbell, 1998; Bloom, 1998; Strasser, 2019). Other journals quickly followed suit. We code the maturation time as missing if the structure was deposited prior to 1999 to ensure a clear interpretation of this variable.

to crystallize the protein, we cannot measure this part of the process. Therefore our maturation variable does not capture the full interval of time spent working on a given project.

Measuring Investment

There is no clear way to measure the total resources that a researcher invests in starting a project using data from the PDB. However, one scarce resource that scientists must decide how to allocate across different projects is lab personnel. We can measure this, because every structure in the PDB is assigned a set of "structure authors." We take the number of structure authors as one measure of resources invested in a given project. In addition, we can also count the number of paper authors on structures with an associated publication. To understand the difference between structure authors and paper authors, note that structure authors are restricted to authors who directly contributed to solving the protein structure. Therefore, the number of structure authors tends to be smaller than the number of paper authors on average (about five versus about seven in our main analysis sample), because paper authors can contribute in other ways, such as by writing the text or performing complementary analyses. Appendix Figure A7 shows the histogram of the difference between the number of paper authors and structure authors. While we view the number of structure authors as a cleaner measure of investment, because these authors contributed directly to solving the protein structure, we will use both in our analysis.

Measuring Competition

Our measure of competition leverages the fact that the PDB assigns each protein to a "similarity cluster" based on the protein's amino acid sequence. Two identical or near-identical proteins will both belong to the same similarity cluster.[12] Therefore, we are able to count the number of PDB deposits within a similarity cluster, which gives some measure of the "crowdedness" or competition

[12] More specifically, there are different "levels" of sequence similarity clusters. Two proteins belonging to the same 100 percent similarity cluster share 100 percent of their amino acids in an identical order. Two proteins belonging to the same 90 percent similarity cluster share 90 percent of their amino acids in an identical order. We use the 100 percent cluster. For more detail, see Hill and Stein (2020b).

for a given protein.

However, these deposits may not represent concurrent discoveries or races if they were deposited long after the first structure was deposited. Therefore, we instead count the number of deposits in the PDB that appear within the first two years of when the first structure was deposited. We choose two years as our threshold, because the average maturation period is 1.75 years on average. Therefore, we believe that structures deposited within two years of the first structure likely represent concurrent work. This two year cutoff is admittedly ad hoc, and so we construct some alternative competition measures and show in Appendix A.3 that our results are not sensitive to this particular cutoff.

This measure is meant to proxy for g, the equilibrium probability that a competitor has also started the project. However, we cannot directly measure the ex-ante probability of competition, and so instead we measure ex-post realized competition. This implies that our measure of competition will be noisy estimate of g — the researcher's perceived competition — which is the relevant variable for dictating researcher decision-making and behavior. We flag this measurement issue because it will lead to attenuation bias if this proxy is used as an independent variable in a regression.

Complexity Covariates

Proteins can be difficult to solve because (a) they are hard to crystallize, and (b) once crystallized, they are hard to model. In general, predicting whether a protein will be easy or hard to crystallize is a difficult task. Researchers have failed to discover obvious correlations between crystallization conditions and protein structure or family (Chayen and Saridakis, 2008). Often, a single amino acid can be the difference between a structure that forms nice, orderly crystals and one that evades all crystallization efforts. However, as a general rule, larger and "floppier" proteins are more difficult to crystallize than their smaller and more rigid counterparts (Rhodes, 2006). Moreover, since these larger proteins are more complex, with more folds, they are harder to model once the experimental data are in hand. Therefore, despite the general uncertainty of protein crystallization, size is a predictor of difficulty.

The PDB contains several measures of structure size, which we use as covariates to control for complexity. These include molecular weight (the structure's weight), atom site count (the number of atoms in the structure), and residue count (the number of amino acids the structure contains). Because these variables are heavily right-skewed, we take their logs. We then include these three variables and their squares as complexity controls.[13]

Other Descriptive Covariates

For each structure, the PDB includes detailed covariates describing the molecule. Some of these covariates are related to structure classification — these include the macromolecule type (protein, DNA, or RNA), the molecule's classification (transport protein, viral protein, signaling protein, etc.), the taxonomy (organism the structure comes from), and the gene that expresses the protein. We use these detailed classification variables to estimate a protein's scientific relevance, a topic discussed in more detail in Section 1.4.1.

1.3.3 Other Data Sources

Web of Science

The Web of Science links over 70 million scientific publications to their respective citations.[14] Our version of these data start in 1990 and end in 2018. Broadly, we are able to link the Web of Science citations data to the PDB using PubMed identifiers, which are unique IDs assigned to research papers in the medical and life sciences by the United States National Library of Medicine. The PDB manually links all structures to the published paper that "debuts" the structure, and includes

[13] A key exception to the discussion above is membrane proteins. Membrane proteins are embedded in the lipid bilayer of cells. As a result, membrane proteins (unlike other proteins) are hydrophobic, meaning they are not water-soluble. This makes them exceedingly difficult to purify and crystallize (Rhodes, 2006; Carpenter et al., 2008). This has made membrane protein structures a rarity in the PDB — although membrane proteins comprise nearly 25 percent all proteins (and an even higher share of drug targets), they make up just 1.5 percent of PDB structures. We drop membrane proteins from our sample, though their inclusion or exclusion do not meaningfully impact our results.

[14] The Web of Science is owned by Clarivate Analytics since 2016.

the PubMed ID in this linkage. The Web of Science includes a paper-PubMed ID crosswalk. This allows us to link the Web of Science to the PDB.

We then use these linked data to compute citation counts for PDB linked papers. We compute citations by counting citations in the three years following publication,[15] and exclude any self-citations.[16] By restricting to citations in the three years since publication (rather than total cumulative citations) we avoid the problem that older papers have had more time to accumulate citations. Note that these citation variables are unique at the *paper* level, rather than at the structure level. Structures are linked to papers in a many-to-one fashion. In other words, while some papers only have one affiliated structure, other papers may have multiple affiliated structures. We discuss how we handle multiple matching of structures to a single paper in Section 1.3.4.

UniPROT Knowledgebase

The UniPROT Knowledgebase is a database of over 120 million proteins from all species and branches of life (The UniProt Consortium, 2019). The PDB only contains entries for proteins whose structures have been solved. Therefore, the UniPROT data represents a superset of proteins found in the PDB. For each protein, the data contain the amino acid sequence, protein name, and PubMed IDs for all of the academic papers that reference the protein. Importantly, each entry also includes a PDB ID if the protein has an associated structure in the PDB. This allows us to link the UniPROT data to the PDB.

Scientists often study and publish papers about proteins long before their structures are solved. Therefore, we can count the number of papers that were published about a protein *prior* to the protein's structure publication. We view this as a measure of ex-ante demand for the protein's

[15] We only count citations that have been assigned a PubMed ID. Because structural biology falls squarely in the medical and life sciences, this restriction has little impact.

[16] Following Wuchty et al. (2007), we define a self-citation as any citation citation where a common name exists in the authorship of both the cited and the citing papers. Common names are defined as when the first initial and last name match. This method can also eliminate citations where the authors are different people but share the same name. However, Wuchty et al. (2007) perform Monte Carlo simulations on the data, and find that such errors occur in less than 1 of every 2,000 citations. Thus, any errors introduced by this procedure appear negligible.

structure. In other words, if a protein is heavily studied before anyone has solved and released its structure, there is probably more interest in the structure. We use this to help proxy for a protein's importance, a topic discussed in more detail in Section 1.4.1.

DrugBank

DrugBank is a comprehensive database containing information on both drugs, their mechanisms, their interactions, and their protein targets. It is widely used by researchers, physicians, and the drug industry (Wishart et al., 2018). The current release contains over 11,000 drugs, including about 2,600 approved drugs (approved by the FDA, Health Canada, EMA, etc.), 6,000 experimental (i.e., pre-clinical) drugs, and about 4,000 investigational drugs (in Phase I/II/III human trials).[17] Importantly for us, beyond just linking to the target protein, DrugBank provides the PDB ID(s) for any target structure that has been deposited in the PDB. This allows us to link structures to the drugs that target them.

1.3.4 Sample Construction

We begin with the full sample of 128,876 PDB structures that were deposited and solved using x-ray crystallography between 1971 and 2018. These structures are linked to 63,809 unique publications. From here, we make a series of sample restrictions to construct our final analysis sample. Key variables in our data are indexed at two distinct levels: the structure level and the paper level. Therefore, we start by restricting to publications with just one structure. This leaves us with 35,625 structures linked to 35,625 papers (or "projects" in the case of structures without an associated publication).[18] The resulting data have a one-to-one mapping between a given paper and structure. This restriction allows us to assign paper-level characteristics, such as expected citations, directly to individual structure deposits in the PDB.

[17] Some drugs fall into more than one category.

[18] For structures without an associated publication, we attempt to predict whether the structure would have have been the only structure in a paper *had it been published*. See Appendix A.2 for details. Appendix Figure A8 suggests that we are able to correctly classify these structures the majority of the time.

Because we are interested in the behavior of scientists who are potentially racing, we further restrict our analysis sample to new structure discoveries. In other words, we drop PDB deposits if a structure of the protein had previously been deposited. In practice, we use the similarity clusters and only keep the first protein to be released in each cluster. This leaves us with 25,620 structures. Finally, we drop structures that are missing any of our three quality measures. We also drop membrane proteins.[19] This leaves us with a final sample of 21,951 structures.

Table 1.1 provides summary statistics for both the full sample and our analysis sample. Panel A presents structure-level statistics and Panel B presents paper-level statistics. Although our analysis sample comprises a small subset of the total structures, it appears fairly representative of the full sample. There are a few exceptions to this claim. The maturation period (years between collection and deposit) is shorter in the analysis sample, likely because we focus on the first deposit of a given protein, and so racing is more likely. Competition (deposits per similarity cluster within two years) is smaller in the analysis sample, but this occurs mechanically because we drop all deposits after the first structure deposition.[20] Similarly, the number of UniPROT papers (i.e., papers published prior to the first structure discovery) is lower in the analysis sample because there are more UniPROT papers for structures in crowded clusters. For more detail on the full distributions of our key outcome variables, see the histograms in Appendix Figure A9.

1.4 Testing the Model: Empirical Strategy and Results

In this section, we test the predictions laid out by the model in Section 1.2. We start by focusing on Propositions 4 and 5, which rely on cross-sectional variation in potential. Propositions 4 states that high-potential projects should generate more investment and therefore more competition. Proposition 5 states that high-potential projects should therefore be more rushed and lower quality.

[19] We drop membrane proteins because they are exceptionally difficult to purify and crystallize (Rhodes, 2006; Carpenter et al., 2008). This exclusion only drops 357 structures and does not meaningfully impact our results.

[20] So in a cluster with 100 deposits we drop 99, while in a cluster with 2 deposits, we only drop 1. This will mechanically lower the average number of deposits per cluster.

We provide a variety of evidence which points to increased competition and rushing — rather than other omitted factors — as the primary channel.

Finally, we return to Proposition 3, which states that more competitive projects (projects at higher risk of having multiple teams competing simultaneously) are more likely to be rushed and lower quality. We do not have a clean measure of ex-ante competition — as discussed in Section 1.3.2, we only measure ex-post realized competition. This noise will lead to attenuation bias in our estimates. However, the model sets up a natural instrumental variables specification: we can instrument for competition with project potential. Proposition 4 functions as the first stage, while Proposition 5 is the reduced form.

1.4.1 Defining Project Potential

Before we can begin testing the model, we need to define an empirical analog to the project potential variable in our model. Project potential captures the notion that ex-ante, some proteins are likely to be heavily cited. Scientists are usually aware of which projects, if successfully completed, will publish well and garner many citations, and this information guides their choices over which projects to pursue. For example, the COVID-19 pandemic which began in 2019 spurred a sudden and large interest in a particular virus and its associated proteins (Corum and Zimmer, 2020). The scientists who successfully determined the structures of these key proteins were ex-ante likely to publish in the top science journals and receive high levels of citations, acclaim, and publicity — indeed, the first structure-paper pair to describe the structure of the SARS-CoV-2 viral spike protein has received over 2,000 citations in the six months since publication (Wrapp et al., 2020; also see PDB ID 6VSB). While not all important proteins are related to a specific disease, many other features of proteins are predictive of the ex-ante demand for their structure.

While project potential is a key variable in our model, it cannot be observed directly in the data. Therefore, we estimate it. We use the rich structure-level data in the PDB to predict which proteins will be highly cited, based only on ex-ante characteristics of the protein. The predicted citation value serves as our measure of potential, corresponding to P in the model.

This kind of prediction is possible due to extremely detailed data describing and categorizing every structure in the PDB. Each structure is given a detailed classification (over 500 different classifications, such as "transcription protein" or "signaling protein"), a taxonomy (over 1,000 different organisms, such "homo sapiens" (human) or "mus musculus" (mouse)), and a link to the gene which codes for the protein (over 2,500 different genes). We also take advantage of the UniPROT prior paper measure (described in Section 1.3.3) as an additional predictor. For each structure, we compute the number of citations that the associated publication accrued over the first three years since publication (excluding self-citations). Since the citation counts are heavily right-skewed, we transform these counts into percentiles. We then use these detailed data to predict citation percentiles for each structure. It is worth pointing out that we explicitly *exclude* our complexity covariates from this prediction, in an effort to create a measure of potential that is uncorrelated with project complexity.

In this context, the number of predictors is large (over 4,000 variables) relative to the number of observations. Therefore, to avoid overfitting, we implement Least Absolute Shrinkage and Selection Operator (LASSO) to select predictors in a data-driven manner. LASSO regularization helps avoid overfitting, but it also shrinks the fitted coefficients towards zero. To remove this bias, we re-estimate an ordinary least squares regression using the LASSO-selected covariates (Belloni and Chernozhukov, 2011). We then use the post-LASSO coefficients to generate predicted citations.

In our analysis sample of of 21,951 structures, 8,667 (about 40 percent) do not have a three-year citation count. This happens because either the associated paper was published after 2015 (since our citation data only runs through 2018), or because the structure has no associated paper. Rather than drop these observations, we use the LASSO coefficients to impute the predicted citation percentiles, just as we do for the observations with non-missing citation counts.

Figure 1.4 compares actual versus predicted citation percentiles, to help assess the prediction quality. Panel A shows a histogram of actual versus predicted percentiles. While the predicted values are more clustered toward the middle percentiles, we are able to generate fairly good dispersion. Panel B shows the binned scatterplot of actual percentiles on the y-axis versus predicted percentiles on the x-axis. The fit along the $y = x$ line appears quite good throughout the distribution. Taken together, these figures suggest our prediction exercise is reasonably successful. Appendix Table A2 shows the LASSO-selected covariates and the post-LASSO ordinary least squares coefficients. While many of the coefficients are difficult to interpret, it is reassuring to see some common-sense coefficients — for example, proteins that had more prior papers written before the structure discovery tend to be more highly cited. The R^2 from the post-LASSO ordinary least squares regression suggests that we are able to capture about 17 percent of the variation in actual citation percentile with our predictions.

1.4.2 The Relationship between Potential and Competition

Proposition 4 predicts that scientists will invest more in starting high-potential projects, which will generate more competition for completing these projects. We measure investment using the number of structure authors and paper authors, as discussed in Section 1.3.2. We proxy for competition by counting the number of times the structure was deposited in the PDB within two years of the initial deposit, as discussed in Section 1.3.2. Because this variable is heavily right-skewed, we take the log.

Figure 1.5 shows the relationship between investment and potential. We illustrate the relationship using a binned scatterplot. To construct this binned scatterplot, we first residualize investment and potential with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of investment against the mean of potential in each group. Finally, we add back the mean investment period to make the scale easier to interpret after residualizing. As Figure 1.5 demonstrates, high-potential projects have both more structure authors and more paper authors, suggesting that researchers allocate more scarce personnel to more important projects. The highest-potential structures have about 4.8 structure authors and 7.5 paper authors on average, while the lowest-potential structures have about 4.5 structure authors and 6.3 paper authors on average.

Figure 1.6 is similar to Figure 1.5, but shows the relationship between potential and competition. The highest-potential structures have about 1.5 deposits per similarity cluster,[21] while the lowest-potential structures have about 1.1 deposits in the similarity cluster.

Table 1.2 formalizes these relationships. For structure *i* deposited in year *t*, we estimate:

$$Y_{it} = \alpha + \beta P_{it} + X_{it}'\gamma + \tau_t + \varepsilon_{it} \tag{1.12}$$

where Y is our outcome of interest (either investment or competition), P is our measure of potential (the predicted citation percentile), X is a vector of structure covariates, τ is a deposition year fixed effect, and ε is the idiosyncratic error term. β is the coefficient of interest, because it describes the relationship between potential and investment or potential and competition.[22]

Panel A presents the estimates of β with deposition year fixed effects, which corresponds to the plots shown in Figures 1.5 and 1.6. Throughout the remainder of this paper, we will find it convenient to benchmark effect sizes by comparing structures in the 90^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 31^{st} percentile of the citation distribution, as shown in Panel A of Figure 1.4) to structures in the 10^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 63^{rd} percentile of the citation distribution). We will term these "high-potential structures" and "low-potential structures" respectively. Columns (1) and (2) focus on the effect of potential on investment. The coefficient of 0.008 in column (1) implies that high-potential structures have 0.25 more structure authors than

[21] We arrive at this by noting that $e^{0.4} = 1.5$.

[22] We report heteroskedacity-robust standard errors. However, as argued by Pagan (1984) and Murphy and Topel (1985), because our measure of potential is a generated (i.e., estimated) regressor, OLS standard errors will be too small. In Appendix Tables A3 and A5, we re-compute the standard errors using a two-step bootstrap procedure. First, we randomly draw from our sample with replacement, creating a new sample with the same number of observations as the original sample. We use this new sample to re-generate our potential variable, allowing LASSO to re-select the model. Second, we use these generated potential measures and the same sample to estimate the OLS relationship between potential and our dependent variable. We repeat this procedure 200 times. The standard deviation in the sample of 200 coefficient estimates is our bootstrapped standard error. In practice, the boostrapped standard errors do not differ meaningfully from those reported in the main text.

low-potential structures.[23] Similarly, column (2) implies that high-potential structures also have about one additional author compared to low-potential structures. Both coefficients are statistically significant at the one percent level.

Columns (3) turns to the effect of potential on competition. The coefficient of 0.009 in column (3) suggests that high-potential structures have about 30 percent more deposits in their similarity cluster than low-potential structures.[24] Again, this effect is statistically significant at the one percent level. Appendix Table A4 provides similar estimates for alternative measures of competition.

Collectively, these results suggest that researchers are interested in maximizing their citations, and rationally choose which projects to invest in and pursue with citations in mind. In other words, it does *not* appear that researchers simply choose topics they are interested in, with no regard for the citations or acclaim their work will garner. This provides credibility for the setup of our model, where we assume that researchers are behaving as strategic citation-maximizers.

1.4.3 The Relationship between Potential and Quality

In this section, we turn to the core predictions from our model. The first part of Proposition 5 predicts that high-potential projects will be completed more quickly, as scientists internalize the fact that they are more likely to face competition for these projects. The second part of Proposition 5 predicts that this decrease in maturation will lead to lower quality among the high-potential projects. Figure 1.7 shows the relationship between maturation and potential, controlling for deposition year. The highest-potential projects have maturation periods of about 1.7 years, while the lowest-potential projects have maturation periods of about 1.9 years — a difference of just over two months. Figure 1.8 illustrates the relationship between potential and quality. Across all four quality measures, we see that higher potential is associated with lower quality. The magnitude of these correlations is notable. In Panel A, for example, we see that the highest-potential projects have resolution measures that are nearly a full standard deviation lower than the lowest-potential

[23] We calculate this by taking $0.008 \times (63-31) = 0.25$.

[24] We calculate this by taking $e^{0.009 \times (63-31)} = 1.3$.

projects. These trends are fairly consistent across the different quality measures.

Table 1.3 presents these relationships in regression form. We estimate the same regression as in Equation 1.12, but replace the dependent variable Y with our measures of maturation and quality. β remains the coefficient of interest, because it describes the relationship between potential and maturation or potential and quality. Focusing on Panel A, column (1) shows that higher-potential projects have shorter maturation periods. The coefficient of -0.005 implies that high-potential structures are completed about 0.17 years (or just over two months) faster than low-potential structures. Since the typical low-potential structure takes has a maturation period of about 1.9 years, this represents a decline of about nine percent. This effect is statistically significant at the one percent level.[25]

Columns (2) to (5) of Table 1.3 measure the effect of potential on quality. Again looking at Panel A and focusing on the aggregate quality index in column (5), the coefficient of -0.021 implies that high-potential structures have quality index scores that are about 0.7 standard deviations below their low-potential counterparts. The magnitudes are similar across the other quality measures in columns (2) to (4), and all the coefficients are statistically significant at the one percent level.

Together, these results suggest that high-potential projects are more likely to be finished quickly, which translates to lower quality on average. However, as discussed in Section 1.4.6, this negative relationship could be driven by omitted variables bias. In this setting, we are particularly concerned that high-potential structures are more complicated, and this complexity — not rushing — is what drives the lower quality. This motivates our work in the following two sections.

[25] As discussed in Section 1.3.2, our measure of maturation is imperfect. For one, it measures elapsed time, but not necessarily the hours spent working on any particular project. In addition, it only measures the time between when the scientist collects her experimental data and when she submits a draft. It does not include the time spent isolating and crystallizing the protein. Anecdotally, crystallization can be the most difficult and lengthly part of the process. Therefore, the estimates above represent the shortening of a *part* of the project lifespan.

1.4.4 Competition or Complexity?

Our model suggests that the negative relationship we document between potential and quality is caused by scientists rushing. However, an alternative explanation is that high-potential proteins might be more complex and therefore difficult to solve with high quality. If potential is positively correlated with complexity, our results could suffer from omitted variables bias, which would bias our estimate of β down.

In this and the following section, we will provide three distinct pieces of evidence which together suggest that complexity alone cannot explain the negative relationship we observe. We start by pointing out the negative relationship between potential and maturation shown in Figure 1.7. If scientists are agnostic toward priority rewards, but high-potential structures are more complex, then we would expect that scientists spend *longer* on these complex structures. In fact, we find the exact opposite, as discussed in Section 1.4.3. Researchers spend *less* time on the high-potential structures. This suggests that complexity alone cannot explain the negative relationship between potential and quality.

In general, our estimates of β in Equation 1.12 will be biased if the conditional independence assumption fails. In this context, the conditional independence assumption requires that our outcome of interest (maturation or quality) is independent of potential, conditional on controls. Therefore, our next strategy is to include controls for structure complexity, in an effort to achieve conditional independence. These controls, which are outlined in Section 1.3.2, proxy for the size of the protein structure. While it is generally difficult for researchers to anticipate which structures will be difficult to solve, larger structures tend to be more challenging.

Panel B of Table 1.3 illustrates the effect of adding these complexity controls in Equation 1.12 when quality is the dependent variable. To start, we note that these controls are powerful predictors of project quality. The R^2 dramatically increases in columns (2) through (5) with the inclusion of these controls. For example, in column (5), the R^2 increases by over a factor of three (going from 0.065 in Panel A to 0.215 in Panel B).

At the same time, the inclusion of these controls does not have a large effect on our estimated coefficients. Comparing Panels A and B in Table 1.3, we observe that the coefficients remain stable. For example, looking at our quality index outcome in column (5), we see that complexity controls reduce the magnitude of our estimate by just ten percent. Across all four quality outcomes, the coefficients remain negative and statistically significant at the one percent level.

Taken together with our maturation results, this suggests that scientific complexity is not the main driver of the negative correlation between project potential and project quality. Rather, it appears that competition and rushing play a significant role. However, in an effort to cleanly isolate the effect of competition alone, we take advantage of the fact that different researchers face different competitive incentives. This is the subject of the next section.

1.4.5 Investigating Structural Genomics Groups

In this section, we contrast structures deposited by structural genomics (SG) groups and those deposited by other researchers, in order to separate the effect of researcher rushing from other omitted factors (in particular, project complexity). As we discuss below, researchers in SG groups are less focused on competing for priority. Therefore, the optimization problem these researchers face in selecting the maturation period is similar to the no competition benchmark of the model, presented in Section 1.2.2. The model predicts that in this case, Proposition 5 should no longer hold. In other words, without competitive incentives, we no longer expect to see a negative relationship between potential and maturation or quality.[26] Comparing the SG and non-SG structures is helpful, because it allows us to "net out" potential omitted variables bias. Intuitively, if we are concerned that the negative relationship between potential and quality is driven by structure complexity, that concern likely applies to both the SG and non-SG samples. Therefore, the *difference* in slopes between the two samples is not driven by complexity, but rather by differing levels of concern over competition.

[26] This test, which takes advantage of the differing motives between the two groups, is similar in spirit to the public versus private clinical trial comparison in Budish et al. (2015).

Background on Structural Genomics Consortia

We focus on structural genomics (SG) groups because we argue that researchers in these groups face different competitive incentives than the typical academic lab. Since the early 2000s, SG consortia around the world have focused their efforts on solving and depositing protein structures in the PDB. Inspired by the success of the Human Genome Project, SG groups have a different mission than university and private-sector labs. These groups focus on achieving comprehensive coverage of the protein folding space, and eventually full coverage of the human "proteome," the catalog of all human proteins (Grabowski et al., 2016). Even without solving the structure of every protein, SG groups have achieved broader coverage of the "protein folding space," which has allowed subsequent structures to be solved more easily. For a more complete history of these structural genomics consortia, see Burley et al. 2008; Grabowski et al. 2016. All told, these initiatives have produced nearly 15,000 PDB deposits.

Importantly for our purposes, SG groups are less focused on winning priority races than their university counterparts. Indeed, the vast majority of structures solved by structural genomics groups are never published, suggesting that researchers in these groups are focused on data dissemination rather than priority. For example, The Structural Genomics Consortium (an SG center based in Canada and the United Kingdom) describes its primary aim as "to advance science and [be] less influenced by personal, institutional or commercial gain." Therefore, we view structures deposited by SG groups as a set of structures which were published by scientists who were not subject to the usual level of competition for priority.

We are able to identify SG deposits in our data by looking at the structure authors in the PDB. If the structure was solved by an SG group, that group name will be listed as the last structure author (for example, the last author might be "The Joint Center for Structural Genomics"). We use the list of SG centers tabulated by Grabowski et al. (2016) to flag structures deposited by these groups.

Table 1.4 provides summary statistics for our analysis sample separately for non-SG structures and SG structures. SG structures comprise about 20 percent of the analysis sample. The two groups differ in several ways. The SG deposits appear to be higher quality (lower refinement resolution, R-free, and Ramachandran outliers, all of which correspond to higher quality). However, these deposits also appear to be less complex. They have fewer entities, and lower molecular weight, residue count, and atom site count — all of which point to these structures being smaller and simpler to solve than their non-SG counterparts. SG structures are completed more quickly, and have more authors. In line with their stated mission, the SG structures appear to be less studied, with fewer UniPROT papers and fewer deposits within their similarity cluster. Only 20 percent of SG deposits have an associated publication, compared with 83 percent of non-SG deposits. When they do publish, they receive fewer citations.

Given these facts, it is not surprising that SG structures are lower-potential on average. This is in line with mission of the SG groups, which seek to provide coverage for less-studied proteins. However, Figure 1.9 plots the potential distributions for SG and non-SG structures. Here we see that despite the difference in means, the histograms show that the two distributions have overlapping supports. This suggests that we can draw reasonable comparisons between how SG and non-SG structures are impacted by competition and potential.

Analysis of Structural Genomics Consortia

Figure 1.10 compares the relationship between potential and maturation for both SG and non-SG structures. The two binned scatterplots are constructed separately and overlaid on the same set of axes. Because we bin each series separately, there are the same number of observations in each marker within the same series (but not across series). The fact that the markers do not line up vertically over the *x*-axis reflects the fact that the two series have different supports.

The level shift between the two groups is immediately apparent: at all levels of potential, SG structures have shorter maturation periods. The difference is over a full year on average. This gap is consistent with the mission of the SG groups, and is likely driven by their very low publication rates (20 percent of SG structures have an associated publication). These groups endeavor to get their results into the scientific domain as quickly as possible, and often do not write or release a paper to accompany the structure. Non-SG scientists, on the other hand, typically do not deposit their structures until they have a draft manuscript ready to submit.

However, the key takeaway from Figure 1.10 is that there is also a visible difference in slopes. As previously illustrated, the higher-potential non-SG structures are have shorter maturation periods (are completed more quickly). By contrast, the higher-potential SG structures appear to have have slightly *longer* maturation periods.

Figure 1.11 is isomorphic, but presents the the effects on quality. Across all four quality measures, we see that the negative relationship between potential and quality is more negative for the non-SG (i.e., more competitive) structures than it is for the SG (i.e., less competitive) structures. It is interesting to note that at low levels of potential, the quality is very similar across both groups. This suggests that non-SG researchers working on less important (and therefore less competitive) structures behave like their SG counterparts. It is only at high levels of potential (and therefore high levels of competition) that the gap becomes meaningful.

We formalize the trends shown in Figures 1.10 and 1.11 using a differences-in-differences framework. For structure i deposited in year t, we estimate the following regression:

$$Y_{it} = \alpha + \beta P_{it} + \lambda NonSG_{it} + \delta(P_{it} \times NonSG_{it}) + \tau_t + X_{it}'\gamma + \varepsilon_{it} \tag{1.13}$$

where Y is our outcome of interest (maturation or quality), and $NonSG$ is defined as an indicator equal to one for structures that were *not* deposited by an SG group. We choose to use SG deposits as the "control" group and non-SG deposits as the "treated" group, because we can think of non-SG deposits as being "treated" with competition. All other variables are the same as previously defined. β describes the relationship between potential and the outcome for the SG group. λ measures the average difference in outcomes for non-SG structures relative to SG structures. δ, the coefficient of particular interest, measures the difference in the potential-outcome correlation for non-SG structures relative to SG structures.

Table 1.5 presents the results. Focusing first on column (1) of Panel A, we see that our estimate of β (the coefficient on potential) is positive, reflecting the fact that SG groups spend *longer* in high-potential projects. We also see that our estimate λ (the coefficient on the non-SG indicator) is positive, reflecting the fact that non-SG structures are completed more slowly on average (due to higher rates of associated paper publication). However, our estimate of δ, the interaction between potential and non-SG, is negative and statistically significant. The negative estimate of the δ coefficient suggests that relationship between potential and maturation is more negative for non-SG structures relative to SG structures. In fact, it is large enough to more than offset β, implying that non-SG researchers spend less time on high-potential structures, in contrast with their SG counterparts.

If we believe that our estimates of β are contaminated by omitted variables bias, then the difference in the slopes between the SG structures $(\beta + \delta)$ and the non-SG structures (β) yields the causal effect of potential via competition. This comparison assumes that both groups suffer from the same omitted variables bias, and so it is "netted out" when we take the difference. Interpreting δ in this way implies that competition causes high-potential structures (structures that fall in the 90^{th} percentile of the potential distribution) to be completed over four months faster than low-potential structures (structures that fall in the 10^{th} percentile of the potential distribution). Recall that the average non-SG structure has a maturation period of about a 1.75 years, so this represents a meaningful (20 percent) reduction.

Columns (2) to (5) focus on the quality outcomes. Starting with Panel A, the negative estimates of β imply that even among the SG structures, there is a negative relationship between potential and quality. The positive estimates of λ reflect the fact that they y-intercept of the non-SG structures lies above the SG structures. However, more relevant is where the two series intersect at the minimum value of P (which recall is at about $P = 30$, rather than $P = 0$). If we rescaled our measure of P, the main effect of non-SG would in fact be close to zero, suggesting that quality is similar across two groups at the lowest level of potential.[27]

The estimates of the primary coefficient of interest, δ, are negative across all four quality measures and statistically significant at the one percent level. This implies that the negative relationship between potential and quality is stronger for the non-SG (i.e. more competitive)

[27] Focusing on column (5) and plugging in $P = 30$, we see that $\hat{Q}_{SG}(30) = constant - 0.009 \times 30 = constant - 0.26$ while $\hat{Q}_{NonSG}(30) = constant + 0.273 - (0.009 + 0.012) \times 30 = constant - 0.35$.

researchers. Focusing on column (5), we can interpret the the estimated δ coefficient as implying that among the non-SG structures, competition causes high-potential structures to be 0.4 standard deviations lower quality than low-potential structures, relative to SG structures. The magnitudes of the estimates are consistent across all of our quality measures. The inclusion of complexity controls in Panel B does not alter the estimates meaningfully.

The fact that the relationship between potential and quality remains negative even among the SG structures (i.e., the fact that $\beta < 0$) merits further discussion. If researchers in these groups are truly agnostic toward competition, then we would expect there to be no relationship. There are two possible explanations for this negative slope. First, perhaps researchers in SG groups *do* care about competition, but to a lesser extent than their non-SG counterparts. This could lead to negative but less steep slope. If this lesser (but non-zero) competition is the reason for the negative slope, then the effect of potential on quality due to competition in the non-SG group would be $\beta + \delta$ — in other words, we would not want to net out β.

Alternatively, SG researchers may be fully indifferent to competition, but there is a correlation between potential and unobserved complexity in both groups. Then netting out β strips the omitted variables bias from our estimates, and δ is the correct estimate. In reality, both effects may be at play. The fact that maturation is positively correlated with potential in the SG groups suggests that there may indeed be a correlation between unobserved complexity and potential. We view δ as our preferred estimate, but flag that it is likely a conservative lower bound.

1.4.6 The Relationship between Competition and Quality

Competition is the channel by which high-potential projects are ultimately executed with lower quality. This is clarified by Proposition 3, which predicts that more competitive projects are rushed and are therefore lower quality. However, as emphasized by the model, the relevant measure of competition is the researcher's perceived threat of having another researcher in the race. We cannot measure this risk, as discussed in Section 1.3.2. Instead, we measure ex-post realized competition. This noisy proxy may lead to attenuated estimates of the effect of competition on quality. Moreover, realized competition may be correlated with unobserved factors that also correlate with quality.

However, the model also suggests a solution: we can instrument for competition using project potential. Empirically, we have already demonstrated that there is a first stage (Section 1.4.2) and a reduced form (Section 1.4.3). This is enough to tell us that the relationship between competition and quality must be negative. Still, it is informative to recover the magnitudes.

We start by estimating the ordinary least squares regression using our noisy measure of ex-post competition. For structure i deposited in year t, we estimate:

$$Y_{it} = \alpha + \beta C_{it} + X_{it}'\gamma + \tau_t + \varepsilon_{it} \tag{1.14}$$

where Y is our outcome of interest (maturation or quality) and C is our proxy for competition. All other variables are the same as previously defined.

However, we also estimate a separate specification, using two-stage least squares and instrumenting for competition using project potential. The first stage regression is identical to Equation 1.12, with competition (measured as the log number of structures deposited in the same cluster within two years) as the dependent variable. The second stage regression for structure i deposited in year t is given by:

$$Y_{it} = \tilde{\alpha} + \tilde{\beta}\hat{C}_{it} + X_{it}'\tilde{\gamma} + \tilde{\tau}_t + \eta_{it} \tag{1.15}$$

where Y is the outcome of interest (maturation or quality), $\hat{C}$ is the fitted measure of competition from the first stage, X is our vector of complexity controls, $\tilde{\tau}$ is the deposition year fixed effect, and η is the idiosyncratic error term. $\tilde{\beta}$ is the coefficient of interest, as it measures the causal effect of competition on quality. The exclusion restriction in this case is that project potential only affects project quality (or maturation) through its impact on competition, conditional on controls. In other words, potential is not correlated with unobserved factors that impact quality directly once we condition on X. Our results in Section 1.4.4 and 1.4.5 help bolster this case.

Table 1.6 shows the results from both of these specifications. Comparing the coefficients of β (in Panel A) and $\tilde{\beta}$ (in Panel B), we see that competition is correlated with shorter maturation periods and lower quality in both specifications. However, as perhaps expected, we see that the estimates in Panel A are attenuated. To interpret the coefficients in Panel B, consider one structure where the expected number of researchers working is 1.25 and another more competitive structure where the expected number of researchers working is 1.5. This can roughly be interpreted as a 25 percentage point increase in the probability of a competitor. The coefficient in column (1) implies this second structure would be completed one to two months faster.[28] The coefficient in column (5) implies the second structure would score 0.4 standard deviations lower using our quality index.

1.4.7 Benchmarking the Quality Estimates

Are the negative quality effects we estimate large enough to matter for overall scientific productivity in our setting? Rushing leads to lower quality structures, but are these structures low enough quality to prevent researchers from drawing useful conclusions or using the structure in follow-on work? According to structural biologists, the answer depends on what the researcher wishes to do with the structure. If the researcher simply wants to understand the protein's function, a lower-quality structural model may be sufficient. However, if a scientist hopes to use a protein structure for structure-based drug design, then a high-quality structure is required. Anderson (2003) suggests that in order to be useful for structure-based drug design, the structures must have a resolution of 2.5Å or lower, and an R-free of 0.25 or lower.[29] While these cutoffs may not be hard-and-fast, they tell us something about the usefulness of a structure given its quality. It is not uncommon for structures to fall below these thresholds. About 35 percent of the non-SG structures in our analysis sample lie below this resolution cutoff. About 45 percent of these same structures lie below the R-free cutoff.

Drugs typically work by binding to proteins, changing the protein's function. The protein that the drug binds to is known as the "target." In an effort to empirically validate these claims, we use DrugBank to link drugs to their protein targets, and these targets to their PDB ID(s). For every structure in the PDB, this allows us to count the number of drugs that target that particular structure.

[28] $-0.610 \times (\ln 1.5 - \ln 1.25) = 0.11$ years or 1.33 months.

[29] Recall that for the raw resolution and R-free measures, lower values correspond to better quality.

If quality is important for drug development, we would expect high-quality structures (especially structures that surpass the Anderson (2003) criteria) to be targeted more frequently by drugs, all else equal.

Panel A of Figure 1.12 shows the relationship between drug development and resolution in a binned scatterplot.[30] Here we plot unstandardized resolution, so recall that lower values correspond to higher quality. We also plot the 2.5Å cutoff for reference. There is a clear positive relationship between higher levels of drug development and lower (i.e., better) resolution. The relationship is nonlinear, with a sharp drop off at around 2.0Å, which is slightly lower (i.e., better) than the 2.5Å cutoff. Panel B repeats this procedure with R-free (again, lower values unstandardized R-free correspond to higher quality). We again see a sharp drop off in drug development at lower quality. Here that drop off occurs at an R-free of about 0.23, which is slightly lower (i.e., better) than the 0.25 threshold proposed by Anderson (2003). Still, taken together with the conventional wisdom from the literature, these figures suggest that a certain level of quality is necessary for drug development. Moreover, this threshold is stringent enough that many of the structures in our data do not meet or surpass it. This suggests that the negative quality effects we measure are large enough to impact downstream drug development.

1.5 Welfare Implications

Thus far, we have been focused entirely on the positive predictions of the model. Normative conclusions are more difficult to draw. Nevertheless, in the first part of this section, we make the case that researchers cannot easily "fix" low-quality structures, and so the quality effects we measure capture a real inefficiency in the generation of new scientific knowledge. While many low-quality structures are improved over time, offsetting some of the detrimental effects of racing, this comes at a substantial cost. Next, we turn to the question of optimal policy. We show that

[30] If a structure has been deposited multiple times, we use resolution form the best (i.e., highest-quality) structure. The idea is that a pharmaceutical firm would always use the best structure available. We discuss this in more detail in Section 1.5.1.

the current allocation of investment and maturation chosen by racing teams falls short of idealized first-best, but it may represent a constrained second-best allocation. We discuss alternative policies that might improve quality and investment levels in science.

1.5.1 Will Follow-On Work Fix the Problem?

Even if the quality effects we measure are meaningful, is the rush to publish and the subsequent lower-quality work necessarily bad for science? Society values speed of disclosure as well as quality, in part because the quality of a discovery might be improved upon over time. Therefore, in certain circumstances, a rushed low-quality discovery might be preferable to a higher-quality breakthrough that takes longer to develop. The overall costs and benefits of rushing depends in part on the knowledge production model. If science progresses like a quality ladder, where each researcher can build frictionlessly on existing work (Grossman and Helpman, 1991), then quick-and-dirty work is likely not bad for science. To fix ideas, consider the example of ornithologist and molecular biologist Charles Sibley. In 1958, he began collecting egg white samples from as many birds as possible in order to better understand the differences between species. In 1960, he published a survey of over 5,000 proteins from over 700 different species (Sibley, 1960; Strasser, 2019). Now, suppose Sibley had been concerned that a competitor was working on a similar project, and instead released his survey a year earlier, in 1959, with proteins from only 350 different species. Another ornithologist (or indeed, Sibley himself) could add to the survey without having to regenerate any of the existing work.

On the other hand, consider a structural biologist working on a new protein structure. Suppose, for example, that she has a choice: she could spend a year growing her protein crystals and solving and refining her structure, which would yield a 2.5Å structure. Alternatively, she could rush — spending just six months, she could generate a 3.0Å structure. If she rushes, consider the incentives for another researcher to improve the structure from 3.0Å to 2.5Å. This researcher would have to start from scratch, growing new crystals, generating new experimental data, and creating a structural model. The new researcher would have to sink an entire year — not to mention the financial cost — to achieve the marginal 0.5Å quality improvement. Even if the new researcher decides the improvement is worth the cost, it is inefficient. The first researcher could have achieved the 2.5Å structure with a year of work. Instead, the combined researchers spend a year and a half. The key point is that — in contrast to quality ladder models (and the toy naturalist example above), which assume that researchers can frictionlessly build on most current work — the new researcher has to re-sink the same costs in order to generate a marginal improvement.

Bringing this logic into the context of our model, suppose a follow-on researcher is considering whether to improve the quality of a project with potential P and quality $Q(m^{C^*})$. If she generates higher quality by letting the project mature for $m^{IMP} > m^{C^*}$, then she will be rewarded for her marginal quality improvement. Therefore, the present discounted value of this improvement is

$$e^{-rm^{IMP}} P \left[Q(m^{IMP}) - Q(m^{C^*}) \right]. \tag{1.16}$$

The optimal maturation period for the improved structure, m^{IMP^*}, is given by[31]

$$m^{IMP^*} \in \underset{m^{IMP}}{\arg\max} \left\{ e^{-rm^{IMP}} P \left[Q(m^{IMP}) - Q(m^{C^*}) \right] \right\} \tag{1.17}$$

which yields the first-order condition

$$\frac{Q'(m^{IMP^*})}{[Q(m^{IMP^*}) - Q(m^{C^*})]} = r. \tag{1.18}$$

Lemma 1. *The present discounted value of improving a project is increasing in P, project potential.*

Proof. See Appendix A.1.1. The intuition is that the present discounted value of improving a project depends primarily on the project's potential (P) and the quality improvement ($Q(m^{IMP^*}) - Q(m^{C^*})$). Both of these are increasing in P, so the effect on the present discounted value is positive. □

[31] Here we are ignoring racing concerns. We think this is reasonable when focusing on new deposits of an already-solved structure that occur some time after the initial structure deposit.

This above analysis of the maturation decision is conditional on successfully starting the project. However, before entering the project the researcher must first sink an investment cost I. As we discussed in the ornithologist versus structural biologist example above, the follow-on researcher in our setting must re-sink this cost — she cannot take advantage of the fact that a previous researcher already invested. As before, if a researcher invests I, she has probability $g(I)$ of successfully starting the project where $g(\cdot)$ is an increasing, concave function. The optimal value of this investment, I^{IMP^*}, is given by

$$I^{IMP^*} \in \underset{I^{IMP}}{\arg\max} \left\{ g(I^{IMP}) e^{-rm^{IMP^*}} P \left[Q(m^{IMP^*}) - Q(m^{C^*}) \right] - I^{IMP} \right\} \tag{1.19}$$

which yields the first-order condition

$$g'(I^{IMP^*}) = \frac{1}{e^{-rm^{IMP^*}} P \left[Q(m^{IMP^*}) - Q(m^{C^*}) \right]}. \tag{1.20}$$

This immediately gives us Proposition 6.

Proposition 6. *The optimal level of investment for a project that involves re-solving an existing structure (I^{IMP^*}) is increasing in project potential (P). Therefore, high-potential projects are more likely to be re-solved.*

Proof. This comes immediately from noting that $g'(\cdot)$ is decreasing and applying Lemma 1. □

To document whether Proposition 6 is true empirically, we need to identify when a project in our analysis sample is re-solved.[32] We are once again able to use the PDB's cluster classification. If we see that a structure in our analysis sample has another structure in its same similarity cluster that was deposited two years or later than the initial structure, we say that structure was re-solved.[33] We use this "two year" rule in an effort to separate contemporaneous work from replications or

[32] Recall that our analysis sample restricts to structures that were solved for the first time.

[33] In practice this is complicated by the fact that clusters are assigned at the entity level which is a smaller unit of analysis than a structure (one structure can have multiple entities). We discuss the details in Appendix A.2.

re-deposits. Panel A of Figure 1.13 plots the probability a structure is re-solved as a function of project potential. We observe exactly what Proposition 6 predicts — higher *P* structures are more likely to be re-solved. Scientists are more willing to invest in re-solving these structures because (a) they are more valuable and (b) there is more room for improvement.

We can use the re-solved structures within a cluster to find the best quality ever produced for a particular protein. What does Proposition 6 tell us about the relationship between the *maximum* quality of a structure and *P*? At a given value of *P*, the average maximum quality of all structures with potential equal to *P* will be given by

$$\overline{Q}_{max}(P) = Q(m^{C^*}) + g(I^{IMP^*})\left[Q(m^{IMP^*}) - Q(m^{C^*})\right]. \tag{1.21}$$

The first term represents the initial quality, while the second term represents the probability there is an improved structure, times the quality improvement. Note that m^{C^*}, I^{IMP^*}, and m^{IMP^*} all depend on *P*. What happens to $\overline{Q}_{max}$ as *P* increases? This leads to the following proposition:

Proposition 7. *As P increases, the sign of the effect on $\overline{Q}_{max}$ is ambiguous. However, the slope of $\overline{Q}_{max}$ versus P is higher than the slope of $Q(m^{C^*})$. In other words, $\frac{d\overline{Q}_{max}}{dP} > \frac{dQ(m^{C^*})}{dP}$.*

Proof. See Appendix A.1.1. Intuitively, both $g(I^{IMP^*})$ and $Q(m^{IMP^*}) - Q(m^{C^*})$ are increasing in *P*. This must at least partially offset the negative relationship between $Q(m^{C^*})$ and *P*. □

Panel B of Figure 1.13 tests this proposition. The first series on the plot (the dots) shows the relationship between potential and a structure's initial quality, as in Figure 1.8. However, the second series (the diamonds) shows the relationship between potential and the structure's *maximum* quality, when looking across all structures within a similarity cluster. The vertical distance between the red and blue series represents the average quality improvement. As predicted by Proposition 7, the relationship between potential and maximum quality is less negative than the relationship between potential and initial quality. In fact, the relationship between potential and maximum quality is U-shaped. The intuition is that at low values of *P*, the incentives to re-solve are low, but the initial quality is high. At high values of *P*, the incentives to re-solve are high. This leads to high maximum quality at the extremes of the potential distribution, and lower maximum quality in the middle of the distribution.

Returning to our concerns about project complexity in Section 1.4.4, it is comforting to see that the maximum quality values at the top end of the potential distribution are nearly as high as the maximum quality values at the bottom of the potential distribution, because it suggests that high quality is possible for these high-potential structures. If the negative relationship between potential and initial quality were driven purely by structure complexity, we might expect that it is simply impossible to solve these high-potential structures at the same level of quality.[34]

Together, Panels A and B of Figure 1.13 suggest that there are three distinct sources of welfare loss associated with rushing in structural biology. First, there is the loss of structure quality, which translates to lost downstream innovation. However, Panel B shows that without taking into account the subsequent re-deposits, we will overestimate the magnitude of this lost quality as much of it (particularly for the highest potential structures) is made up in future work. Second, there is the time cost associated with the re-deposits. While much of the lost structure quality is eventually reclaimed via follow-on work, this takes additional time. Finally, there is the monetary cost associated with re-solving the same structures. The PDB estimates that the average cost to replicate a structure is about $100,000 (Sullivan et al., 2017).

1.5.2 Optimal Policy

The Infeasible First Best

We start our optimal policy analysis by considering how equilibrium maturation and investment that arises from researchers competing for priority (i.e., m^{C^*} and I^{C^*}) compares to the outcome preferred by an unconstrained social planner. In this setting, an unconstrained social planner would like to dictate both investment (I) and maturation (m) to researchers. The social planner's objective differs from an individual researcher's objective in two ways: first, the social planner only cares

[34] This is not a perfect test, because technology may have improved between when the original structure was deposited and when the new structure was deposited, enabling better quality structures. Nevertheless, it is a reassuring data point.

that at least one researcher successfully starts the project. If both researchers start the project, the planner is indifferent as to which researcher completes the project first, and the second (replicated) structure adds no additional social value. This wedge is similar to the inefficiency identified by Dasgupta and Maskin (1987). Second, consistent with the notion of research generating positive spillovers, the social value of a given project is greater than the private value. We operationalize this by assuming that the social planner's PDV of the project at completion is $e^{-rm}kPQ(m)$, rather than $e^{-rm}\overline{\theta}PQ(m)$ or $e^{-rm}\underline{\theta}PQ(m)$ (the first- and second-place researcher's private PDV, respectively). We further assume that k is large relative to $\overline{\theta}$ and $\underline{\theta}$ (we put more formal bounds on k in the analysis below). Putting these facts together, we have the social planner's objective function:

$$\max_{m,I}\left\{ \underbrace{\left(1-(1-g(I))^2\right)}_{\text{probability at least one researcher successfully starts}} \cdot \underbrace{e^{-rm}kPQ(m)}_{\text{social PDV of project}} - \underbrace{2I}_{\text{investment costs}} \right\}. \tag{1.22}$$

Contrast this with the individual researcher's objective function (Equation 1.9, reproduced and slightly re-arranged below):

$$\max_{m_i,I_i}\left\{ \underbrace{g(I_i)}_{\text{probability } i \text{ successfully starts}} \underbrace{\cdot e^{-rm_i}\left[\overline{\theta}-\frac{1}{2}g(I_j)(\overline{\theta}-\underline{\theta})\right]PQ(m_i)}_{i'\text{s expected private PDV of project}} - \underbrace{I_i}_{i'\text{s investment cost}} \right\}. \tag{1.23}$$

The socially optimal value of m, denoted m^{SP^*}, is defined by the first-order condition of Equation 1.22 with respect to m:

$$\frac{Q'(m^{SP^*})}{Q(m^{SP^*})} = r. \tag{1.24}$$

Notice that this is identical to the first-order condition which defines the optimal value of m in the absence of competition (m^{NC^*}, see Equation 1.4). Therefore, we know that $m^{SP^*} > m^{C^*}$. In other words, the social planner wants projects to mature for longer than researchers will allow them to in a competitive environment. This happens precisely because the social planner — unlike the individual researcher — does not care who finishes the project first. Concerns over priority distort the individual researcher's choice of m away from the social optimum.

The socially optimal value of I, denoted I^{SP^*}, is defined by the first-order condition of Equation 1.22 with respect to I:

$$g'(I^{SP^*}) = \frac{1}{e^{-rm^{SP^*}}kPQ(m^{SP^*})(1-g(I^{SP^*}))}. \tag{1.25}$$

Comparing this equation with the first-order condition that defines I^{C^*} (Equation 1.10), we can see that if k is sufficiently large,[35] then $I^{SP^*} > I^{C^*}$. Intuitively, if the social planner values the project sufficiently more than the researcher, the social planner will want the researcher to invest more than the privately optimal level.

The empirical evidence supports the theoretical argument that individual researchers distort their behavior away from the social optimum. More specifically, Equation 1.24 implies that if we were at the first best, then the relationship between potential and quality should be flat. Instead, we observe a negative relationship between potential and quality, consistent with researchers distorting their behavior in an effort to complete their projects first.

The Feasible Second Best: Using Credit Share as a Policy Lever

The social planner cannot realistically dictate I and m for each project. Monitoring the progress of every scientific team as they work on their projects requires too much information to be feasible. Instead, a more reasonable lever for the social planner might be $\overline{\theta}$ or $\underline{\theta}$, the share of credit allocated to the first and second-place team, respectively. While the literature has often assumed that priority races are winner-take-all, implying that $\underline{\theta} = 0$ (for example, Merton (1957); Fudenberg et al. (1983); Bobtcheff et al. (2017)) empirical evidence suggests that this is not the case. Hill and Stein (2020b) find that in structural biology, winning teams involved in priority races receive about 55 percent of the credit (as measured by citations) — a far cry from 100 percent. While that same paper provides survey evidence to suggest that structural biologists are more pessimistic about the

[35] More precisely, if $k > \frac{\overline{\theta}-\frac{1}{2}g(I_j)(\overline{\theta}-\underline{\theta})}{1-g(I^{SP^*})}$ then k meets the criteria of "sufficiently large."

costs of being scooped (the surveyed authors estimated the winning paper would accrue about 70 percent of the total citations), the 100 percent benchmark does not appear to be correct in this setting.

Moreover, it appears that the bulk of this credit disparity is driven by journal placement rather than citation behavior. This suggests that the priority premium is primarily driven by journal editors and reviewers, who could perhaps be influenced to change their policies. Indeed, a handful of journals have begun to do exactly this — changing their policies to explicitly state that they will treat recently scooped papers the same as novel papers. Concerns about competition harming the quality of submitted work appear to be top of mind. For example, in 2017 the journal *eLife* released the following statement:

> "We all know graduate students, postdocs and faculty members who have been devastated when a project that they have been working on for years is 'scooped' by another laboratory, especially when they did not know that the other group had been working on a similar project. And many of us know researchers who have rushed a study into publication before doing all the necessary controls because they were afraid of being scooped. Of course, healthy competition can be good for science, but the pressure to be first is often deleterious, not only to the way the science is conducted and the data are analyzed, but also for the messages it sends to our young scientists. Being first should never take priority over doing it right or the search for the truth. For these reasons, the editors at *eLife* have always taken the position that we should evaluate a paper, to the extent we can, on its own merits, and that we should not penalize a manuscript we are reviewing if a paper on a similar topic was published a few weeks or months earlier" (Marder, 2017).

Other journals have released similar policies.[36] In light of these changes, the distribution of credit

[36] For example, in January 2018, *PLOS Biology* released a statement reading, "scientific research can be a cutthroat business, with undue pressure to publish quickly, first, and frequently. The resulting race to publish ahead of competitors is intense and to the detriment of the scientific endeavor. Just as summiting Everest second is still an incredible achievement, so too, we believe, is the scientific research resulting from a group

is a particularly interesting and relevant policy tool to study. However, the precise way in which we allow the social planner to manipulate the distribution of credit will have different implications for optimal policy. We consider two cases in turn.

Case 1: Total Rewards are Fixed. In the first case, we consider a social planner who can manipulate $\overline{\theta}$ and $\underline{\theta}$, but cannot change the size of the total private value of the project. In other words, $\overline{\theta}$ and $\underline{\theta}$ can vary, but $\overline{\theta}+\underline{\theta}$ is fixed. To fix notation, let $\overline{\theta}+\underline{\theta}=V$. In this case, the fact that $\overline{\theta} \geq \underline{\theta}$ implies that $\overline{\theta} \geq \frac{V}{2}$ and $1-\overline{\theta} \leq \frac{V}{2}$.

Here we are allowing the social planner to manipulate one parameter ($\overline{\theta}$) in an effort to target two choice variables (m^{C^*} and I^{C^*}). In other words, the social planner would like to pick a value of $\overline{\theta}$ that will induce researchers to select $m^{C^*}=m^{SP^*}$ and $I^{C^*}=I^{SP^*}$. However, as we will show below, no value of $\overline{\theta}$ makes this possible. With just $\overline{\theta}$ at the social planner's disposal, the planner cannot attain the first best.

Lemma 2. *If the social planner sets $\overline{\theta}=\underline{\theta}=\frac{V}{2}$, then researchers will select the optimal maturation period. However, if k is sufficiently large, then investment will be too low.*

Proof. Recall that the social planner would like the researcher to behave as if there is no competition. In other words, $m^{SP^*}=m^{NC^*}$. Intuitively, if we equate the rewards for the first- and second-place researcher, we have eliminated competition, and so researchers will let their projects mature optimally. However, this results in investment below the socially optimal level. See Appendix A.1.1 for more detail. □

By setting $\overline{\theta}=V-\overline{\theta}=\frac{V}{2}$, the social planner is able to select the optimal maturation period, but investment is too low. Next, we will show that as the social planner raises $\overline{\theta}$ — making priority

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* (The PLOS Biology Staff Editors, 2018). In November 2018 the editor of *Cell Systems* released a statement saying "*Cell Systems* thinks it is valuable — as well as simply humane — to welcome strong experimental studies that are "scooped" (Justman, 2018).

rewards more lopsided — maturation periods become shorter, but investment may increase. This sets up a tradeoff for the social planner: more unequal priority rewards lead to shorter maturation periods (moving us away from the optimal maturation level), but potentially higher investment levels (moving us closer to the optimal investment level). This implies that optimal priority rewards may be unequal. Proposition 8 below formalizes this logic.

Proposition 8. *If we restrict $\overline{\theta} + \underline{\theta}$ to sum to a fixed value V, then the researcher's optimal maturation period m^{C^*} is decreasing in $\overline{\theta}$, while the researcher's optimal investment level I^{C^*} may be increasing in $\overline{\theta}$. This implies that the optimal choice of $\overline{\theta}^*$ may lie between $\frac{V}{2}$ and 1. The resulting values of $m^{C^*}(\overline{\theta})$ and $I^{C^*}(\overline{\theta})$ will not achieve the social optimum, with $m^{C^*}(\overline{\theta}^*) < m^{SP^*}$ and $I^{C^*}(\overline{\theta}^*) < I^{SP^*}$.*

Proof. See Appendix A.1.1. □

Proposition 8 helps us interpret the welfare implications of the negative relationship between potential and quality that we document in our empirical results. As clarified by the model, this negative relationship is a product of the unequal priority rewards — in other words, it will exist as long as $\overline{\theta} > \underline{\theta}$. However, proposition 8 illustrates that the optimal choice of $\overline{\theta}^*$ may in fact result in lopsided priority rewards, and so the negative relationship between potential and quality — while inconsistent with an unconstrained social optimum — *is potentially consistent* with a constrained second-best solution. In other words, the negative relationship between potential and quality does *not* imply that a constrained social planner could increase overall welfare.

Case 2: Total Rewards Can Vary. In this case, we consider a social planner who can manipulate $\overline{\theta}$ and $\underline{\theta}$ independently, with no restrictions on $\overline{\theta} + \underline{\theta}$. Intuitively, the social planner has more freedom in this case because $\overline{\theta}$ and $\underline{\theta}$ are independent. In this case, we are allowing the planner to manipulate two parameters ($\overline{\theta}$ and $\underline{\theta}$) in an effort to target two choice variables (m^{C^*} and I^{C^*}). This allows the social planner to achieve the socially optimal investment and maturation, as shown in Proposition 9 below.

Proposition 9. *If we allow the social planner to select $\overline{\theta}$ and $\underline{\theta}$ independently, then the planner can achieve the optimal m^{C^*} and I^{C^*} by setting $\overline{\theta}^* = \underline{\theta}^* = k(1-g(I^{SP^*}))$, which is increasing in k.*

Proof. Setting $\overline{\theta} = \underline{\theta}$ ensures that we achieve the socially optimal maturation, as shown in Lemma 2. Allowing $\overline{\theta} + \underline{\theta}$ to be unconstrained means we can induce the appropriate amount of investment. Intuitively, if the social value of a project is high, then $\overline{\theta} + \underline{\theta}$ will be larger. See Appendix A.1.1 for details. □

Of the two cases outlined above, which represents a more realistic policy lever that a social planner or policy maker could dial up or down? In the basic sciences, where rewards come primarily in the form of credit, we argue that Case 1 is more relevant. Credit is a fickle thing — not handed down by a particular individual, but rather assigned by the community. Reputations are bolstered by awards, prizes, and rankings which are necessarily zero-sum, making manufacturing additional credit (i.e., increasing $\overline{\theta} + \underline{\theta}$) difficult. While journal editors and reviewers can endeavor to bring more attention to scooped researchers via some of the example journal policies outlined above, this likely comes at the expense of the credit granted to the first-place researcher, who is now viewed as more of a co-discoverer rather than the sole discoverer.

On the other hand, in settings where researchers are primarily remunerated with wages rather than credit, Case 2 is more relevant. Wages, unlike credit, are easy to manipulate. A firm can simply choose to set wages optimally, and recover the first-best investment level and maturation period. It is worth noting that if k is large, then optimal wages will be high. Firms will only choose to set these high wages if they capture the full social surplus (in other words, if there are not positive spillovers outside the firm). Still, this highlights one advantage of conducting research inside of firms. As emphasized by Holmstrom (1999), it allows for “access to more instruments,” leading to a better set of incentives.

An Alternative Policy: Ending Races Early

Another policy option would be to end priority races when the first team successfully starts the project, and let that team carry out the maturation phase without threat of competition. In other words, once one team successfully started the project, other teams would be barred from entering. This would lead to teams choosing the optimal maturation period (recall that the maturation period selected in the absence of competition is the same as the socially optimal maturation period). Investment levels would depend on the payoff that the winning team receives, but they would be higher than in the standard competitive case, because the projects are more valuable when allowed to fully mature.

This policy works because of the somewhat specific nature of our model. In particular, all the uncertainty occurs in the investment stage, while the maturation stage is purely deterministic. Having two teams competing during the investment stage can be helpful, because it increases the probability that at least one team successfully starts the project. But once at least one team has entered the project, there is no more uncertainty, and so the second team no longer brings a benefit. Yet, despite the model-specific nature of this policy, we highlight it because it is relevant in structural biology — so relevant in fact, that an informal policy along these lines once existed in the field.

Recall that when solving protein structures, the most difficult and risky part of the process is growing the protein crystal. Researchers may try to crystallize a protein under a variety of conditions and simply fail to generate a usable crystal. Therefore, growing the crystal is analogous to the investment stage of the model. Researchers sink resources, which increases the odds they successfully crystallize their protein and can start building their model. By contrast, building the atomic model from the diffraction data is a more deterministic process, akin to the maturation phase. Therefore, the analog of ending priority races early in this setting would be to let researchers "call dibs" on a protein structure once they successfully crystallize it. Then they can build the structure from their experimental data, without fear of being preempted.

Barring other teams from entering to solve the structure is akin to increasing patent breadth in models of follow-on innovation (for example, Green and Scotchmer (1995) and Hopenhayn and Squintani (2016)). As pointed out by Horstmann et al. (1985) and Scotchmer and Green (1990) in the patent realm, researchers might ordinarily be reluctant to patent or release any details of their initial project (i.e., the protein crystal) if doing so would give competitors an informational advantage in their efforts to develop a related project (i.e., to solve the structure). However, by giving the team that crystallizes the protein some informal intellectual property over the eventual structure, researchers become willing to share this work.[37]

In fact, in the early days of structural biology, there was a strong, community-enforced norm that if "someone else is working on [a structure] — hands off" (Strasser, 2019). As Ramakrishnan (2018) explains, scientists would announce (often through publication) that they had successfully crystallized a protein, and "there was a tradition that if someone had produced crystals of something, they were usually left alone to solve the problem." This norm exactly parallels the policy of stopping races once the first research has successfully entered the project. However, as the field grew and the number of unsolved structures dwindled, this precedent became too difficult to enforce. Today structural biologists are secretive about what they are working on, knowing that the "hands off" rule no longer applies (Strasser, 2019). Still, it is interesting to note that structural biology organically developed a set of norms which alleviated the problem of rushing and associated lower quality work, even if those norms have not been sustained to the present day.

1.6 Conclusion

This paper documents that in the field of structural biology, competition to publish first and claim priority causes researchers to release their work prematurely, leading to lower quality science. We explore the implications of this fact in a model where scientists choose which projects to work on, and how long to let them mature. Our model clarifies that because important problems in science are more crowded and competitive, perversely it is exactly these important projects that will be

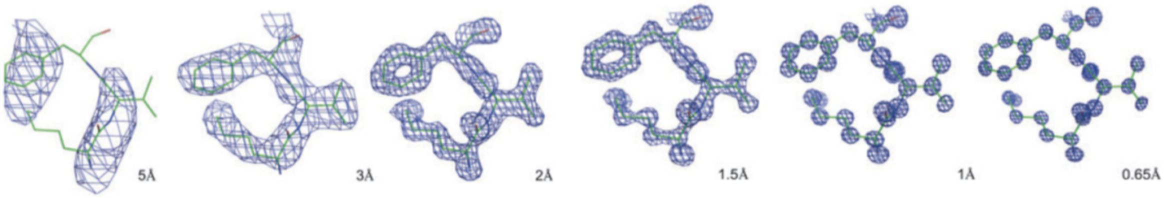
[37] In some contexts, we might be concerned about allowing teams to claim intellectual property prematurely, especially if another team is better suited to carry out the eventual work that is protected. Ouellette (2019) outlines this view in the patent system. We assume this concern away, because our model assumes that all researchers are equally skilled at solving the protein structure given the experimental data, although in practice this may be a concern in our setting and a potential drawback of this policy. Indeed, in cases where researchers felt that the team with the crystal was making insufficient progress, other researchers would violate this norm and also begin to work on the problem (Ramakrishnan, 2018).

the most poorly executed. We find strong evidence of this negative relationship between project potential and project quality in our data. While this negative relationship is inconsistent with an idealized first best, where a social planner can dictate how much investment researchers dedicate to projects and how long they let these projects mature, it *not inconsistent* with a more realistic constrained second best, where the social planner can only dictate how credit is shared between first- and second-place researchers.

We stop short of attempting to calibrate an optimal credit split between first- and second-place scientists. Such a calibration would require assigning dollar values to marginal quality improvements, as well as careful measurement of project investment, both of which are beyond the scope of this project and our data. However, perhaps more importantly, such a calibration would likely be incomplete. Competition shapes the field of science in numerous ways. In this project, we focus on the effect it has on scientific quality, and explore the potential tradeoff a social planner faces between inducing more investment versus longer maturation (and thus higher-quality work). However, other margins are likely important as well. For example, heightened competition may reduce potentially productive collaborations across different labs, promoting secrecy and ultimately slowing the pace of innovation (Walsh and Hong, 2003; Anderson et al., 2007). Competition also may influence who selects into and remains in certain fields of science. Others have expressed concern that increased competition has led to “crippling demands” on scientists’ time, leaving little time for “thinking, reading, or talking with peers” — key ingredients for transformative research (Alberts et al., 2014). These additional margins represent productive avenues for future research, and are also key inputs to consider when determining how best to allocate credit and the optimal level of competition in science.

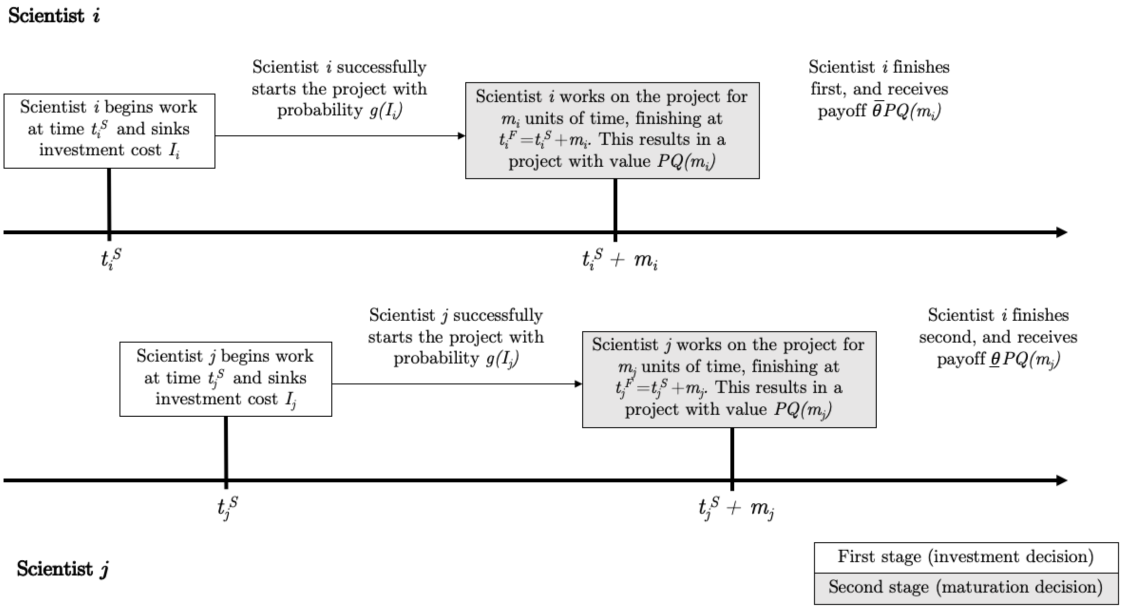
Figures and Tables

Figure 1.1: Illustration of a Protein Structure at Different Refinement Resolutions



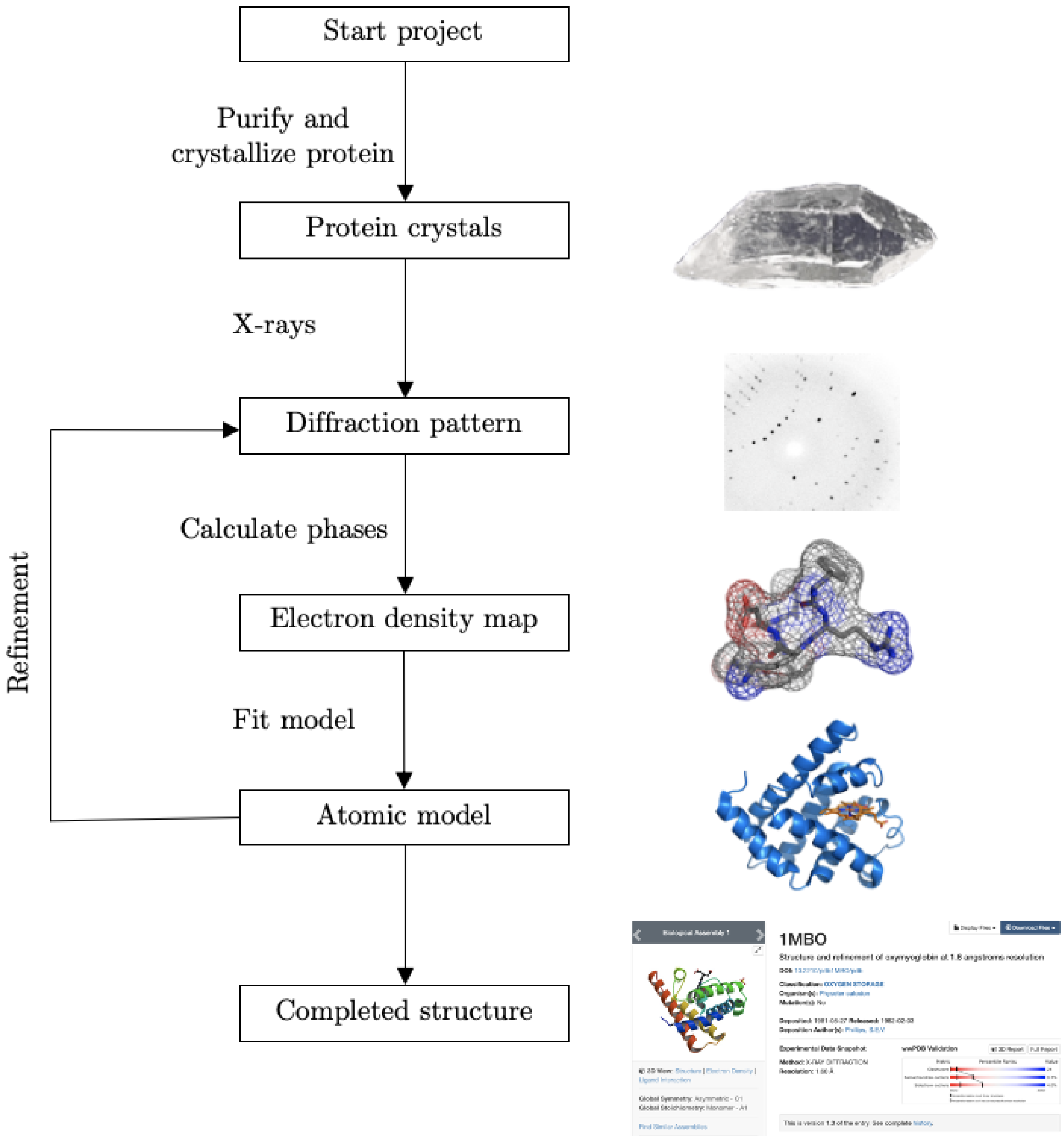
Notes: This figure shows the electron density maps from a fragment of the triclinic lysozyme (PDB ID 2VB1) at different refinement resolutions. The Angstrom (Å) values measure the smallest distance between crystal lattice planes that can be detected in the experimental data. Lower values correspond to better (higher-resolution) structures. Figure taken from Wlodawer et al. (2008).

Figure 1.2: Model Summary


Scientist i
Scientist i begins work at time t_i^S and sinks investment cost I_i
Scientist i successfully starts the project with probability $g(I_i)$
Scientist i works on the project for m_i units of time, finishing at $t_i^F = t_i^S + m_i$. This results in a project with value $PQ(m_i)$
Scientist i finishes first, and receives payoff $\overline{\theta} PQ(m_i)$
t_i^S
$t_i^S + m_i$
Scientist j begins work at time t_j^S and sinks investment cost I_j
Scientist j successfully starts the project with probability $g(I_j)$
Scientist j works on the project for m_j units of time, finishing at $t_j^F = t_j^S + m_j$. This results in a project with value $PQ(m_j)$
Scientist i finishes second, and receives payoff $\underline{\theta} PQ(m_j)$
t_j^S
$t_j^S + m_j$
Scientist j
First stage (investment decision)
Second stage (maturation decision)

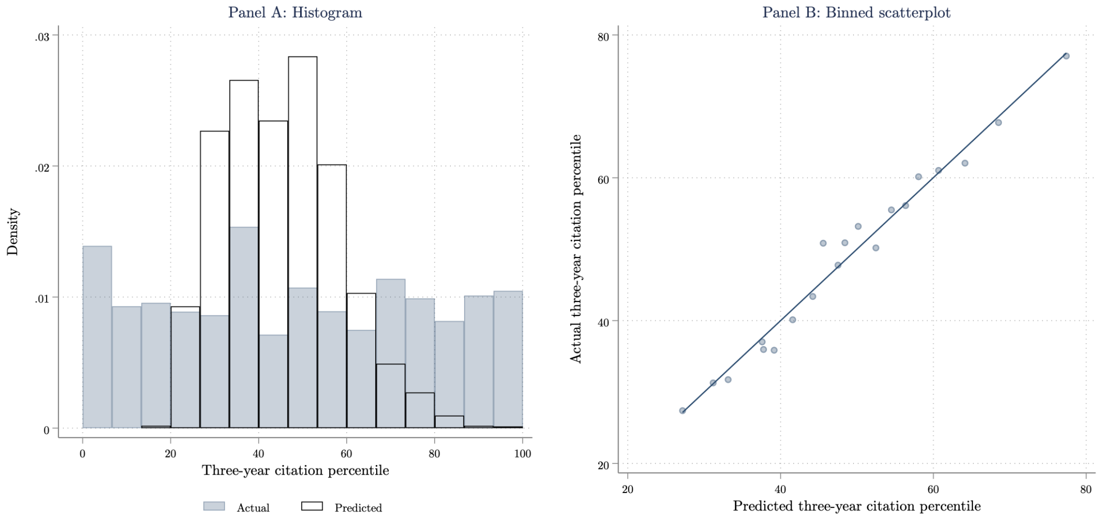
Notes: This figure summarizes the setup of the model described in the text.

Figure 1.3: Summary of the X-Ray Crystallography Process


Start project
Purify and crystallize protein
Protein crystals
X-rays
Diffraction pattern
Calculate phases
Electron density map
Fit model
Atomic model
Completed structure
Refinement
1MBO
Structure and refinement of oxymyoglobin at 1.6 angstroms resolution
Biological Assembly 1
Classification: OXYGEN STORAGE
Method: X-RAY DIFFRACTION
wwPDB Validation
This is version 1.3 of the entry. See complete history.

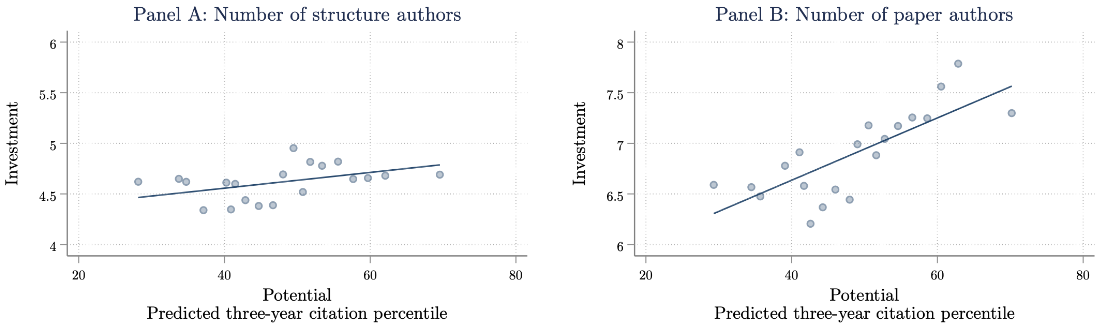
Notes: This figure summarizes the process of solving a protein structure via x-ray crystallography. The images in this figure were taken from Thomas Splettstoesser (www.scistyle.com) and rendered with PyMol based on PDB ID 1MBO.

Figure 1.4: LASSO Validation



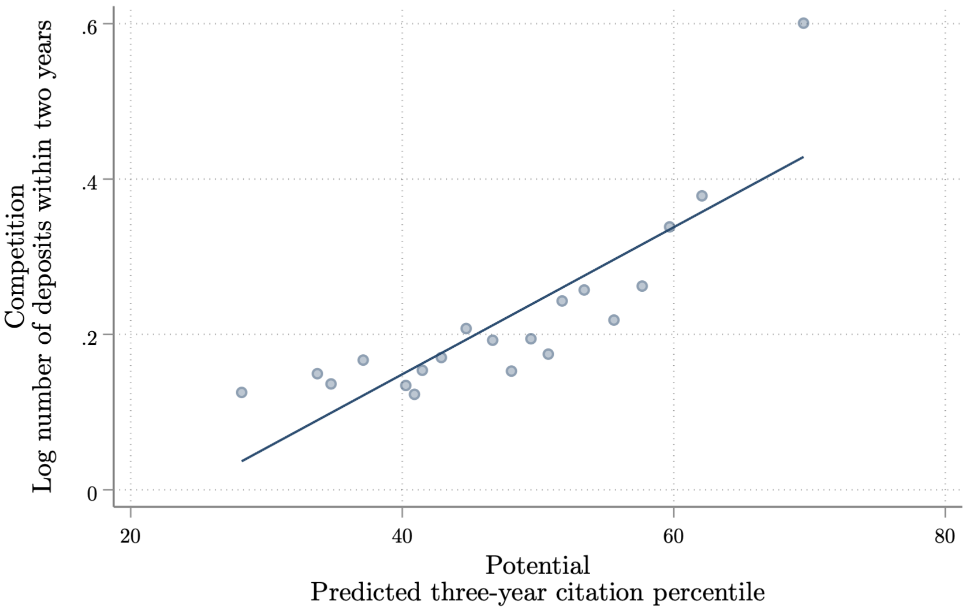
Notes: Panel A of this figure plots the distribution of actual and predicted potential. Panel B presents a graph of actual versus predicted potential as a binned scatterplot. In both panels, potential is measured by the percentile of the structure's three-year citation count. To construct this binned scatterplot, we divide the sample into 20 equal-sized groups based on the ventiles of predicted three-year citation percentile, and plot the mean of actual three-year citation percentile against the mean of predicted three-year citation percentile in each bin. The sample is all structures in the analysis sample that have a three-year citation count.

Figure 1.5: The Effect of Potential on Investment



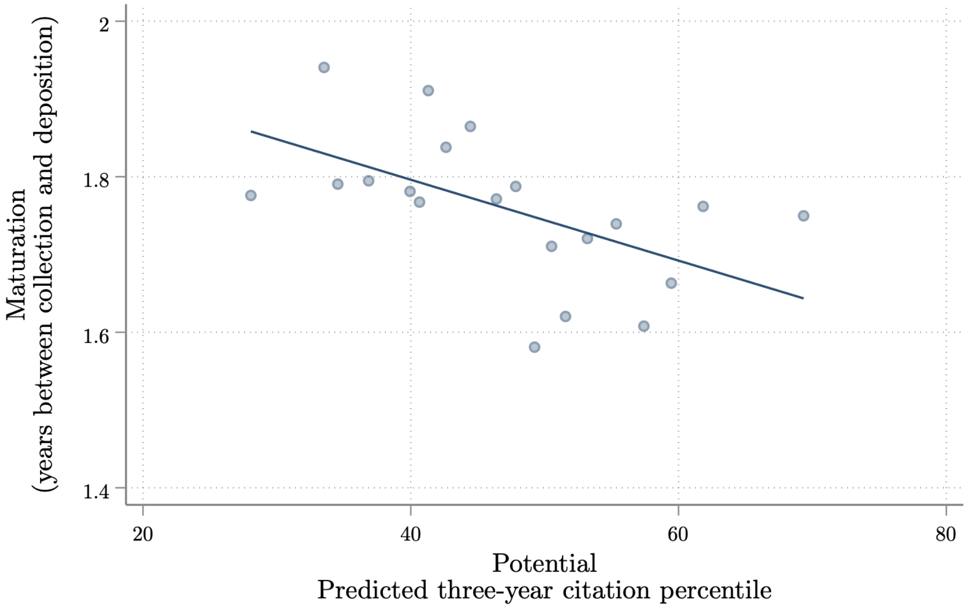
Notes: This figure plots the relationship between potential and investment, testing Proposition 4 of the model. Potential is measured as the predicted three-year citation percentile. Investment is measured as either the number of structure authors or number of paper authors. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we first residualize potential and investment with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of investment against the mean of potential in each bin. Finally, we add back the mean investment to make the scale easier to interpret after residualizing. The sample in Panel A is the full analysis sample as defined in the text, excluding SG deposits. The sample in Panel B is the same, but excludes observations that have no associated publication and therefore no paper author count.

Figure 1.6: The Effect of Potential on Competition



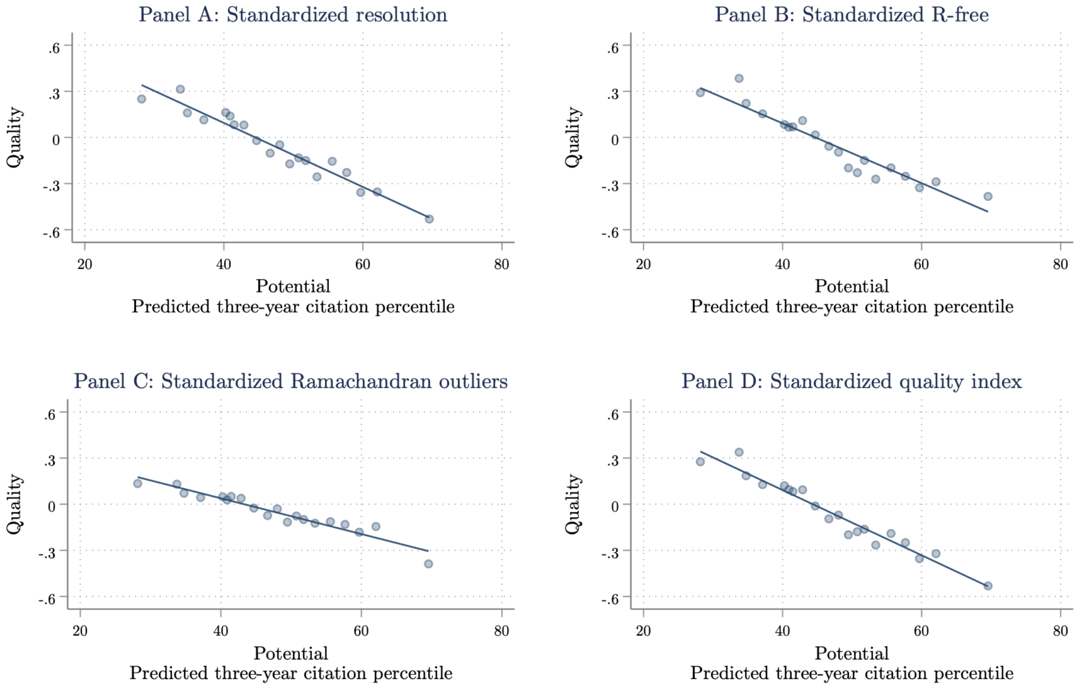
Notes: This figure plots the relationship between potential and competition, testing Proposition 4. Potential is measured as the predicted three-year citation percentile. Competition is measured as the log number of deposits that appear in the 100 percent similarity cluster within two years of the first deposit in the cluster. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we first residualize potential and competition with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of competition against the mean of potential in each bin. Finally, we add back the mean competition to make the scale easier to interpret after residualizing. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure 1.7: The Effect of Potential on Maturation



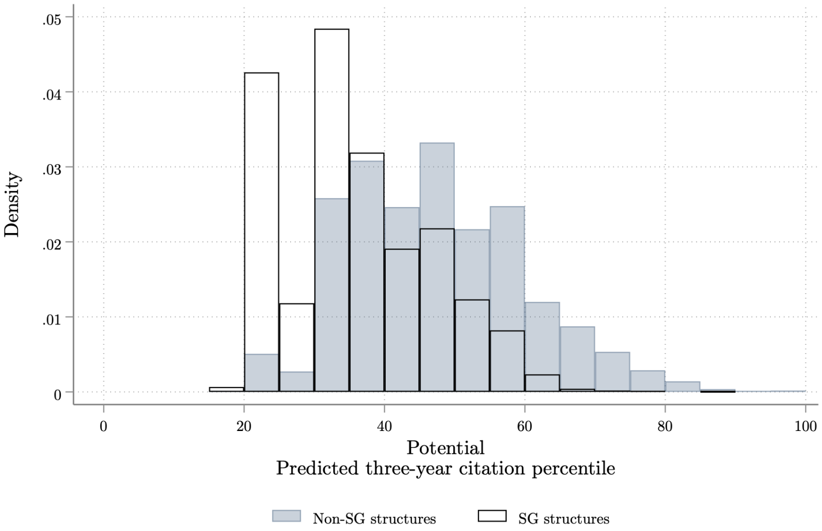
Notes: This figure plots the relationship between potential and maturation, testing Proposition 5. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we first residualize potential and maturation with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of maturation against the mean of potential in each bin. Finally, we add back the mean maturation to make the scale easier to interpret after residualizing. The sample is the full analysis sample as defined in the text, excluding SG deposits and observations where the maturation is missing.

Figure 1.8: The Effect of Potential on Quality



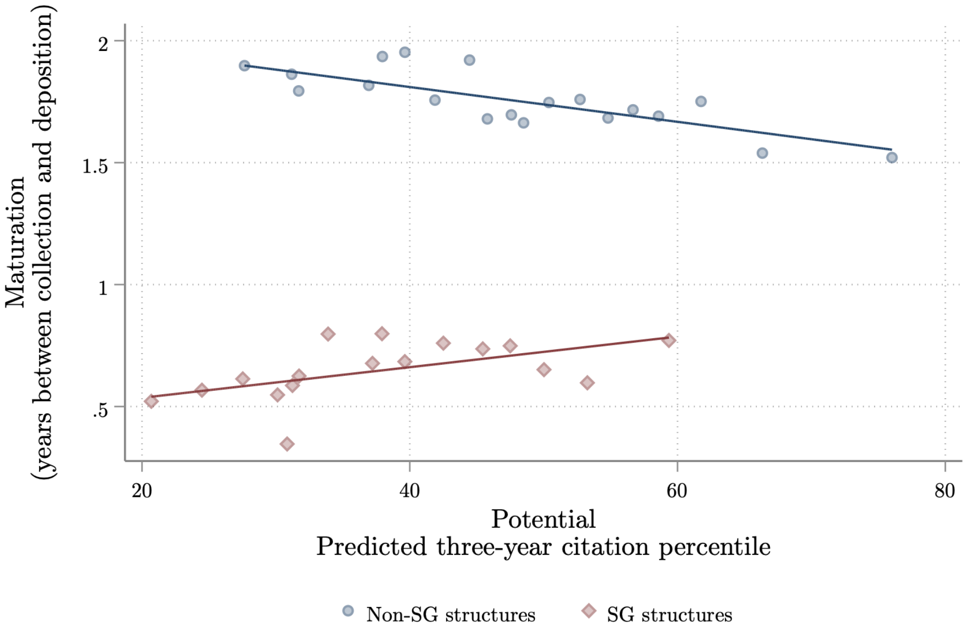
Notes: This figure plots the relationship between potential and quality, testing Proposition 5. Potential is measured as the predicted three-year citation percentile. Quality is measured by our four standardized quality measures described in detail in Section 1.3.2. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we first residualize potential and quality with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of quality against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure 1.9: Potential Distributions by Structural Genomics Status



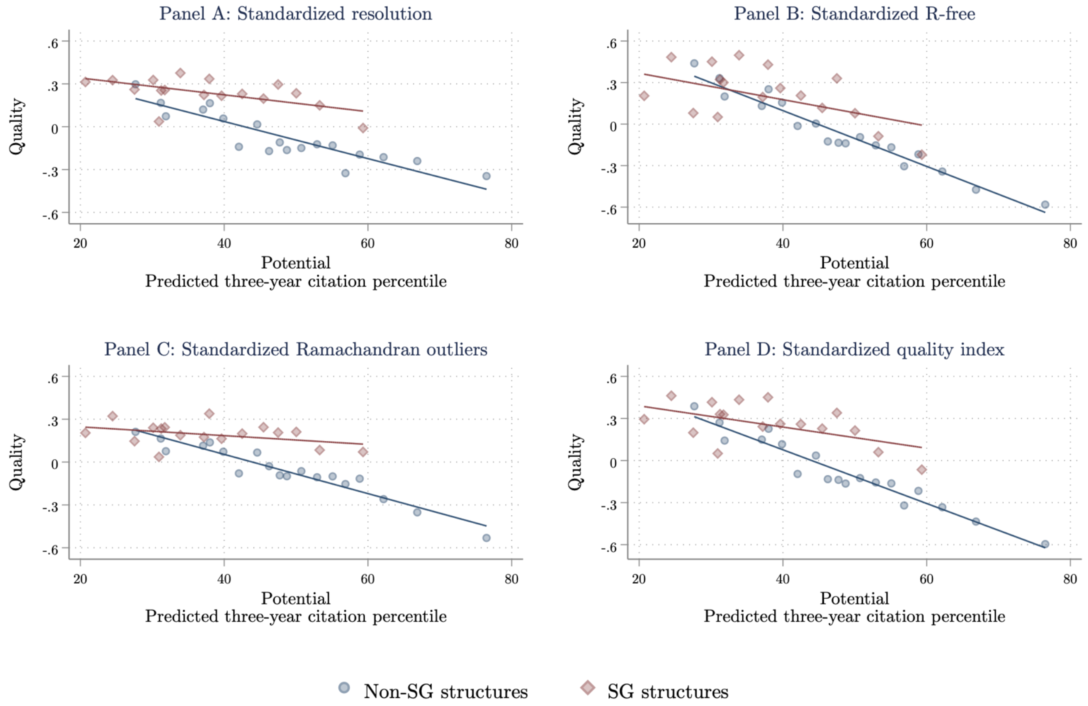
Notes: This figure plots the distribution of potential (measured by predicted three-year citation percentile) for both non-SG and SG structures. The sample is all structures in the analysis sample.

Figure 1.10: The Effect of Potential on Maturation by Structural Genomics Status



Notes: This figure plots the relationship between potential and maturation, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plots are presented as two separate binned scatterplots, overlaid on the same axes. To construct these binned scatterplots, we first residualize potential and maturation with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of maturation against the mean of potential in each bin. Finally, we add back the mean maturation period to make the scale easier to interpret after residualizing. We repeat this procedure separately for the SG and non-SG structures, but plot the resulting series on the same axes. As a result, there are the same number of observations within each point in the same series. The sample is the full analysis sample where the maturation variable is non-missing.

Figure 1.11: The Effect of Potential on Quality by Structural Genomics Status



Panel A: Standardized resolution

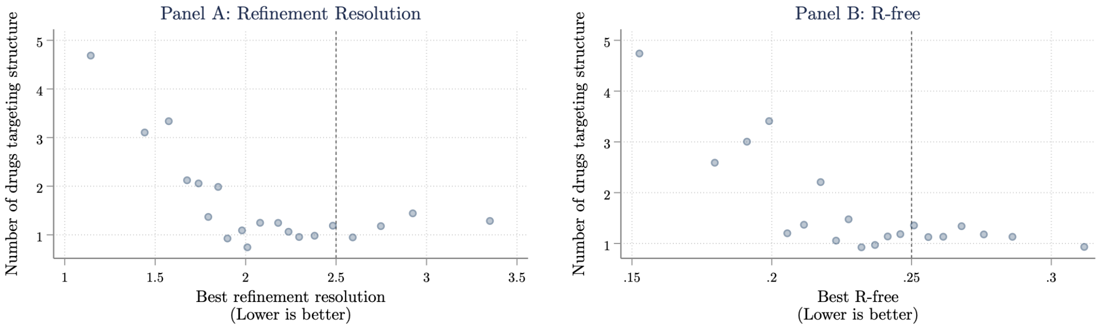
Panel B: Standardized R-free

Panel C: Standardized Ramachandran outliers

Panel D: Standardized quality index

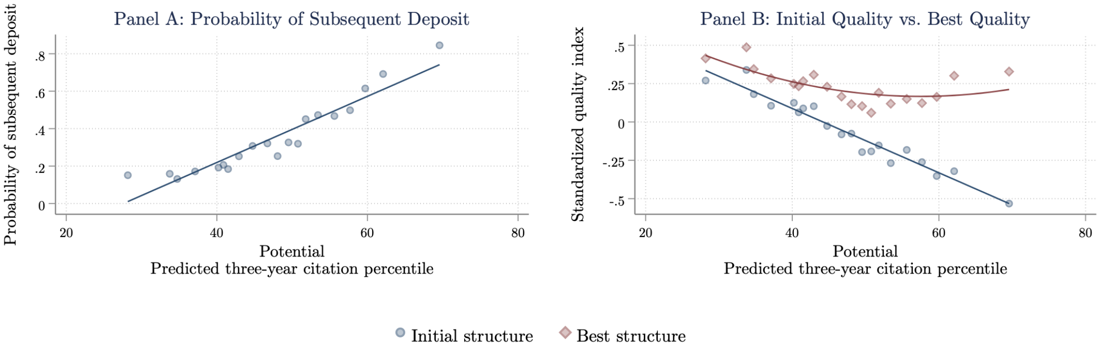
Notes: This figure plots the relationship between potential and quality, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Quality is measured by our four standardized quality measures described in detail in Section 1.3.2. The plots are presented as two separate binned scatterplots, overlaid on the same axes. To construct these binned scatterplots, we first residualize potential and quality with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of quality against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. We repeat this procedure separately for the SG and non-SG structures, but plot the resulting series on the same axes. As a result, there are the same number of observations within each point in the same series. The sample is the full analysis sample.

Figure 1.12: Relationship between Structure Quality and Drug Development



Notes: This figure plots the relationship between structure quality and structure's use in drug design. Quality is measured using unstandardized refinement resolution and R-free, so lower values indicate better quality. In instances where the same structure is deposited in the PDB multiple times, we take the best quality. The results are presented as a binned scatterplot. To construct this binned scatterplot, we divide the sample into 20 equal-sized groups based on the ventiles of resolution or R-free distribution, and plot the mean of the drug count against the mean of quality measure in each bin. The dashed lines indicate the quality thresholds for drug development proposed by Anderson (2003).

Figure 1.13: Subsequent Structure Deposits and Maximum Structure Quality



Notes: This figure plots the relationship between potential and probability of subsequent deposition (Panel A) and the relationship between potential and initial quality and best quality (Panel B). A subsequent deposit is defined as a deposit in the same 100 percent cluster that is deposited in the PDB more than two years after the first deposit. Quality is measured using our quality index described in detail in Section 1.3.2. The plots are presented as binned scatterplots. To construct these binned scatterplots, we first residualize the dependent variable (indicator for subsequent deposit, the initial quality, or the best quality) and potential with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of the dependent variable against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. The sample is the full analysis sample.

Table 1.1: Summary Statistics: Full Sample versus Analysis Sample

	All X-Ray Crystallography Sample						Analysis Sample					
	Mean	Median	Std. Dev.	Min	Max	% Missing	Mean	Median	Std. Dev.	Min	Max	% Missing
Panel A. Structure-level statistics												
Quality measures												
Refinement resolution (lower is better)	2.2	2.0	0.6	0.5	15.0	0.2%	2.2	2.2	0.6	0.6	9.5	0.0%
R-free value (lower is better)	0.24	0.24	0.04	0.05	0.51	5.0%	0.24	0.24	0.04	0.07	0.48	0.0%
Ramachandran outliers (lower is better)	0.6	0.1	1.6	0.0	100.0	4.5%	0.8	0.2	1.9	0.0	75.0	0.0%
Maturation measures												
Years between collection and deposition	1.8	1.2	2.0	0.0	123.0	11.8%	1.5	1.0	1.7	0.0	22.8	8.1%
Competition measures												
Deposits per similarity cluster within two yrs	4.1	2.0	16.5	1.0	297.0	0.0%	1.4	1.0	1.3	1.0	49.0	0.0%
Investment measures												
Authors per structure	4.9	4.0	3.9	1.0	88.0	0.0%	5.3	4.0	3.9	1.0	88.0	0.0%
Authors per paper	8.0	7.0	5.6	1.0	88.0	18.4%	7.1	6.0	4.9	1.0	88.0	29.3%
Complexity measures												
Number of entities	1.5	1.0	3.0	1.0	91.0	0.0%	1.5	1.0	2.5	1.0	86.0	0.0%
Molecular weight (1000s of Daltons)	107.1	51.9	600.1	0.3	97730.5	0.0%	102.0	55.0	444.2	0.4	47370.7	0.0%
Residue count (1000s of amino acids)	0.8	0.5	1.5	0.0	89.2	0.0%	0.8	0.5	1.3	0.0	46.9	0.0%
Atom site count (1000s of atoms)	6.5	3.4	16.4	0.0	717.8	0.0%	5.9	3.6	12.8	0.0	470.6	0.0%
UniProt papers	9.5	4.0	16.9	0.0	199.0	0.0%	6.2	2.0	11.2	0.0	198.0	0.0%
Deposition year	2009.1	2010.0	6.2	1972.0	2018.0	0.0%	2008.6	2009.0	5.6	1993.0	2018.0	0.0%
Total number of structures	128,876						21,951	0	0	0	0	0.0%
Panel B. Paper/project-level statistics												
Number of structures	2.1	1.0	4.3	1.0	860.0	0.0%	1.0	1.0	0.0	1.0	1.0	0.0%
Fraction published	0.76	1.00	0.43	0.00	1.00	0.0%	0.71	1.00	0.46	0.00	1.00	0.0%
Three-year citations	16.6	9.0	28.8	0.0	913.0	36.1%	17.2	9.0	29.8	0.0	811.0	39.5%
Total number of papers/projects	63,809						21,951					

Notes: This table shows summary statistics for the structure-level and paper/project-level data. We present summary statistics for both the full sample and our analysis sample. Structures are grouped into a single paper based on PubMed ID. For structures with no associated PubMed ID / publication, we impute which structures were part of the same project (see text and Appendix for details). Complexity variables (molecular weight, residue count, atom site count) are divided by 1000 for ease of interpretation.

Table 1.2: The Effect of Potential on Investment and Competition

	Investment		Competition
	Number of structure authors	Number of paper authors	Log number of deposits within two years
Dependent variable	(1)	(2)	(3)
Panel A. Without complexity controls			
Potential	0.008***	0.031***	0.009***
	(0.002)	(0.003)	(0.000)
R-squared	0.023	0.063	0.050
Panel B. With complexity controls			
Potential	0.007***	0.033***	0.009***
	(0.002)	(0.003)	(0.000)
R-squared	0.026	0.065	0.081
Mean of dependent variable	4.615	6.896	0.655
Observations	17,688	14,680	17,688

Notes: This table shows the relationship between investment / competition and potential, testing Proposition 4 of the model and estimating regression equation (12) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include molecular weight, residue count, and atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample. The sample in column (2) is smaller because some structures don't have an associated publication. Heteroskedasticity-robust standard errors are in parentheses.

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

Table 1.3: The Effect of Potential on Maturation and Quality

	Maturation	Quality			
		Std.	Std.	Std. Rama.	Std. quality
	Years	resolution	R-free	outliers	index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	-0.005***	-0.021***	-0.019***	-0.012***	-0.021***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.016	0.048	0.077	0.057	0.065
Panel B. With complexity controls					
Potential	-0.005***	-0.018***	-0.019***	-0.009***	-0.019***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.018	0.281	0.162	0.098	0.215
Mean of dependent variable	1.759	-0.060	-0.052	-0.048	-0.065
Observations	15,982	17,688	17,688	17,688	17,688

Notes: This table shows the relationship between maturation/ quality and potential, testing Proposition 5 of the model and estimating regression equation (12) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include molecular weight, residue count, and atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample. The number of observations in column (1) is lower because maturation is missing for a subset of observations. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses. *$p<0.1$, **$p<0.05$, ***$p<0.01$.

Table 1.4: Summary Statistics: Non Structural Genomics Sample versus Structural Genomics Sample

	Non-Structural Genomics Sample						Structural Genomics Sample					
			Std.			%			Std.			%
	Mean	Median	Dev.	Min	Max	Missing	Mean	Median	Dev.	Min	Max	Missing
Panel A. Structure-level statistics												
Quality measures												
Refinement resolution (lower is better)	2.3	2.2	0.6	0.6	9.5	0.0%	2.1	2.0	0.4	0.9	4.3	0.0%
R-free value (lower is better)	0.24	0.25	0.04	0.07	0.48	0.0%	0.23	0.24	0.03	0.12	0.39	0.0%
Ramachandran outliers (lower is better)	0.9	0.3	2.0	0.0	75.0	0.0%	0.4	0.0	1.0	0.0	13.7	0.0%
Maturation measures												
Years between collection and deposition	1.8	1.2	1.8	0.0	22.8	9.6%	0.6	0.2	1.1	0.0	12.6	1.9%
Competition measures												
Deposits per similarity cluster within two yrs	1.5	1.0	1.4	1.0	49.0	0.0%	1.2	1.0	0.7	1.0	13.0	0.0%
Investment measures												
Authors per structure	4.6	4.0	3.0	1.0	88.0	0.0%	8.1	7.0	5.5	1.0	73.0	0.0%
Authors per paper	6.9	6.0	4.0	1.0	88.0	17.0%	11.6	8.0	12.0	2.0	72.0	80.5%
Complexity measures												
Number of entities	1.6	1.0	2.7	1.0	86.0	0.0%	1.1	1.0	0.5	1.0	22.0	0.0%
Molecular weight (1000s of Daltons)	108.8	56.2	492.9	0.4	47370.7	0.0%	73.4	50.5	81.5	5.6	1641.1	0.0%
Residue count (1000s of amino acids)	0.8	0.5	1.4	0.0	46.9	0.0%	0.7	0.4	0.7	0.0	15.5	0.0%
Atom site count (1000s of atoms)	6.2	3.7	14.0	0.0	470.6	0.0%	4.8	3.3	5.4	0.3	113.3	0.0%
UniProt papers	7.1	3.0	12.0	0.0	198.0	0.0%	2.3	0.0	5.6	0.0	103.0	0.0%
Deposition year	2008.6	2010.0	5.9	1993.0	2018.0	0.0%	2008.6	2008.0	3.9	1997.0	2018.0	0.0%
Total number of structures	17,688						4,263	0	0	0	0	0.0%
Panel B. Paper/project-level statistics												
Number of structures	1.0	1.0	0.0	1.0	1.0	0.0%	1.0	1.0	0.0	1.0	1.0	0.0%
Fraction published	0.83	1.00	0.38	0.00	1.00	0.0%	0.20	0.00	0.40	0.00	1.00	0.0%
Three-year citations	17.5	9.0	29.9	0.0	811.0	29.3%	11.9	5.0	27.9	0.0	324.0	81.5%
Total number of papers/projects	17,688						4,263					

Notes: This table shows summary statistics for the structure-level and paper/project-level data. We present summary statistics for both the non-SG sample and the SG sample, within the analysis sample.. Structures are grouped into a single paper based on PubMed ID. For structures with no associated PubMed ID / publication, we impute which structures were part of the same project (see text and Appendix for details). Complexity variables (molecular weight, residue count, atom site count) are divided by 1000 for ease of interpretation.

Table 1.5: The Effect of Potential on Maturation and Quality, by Structural Genomics Status

	Maturation	Quality			
		Std.	Std.	Std. Rama.	Std. quality
	Years	resolution	R-free	outliers	index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	0.006***	-0.007***	-0.010***	-0.004***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Non-structural genomics	1.491***	0.368***	0.194***	0.107**	0.273***
	(0.081)	(0.053)	(0.056)	(0.045)	(0.052)
Potential * Non-structural genomics	-0.011***	-0.013***	-0.009***	-0.008***	-0.012***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.085	0.056	0.086	0.065	0.080
Panel B. With complexity controls					
Potential	0.006***	-0.006***	-0.009***	-0.003***	-0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Non-structural genomics	1.503***	0.343***	0.213***	0.063	0.253***
	(0.081)	(0.048)	(0.054)	(0.044)	(0.048)
Potential * Non-structural genomics	-0.012***	-0.012***	-0.009***	-0.006***	-0.011***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.087	0.274	0.171	0.102	0.221
Mean of dependent variable	1.526	0.000	0.000	0.000	0.000
Observations	20,164	21,951	21,951	21,951	21,951

Notes: This table shows the relationship between maturation / quality and potential, interacted with structural genomics status, estimating equation (13) in the text. The regressions include interactions between potential and an indicator for whether the structure was deposited by a non-structural genomics group. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Structural genomics deposits are defined as described in the text. Complexity controls include molecular weight, residue count, and atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of structures in the analysis sample. The number of observations in column (1) is lower because maturation is missing for a subset of observations. Heteroskedasticity-robust standard errors are in parentheses.

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

Table 1.6: The Effect of Competition on Maturation and Quality

	Maturation	Quality			
		Std.	Std.	Std. Rama.	Std. quality
	Years	resolution	R-free	outliers	index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Ordinary least squares					
Competition	-0.150***	-0.053***	-0.014	-0.053***	-0.049***
	(0.032)	(0.016)	(0.016)	(0.020)	(0.017)
Complexity controls?	Y	Y	Y	Y	Y
Panel B. Two-stage least squares					
Competition	-0.610***	-2.112***	-2.146***	-1.082***	-2.181***
	(0.167)	(0.122)	(0.125)	(0.112)	(0.127)
Complexity controls?	Y	Y	Y	Y	Y
First-stage F statistic	508.5	575.8	575.8	575.8	575.8
Mean of dependent variable	1.76	-0.06	-0.05	-0.05	-0.07
Observations	15,982	17,688	17,688	17,688	17,688

Notes: This table shows the relationship between maturation / quality and competition, testing Proposition 3 of the model. Panel A presents the results from an OLS regression, following equation (14) in the text. Panel B presents the results from a 2SLS regression, where competition is instrumented with potential, following equations (12) and (15) in the text. The level of observation is a structure-paper pair. Competition is measured as the number of deposits within a 100 percent similarity cluster within two years of the first deposit. Complexity controls include molecular weight, residue count, and atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of non-SG structures in the analysis sample. In column (1), we report fewer observations due to missing data in the maturation variable. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses.
*$p<0.1$, **$p<0.05$, ***$p<0.01$.

Chapter 2

Scooped! Estimating Rewards for Priority in Science*

2.1 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)*

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

Basic science is a critical input to innovation, but it may be under-provided in competitive markets because discoveries are not directly marketable and property rights are difficult to enforce. Unlike applied research, basic (or "pure") scientific research advances our fundamental understanding of the world, but typically does not yield immediate opportunities for commercialization (Nelson 1959; Arrow 1962). As a result, *credit* for ideas, rather than direct profits, is a potential motivator of innovative activity (Dasgupta and David 1994). Within academia, there is a widespread notion that the first person to publish a new discovery receives the bulk of the credit. Scientists therefore compete fiercely for priority (Merton 1957). Famous examples of priority disputes include Isaac Newton versus Gottfried Leibniz over the invention of calculus, Charles Darwin versus Alfred Wallace over the discovery of natural selection and evolution, and more recently, Grigori Perelman versus Shing-Tung Yau, Xi-Peng Zhu, and Haui-Dong Cao over the proof of the Poincaré conjecture. This competition for recognition shapes the culture and professional structure of many disciplines, and scientists regularly worry about their work being "scooped" or preempted by a competitor (Hagstrom 1974). Many theoretical papers about innovation races conceptualize the reward structure as winner-take-all (Loury 1979; Fudenberg et al. 1983; Dasgupta and David 1994; Bobtcheff et al. 2017). However, there is little empirical evidence documenting how credit is allocated in science or how rewards are shared between the "winners" and "losers" of these races.

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

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

We construct races using two key dates that are recorded for all PDB projects. First, the deposit date marks when the scientist first uploaded their findings to the PDB. Scientists typically deposit their findings shortly after a manuscript has been submitted for publication. The second is the release date, which closely corresponds to the date of publication and is usually two to six months after deposit. Critically for our design, the data is hidden from the public (and from competing scientists) between deposit and release. To construct races, we find instances where two or more teams had deposited a structure discovery for identical macromolecules independently of each other

[1] 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.

prior to the other competitors' release date. The order of release then defines the outcome of the race. The first team to release is the winner, and the second team is scooped. We identify 1,630 races in our data. These races consist of 3,319 separate projects out of 64,018 total projects in our sample period from 1999 to 2017, suggesting that five percent of all structural biology projects are involved in a late-stage race to publication. These races are composed of a diverse set of scientific teams from different countries, institutional prestige, and experience. Our definition of scooped projects focuses only on late-stage races where both teams are on the cusp of publication. Researchers may worry about being scooped earlier in the research process, such as during the design or data collection of an experiment. We cannot systematically identify these events in our data if the first team publishes before the second team deposits. Nevertheless, focusing only on late-stage scoops is advantageous for the economic interpretation of our results. Since both projects had been completed independently prior to publication, we can infer that the second-place team *would have* published the priority paper in the counterfactual where they had not been scooped. The estimated difference in observed outcomes therefore isolates the premium for novelty awarded by editors and readers.

While getting scooped is not randomly assigned, we use multiple methods to assess the validity of the causal identification assumptions. We estimate the effect of winning a race using the naturally occurring variation in the priority ordering of races. Therefore, omitted variables bias is a threat to the causal interpretation of the estimates. If the winners are positively selected on experience, research ability, or university prestige, our estimates of the scoop penalty will be biased up (in terms of magnitudes). However, we find that the outcome of races — even if not perfectly random — is highly unpredictable. We observe cases of both high-ranked teams scooping low-ranked teams, and low-ranked teams scooping high-ranked teams. Throughout the analysis, we carefully document potential sources of bias and assess treatment balance using the observable team and author characteristics. To further mitigate concerns of omitted variables bias, we use the post-double-selection Lasso method for control variable selection (Belloni et al. 2014).

We find that getting scooped has a moderate-sized impact on the success of the scooped project.

Scooped projects are 2.5 percent less likely to be published. Scooped papers appear in a 0.18 standard deviation lower-ranked journal, and are 19 percent less likely to appear in a top-10 journal. Scooped papers receive 20 percent fewer citations, and are 23 percent less likely to be a "hit" paper, defined as reaching the top 10 percent in citations for that publishing year. While these effect sizes are meaningful, they are far from a winner-take-all division of credit. Focusing on citations as an outcome, our estimates imply that the losing paper receives 45 percent of the total citations accrued by both papers, a much higher share than the zero percent assumed by a winner-take-all model.

Much of the citation effect is driven by journal placement, with only a five percent difference in citations once we control for journal fixed effects. We provide suggestive evidence that editors and reviewers have a strong taste for novelty. Papers that are scooped prior to submission to a top journal are rarely, if ever, accepted for publication. Some scooped papers do appear in top journals, but only if they were far along in the review process on the date they are scooped.

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

Does getting scooped have a detrimental impact on the careers of individual authors? We compare the future publications, citations, and academic longevity of scientists on the winning and losing teams. We find that scientists who are scooped are about five percent less likely to be actively depositing in the PDB five years after they were scooped, but not less likely to be publishing in life and medical sciences as a whole. We do not find significant effects on intensive margin publication rates. However, scooped scientists receive 17 percent fewer citations to their future work, an effect that is stronger for novice scientists (32 percent) than their veteran co-authors (13 percent).

We analyze and discuss how the priority reward system relates to inequality in science. Our sample of races provides unique insight into how reputation affects academic attention, because we see teams of varying reputation and affiliation competing to publish the same discovery first. We find that when a high-reputation lab scoops a relatively unknown lab, they receive 66 percent of the total citations, but when a low-reputation lab scoops a high-reputation lab, they only receive 46 percent of the total citations. We rationalize this asymmetry in priority rewards with a model of academic attention based on the statistical discrimination literature (Phelps 1972; Aigner and Cain 1977). Our model proposes that readers receive a noisy signal of a paper's true quality, and therefore place some weight on the authors' pre-existing reputation. A high-reputation team that wins the race not only receives a premium for priority, but also a boost in citations because of their renown. If a low-reputation team scoops a high team, the winner still receives a priority benefit, but it is fully offset by a penalty for their lower reputation. This relationship between priority credit and reputation suggests that compensation in science is not formulaic, but may be influenced by the attention constraints and biases of editors and readers.

Finally, we benchmark the size of the scoop penalty by comparing it to the perceptions of active structural biologists. We survey 915 corresponding authors of papers linked to the PDB and pose a hypothetical scenario about getting scooped. The respondents estimate a 25 percent probability of getting scooped between submission and publication, much larger than the three percent chance we document in the PDB data. We then ask them to predict the probability of publication and expected citations if they are scooped by a competitor's paper. They predict that they only have a 66 percent chance of publishing the paper, again much lower than the 86 percent of scooped projects that we observe being published in the PDB data. Finally, they estimate a 59 percent penalty in citations compared to the hypothetical winner, much higher than the 20 percent penalty we estimate in the PDB data.[2] These comparisons suggest that scientists may be overly concerned about the probability and cost of getting scooped, and perhaps better information about the true outcome of races might alleviate concerns about risk and competition in academia.

We choose to focus on structural biology because the unique features of the PDB allow us

[2] We also estimate these numbers in a subsample of the PDB data that is most similar to the hypothetical posed in the survey and still find evidence of pessimism. See Table 2.8 for details.

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

In a parallel paper, we focus more broadly on the welfare implications of scientific races (Hill and Stein 2020a), using the estimates from the PDB as an empirical benchmark for the returns to priority. Is the observed difference in priority rewards between winners and losers too large or too small from a social welfare perspective? On one hand, explicitly rewarding priority encourages scientist effort and the timely disclosure of scientific results. On the other hand, sharp priority rewards (or the perception of a scoop penalty) may cause scientists to rush to publication at the expense of the quality of scientific research or the transparency of scientific communication. Indeed, the share of credit given to scooped articles is a salient policy lever for journal editors and funding organizations. Some journals have begun to explicitly offer a grace period where they will consider scooped papers for publication (PLOS Biology Staff Editors 2018,Marder 2017). These policies are aimed toward easing concerns about priority and reducing the risk that scientists face when embarking on competitive projects.

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

Priority races in science are often compared to patent races in industry. However, incentives for basic scientific advances are in many ways distinct from patents. Inventors in a patent race are competing for profits, while researchers in a priority race are competing for journal placement, citations, and recognition from their peers. However, both systems compensate researchers for the production of public goods, incentivize timely disclosure of knowledge, and hasten the pace of discovery. Both systems are usually conceptualized as tournaments for a discrete innovation reward or prize, with the first innovator getting the outsized share of rewards. Theoretical models of patent races have considered how racing affects the amount of R&D investment (Loury 1979; Lee and Wilde 1980) as well as the pace of research and the amount of risk-taking induced by the structure of races (Dasgupta and Stiglitz 1980). Many of these models pre-suppose a winner-take-all reward, which has implications for the outcome of innovation tournaments and the strategic behavior of the participants. The conventional wisdom in the sciences — and the assumption underlying much of the theoretical economics work on the topic — is that the process of scientific discovery is also a winner-take-all tournament, even if the prize is priority recognition rather than a patent. (Merton 1957; Dasgupta and David 1994; Stephan 1996). This reward structure again has implications for the pace of research and the strategic interaction of teams (Bobtcheff et al. 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) studies the disk drive industry in the 1970s and 1980s to test predictions about competing firms' strategic behavior, and finds that firms lagging behind the leader are most likely to innovate. Most related to our work, Thompson and Kuhn (2017) document that winners of patent races do more innovation in the future, and that this innovation is more likely to be related to the original patent. The authors identify patent races by looking for patents that were rejected for lack of novelty. Bikard (2013) studies the phenomenon of simultaneous discovery in science, and documents many cases of papers that are similar in content, are published around the same time, and are frequently cited together. However, our method of using biological details to link competing papers allows us to find simultaneous discoveries where one paper goes unpublished or is cited infrequently in the future.

Our estimates also contribute to work in sociology and economics about how academic reputation interacts with future success. The Matthew Effect, first described by Merton (1968), is a model of path-dependent advantage, whereby success begets future success through increased name recognition, resources, and opportunities. Recent empirical work has documented evidence of the Matthew Effect in science. Azoulay et al. (2013) find that life scientists who win a prestigious award experience a "boost" in citations to their pre-award work relative to similar scientists. Hill (2019) finds that astronomers who experience exogenous bad-weather shocks during their telescope observations publish at lower rates in the future, with larger effects for novice researchers. Jacob and Lefgren (2011) and Bol et al. (2018) find that narrowly winning a post-doc grant early in the career can increase profile and accelerate productivity relative to applicants who were narrowly rejected. On the other hand, Wang et al. (2019) find that near-miss rejections from R01 NIH grants lead scientists to produce more impactful and creative work. They attribute this effect to "grit" or other internal motivation to overcome professional setbacks, which is also a possible response to being scooped that would counter the negative effects of being scooped in the long run. Although scientists may value attention and prestige intrinsically, journal placement and citations also translate to monetary gain in the form of grants, tenure promotions, and salary increases (Hamermesh and Pfann 2012; Ellison 2013). Our estimates of the long-run consequences of getting scooped confirms that there is some amplification of citations after a successful project in our setting. The evidence we present of asymmetric credit for high- and low-reputation teams also agrees with the notion that superstar scientists may be rewarded as much for their past productivity as for their current output.

The remainder of the paper proceeds as follows. Section 2.2 provides some scientific background and a description of our data. Section 2.3 describes the empirical design and identification. Section 2.4 presents results for the short-run impact on publication, journal placement, citations, and alternative attention metrics as well as the long-run career results. We also discuss the role of editors and the timing of races for the distribution of priority rewards. Section 2.5 describes a model of academic attention and reports results for heterogeneity of the scoop penalty by pre-existing reputation. Section 2.6 benchmarks the size of our estimates against the beliefs of surveyed structural biologists about the probability and cost of getting scooped. Section 2.7 concludes.

2.2 Background and Data Construction

2.2.1 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 2019b). Although knowledge about protein structures is useful for applied technologies, the discovery of the structure itself is not patentable.[3] New structures are usually solved by academic researchers at universities or research centers, although 15 percent of the scientists in our sample work at non-profit research laboratories or private companies.

2.2.2 The Protein Data Bank

We focus on structural biology because the Protein Data Bank (PDB) contains detailed, organized, and comprehensive project-level data that is publicly available. The PDB is a worldwide repository of biological macromolecule structures, 95 percent of which are proteins.[4] The PDB was established in 1971 at Brookhaven National Laboratories, with just seven structures. Today, the PDB contains over 150,000 macromolecule structures, and is growing at a rate of about ten percent annually (Berman et al. 2000; Burley et al. 2019). Since the early 1990s, the majority of scientific journals have required that any published structures be deposited in the PDB (Barinaga 1989; Berman et al. 2000, 2016). Furthermore, in 1998, top journals including *Science*, *Nature*, and *PNAS* formalized a policy to ensure simultaneous release of academic papers and PDB details (Campbell 1998; Sussman 1998) as encouraged by the PDB and the International Union of Crystallography.

Because of these strict public disclosure policies, we believe the PDB represents a near-complete census of macromolecule structure discoveries. Whenever a structural biologist completes a project, they upload the structure, experiment, and discovery details to the PDB. This typically happens

[3] The 2013 Supreme Court ruling on the *Association for Molecular Pathology versus Myriad Genetics Inc.* case precludes patents on naturally occurring products such as proteins, genes, and bacteria in the United States. However, even prior to this ruling, patents on the 3D structure of proteins were rare and difficult to obtain (Seide and Russo, 2002; Shimbo et al., 2004).

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

shortly before or after they submit an academic paper describing their findings for publication. An important feature of this process is that the uploaded data is confidential. No other user of the PDB can access the data or see that the deposit has been created. Even the editor and reviewers only receive a receipt of deposit from the PDB and author, and they do not see the underlying structure data until the date of publication. Only at the point of publication is the data released to the public. If any project goes unpublished, the data is released by default after one year (wwPDB 2019).

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

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

2.2.3 Identifying Priority Races: Challenges and Solutions

Identifying priority races in scientific data is difficult for three reasons. First, questions should be well-defined and have a common approach to solving the problem. To underscore the importance of this requirement, consider economics, a field where this is *not* the case. There are many papers on the same topic or question (e.g., what is the effect of raising the minimum wage on employment?), which are often published in close succession (for example, Jardim et al. 2018 and Cengiz et al. 2019). And yet, because there are a variety of methods, settings, and approaches, these papers may be quite distinct. Therefore, the first paper to be published does not necessarily "scoop" subsequent papers that aim to answer the same question. For our purposes, we need a field where the questions are tightly defined with a common approach, a feature that seems more common in the hard sciences than the social sciences. The second challenge is identifying papers that answer the same question. Manually comparing papers to decide whether they address the same question is infeasible at scale. Ideally, we would have some objective measure of scientific proximity, which can tell us whether two teams are working on the identical problem. Finally, the third challenge is that scooped papers are often abandoned without publication. If authors abandon their projects when they see that a similar paper has been published, many scooped papers will never show up in bibliographic data.

The PDB enables us to make significant progress on these three obstacles. First, the questions in structural biology are well-defined, because scientists are typically trying to solve the structure of a known protein. Moreover, the methods are consistent: 85 percent of proteins are solved using x-ray crystallography. This means that if we observe two papers that study the structure of the same protein, these two papers are likely to be very similar in terms of the question, methods, and conclusions. Second, as mentioned in Section 2.2.2, the PDB measures how biologically similar different proteins are to one another. This allows us to link projects based on objective measures of scientific proximity rather than text similarity or citation behavior. Finally, scientists are required to deposit their structures in the PDB *prior* to publication. This gives us the ability to observe some projects that never reach publication. Given that scientists might abandon projects that get scooped, having this record of unpublished projects is a key feature of our data. We will discuss the timeline in more detail in the next section. To the best of our knowledge, we are the first to measure scientific races in a data-driven manner.[5]

2.2.4 Defining Priority Races

Broadly speaking, we define a priority race as an instance where two or more teams are working on the same protein independently and concurrently and are likely uncertain about the identity or progress of their competitors. Following Brown and Ramaswamy (2007), we define "same protein" as meaning two proteins within the same 50 percent or higher sequence similarity group (called a "cluster" in the PDB). This is a conservative cutoff, as 30 percent has been suggested as sufficient similarity for building homology models (Dessailly et al. 2009; Moult 2005). In other words, the first deposit within these 50 percent similarity clusters are highly cited because they provide a novel structure model that other crystallographers can build on to solve very similar proteins. For robustness, we can restrict to scoops by proteins within the same 100 percent cluster, and find similar results which we report in Appendix Table B3. [6] The PDB assigns ID numbers to clusters of similar proteins, and we say that the first deposit released in that cluster is the "priority" deposit. There are often many subsequent deposits that report similar structure coordinates as the priority deposit. These follow-on deposits are either scooped projects, replication projects of the same protein by future teams, or new projects that solve the structure for closely related proteins from different organisms or bonded with different macromolecules in a novel way.[7]

We use the timing to determine whether a follow-on deposit qualifies as scooped by the priority deposit. The PDB provides two key dates at the structure level that outline the timeline of each project and help us determine whether two teams are working concurrently: the deposit date and

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

[6] If a protein is scooped by more than one other protein, we give preference to the protein that is biologically closer (i.e. in the "higher" cluster). See Appendix B.2 for details on the data construction.

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

release date.[8] The deposit date corresponds to the date that the scientist uploaded her solved structure to the PDB. Importantly, the structure is not yet visible to the public. Nearly all scientific journals require that authors upload their structures to the PDB prior to publication, so deposit typically occurs slightly before or after the date that the scientist first submitted their paper. The release date is the date that the PDB deposit is made public. This typically corresponds to the publication date. In cases where the structure is never published, the PDB releases the deposit by default one year after the deposit date. Figure 2.1 provides a visual timeline of these dates, as well as some summary statistics. Throughout this analysis we will always use the release date as the relevant marker of priority. An alternative approach would be to use paper publication dates to determine priority ordering. But these dates are often unavailable, especially for older publications, or are ambiguous in recent data because online publication may come before print edition publication. Further, we treat publication as an outcome variable, leading to potential bias if we condition on publication as a requirement for treatment assignment. Lastly, PDB releases tend to be publicly salient dates that the community pays attention to, so we are comfortable using these dates to mark priority. Appendix Section B.1.4 discusses implications and presents evidence about the concordance between release dates and publication dates in greater detail.

Figure 2.2 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,

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

and 5.8 percent of the races we identify have three or more competitors. Appendix Section B.2 provides a more detailed description of the data work necessary to construct these races in practice.

An important caveat to our approach is that we can only identify races that were "close" enough that both teams had already completed a structure determination and were preparing to publish. Some scientists may claim they were "scooped" if they were working on an incomplete project when another team published a solution first. We cannot observe their setback if they abandoned the project before completion, nor can we infer their counterfactual probability of success had they not been scooped. Therefore our approach specifically identifies the cost of being scooped when both teams are near the finish line. This effect may be smaller or larger than the effect of being scooped earlier in the scientific process.

An Example

To help understand our procedure, consider an example outlined in Table 2.1. 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 2.3 shows the nearly identical biological assembly models that each team deposited independently and confidentially to the PDB. The scientists at Leiden University (3W9C) collected their data a few months before the scientists at University of California, Irvine (4JWS) (February 3, 2012 versus September 14, 2012). However, by the time of deposit, the UC Irvine team had pulled ahead, depositing one week before the Leiden team (March 27, 2013 versus April 3, 2013). Ultimately, UC Irvine won the priority race, with their structure being released two months before Leiden (June 19, 2013 versus August 21, 2013). Importantly, when Leiden deposited their structure on April 3, 2013, UC Irvine had not yet released their structure. This means that Leiden was likely unaware of their competitor's progress or results when they were preparing their publication and depositing the structure. Comparing the outcomes of the winner (4JWS) and the loser (3W9C), we observe that the winning paper was more successful. It was published in a better journal (*Science*, with an impact factor of 31.5 versus

Journal of Molecular Biology, with an impact factor of 4.0) and received about 30 percent more citations over the next five years (Tripathi et al. 2013; Hiruma et al. 2013). In this case, the Leiden authors became aware that they were scooped during the manuscript review. In the conclusion of their paper, they write, "While this manuscript was under review, Tripathi et al. published the crystal structure of the Pdx–P450cam complex that was obtained via cross-linking of the two proteins. It is interesting to compare our complex with those reported in that study. Tripathi et al. found a position and orientation of Pdx relative to P450cam that is essentially identical with ours." (Hiruma et al. 2013) [9]

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.[10] Second, we drop projects that are published in a paper that is linked to 15 or more other structures. Among the set of papers included in our final analysis sample, 46 percent are linked to more than one structure, and the average number of structures per paper is 1.9. Multi-structure papers are at risk of being scooped on a fraction of the full project. This restriction allows for some fractional scoops to enter our data, but ignores papers where each protein becomes a very small fraction of the full contribution of the paper. Finally, we drop races that end in a near or exact tie. Occasionally, two racing papers will be submitted to the same journal and the editor will publish

[9] 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 10 percent cite the winning paper. That probability increases to 60 percent for races with an eight month gap between release dates. See Appendix Figure B1.

[10] 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.2 describes in more detail how we deal with multi-entity structures in the data.

them as companion pieces in the same issue, and we drop these cases. We also drop races where the two papers were released closer than two weeks apart from each other. We make this restriction to help ensure that the first project has a clear claim of priority and that the order of release is more likely to correspond to the order of publication.[11]

2.2.5 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 B.1.

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:

$$JIF^j_t = \frac{Cites^j_{t,t-1} + Cites^j_{t,t-2}}{Articles^j_{t-1} + Articles^j_{t-2}}. \tag{2.1}$$

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

[11] 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 B.1.4 for more details on the correspondence between the PDB release date and publication date.

journals as well as top-ranked life science journals.[12]

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

We construct career histories of variables before and after the priority date of each race to serve as control variables and long-run outcomes. Reconstructing publication records for individual authors is difficult because names are not disambiguated in the PubMed or PDB. We use a dataset called Author-ity, which groups PubMed IDs into distinct author identifiers using co-author and topic patterns (Torvik et al. 2005; Torvik and Smalheiser 2009). However, because not all PDB deposits are published, it is hard to link unpublished deposits to the correct name identity in Author-ity. Therefore, in the long-run results section, we restrict to a subset of authors that have uncommon names and uniquely match to an individual in Author-ity. We also use simple name-matching techniques within the PDB to construct control variables of team productivity prior to treatment, which we can do for all deposits including those that are not published. We describe the name disambiguation procedures in detail in Appendix B.1.6.

For long-run outcomes, we count PubMed publications, PDB-linked publications, top-10 publications,

[12] 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.*

citation-weighted publications, and "hit" publications for the years following the treatment date. Besides analyzing the effects of race outcomes on the intensive margin of publication, we also consider the extensive margin of exit from publishing PubMed papers and PDB-linked papers altogether. We mark an individual as having exited academia if there is a hiatus of at least five years in their publication record that begins in the five years after the priority date. Similarly, we identify individuals that exited structural biology (either changed fields or left academia) as those that have a hiatus of publishing PDB-linked papers in the following five years.

Altmetric.com Getting scooped may not only affect traditional publication outcomes like journal placement and citations, but also the overall engagement with the research by the academic community and general public. There have been many recent efforts to measure broader sources of academic impact by counting metrics such as news and social media engagement, patent citations, and online downloads and readership. We link the PubMed papers in our sample to data provided by Altmetric.com. In Section 2.4.2, we examine the effect of getting scooped in recent years on these non-traditional measures, including Mendeley downloads (a popular citation management software), news article citations, Wikipedia citations, patent citations, Twitter.com mentions, and a composite measure of attention called the Altmetric Attention Score.

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

Editorial Dates In Section 2.4.4, we analyze how the scoop penalty is affected by the timing of the scoop event relative to the journal review and publication timeline. We supplement our data with the received, accepted, and publication dates for papers published in journals owned by a handful of large publishers. While we were not able to obtain these dates for all articles, we chose to focus on journals based on their prevalence in the PDB and the availability of the data for download. The journals included in the subsample are flagship or field journals from the following journal groups: Science, Nature Journals, Cell Press, and Public Library of Science (PLOS). This subsample covers 19 percent of our primary regression sample.

Scientist Survey In order to benchmark the magnitudes of our findings, we surveyed structural biologists about their perceptions of the probability and costs of getting scooped. Email surveys were conducted in September of 2019. We collected email addresses from the Web of Science, which provides a contact email for many of the corresponding authors on academic publications. The recruitment sample was defined as any corresponding author on a PDB-linked publication from 2014-2019 that had an email address available in the Web of Science files. We sent recruitment emails to 8,984 unique email addresses, and encouraged respondents to participate on a volunteer basis. We received 915 responses, for a total response rate of 10.2 percent. Each potential recruit received one initial solicitation and two follow-up reminders to complete the survey. Relevant text of the questionnaire is provided in Appendix B.3.

2.2.6 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 2.2 shows summary statistics at the structure level for both of these samples. Just over five percent of the structures in our sample are involved in a priority race. We look at both team characteristics and deposit outcomes. Teams involved in priority races tend to be smaller, younger, and more likely to come from a top university. The racing scientists were also more likely to work in Asia, and less likely in North America. The deposit outcomes suggest that proteins involved in priority races are scientifically more important. Proteins in the racing sample are more likely to be published, appear in higher-ranked journals, and receive more citations.

2.3 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.

2.3.1 Baseline Specification

Equation 2.2 presents the basic specification for the project-level regressions. For deposit i studying protein p, we estimate

$$Y_{ip} = \alpha + \beta Scooped_{ip} + \mathbf{X_{ip}'}\delta + \gamma_p + \varepsilon_{ip} \tag{2.2}$$

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

In Section 2.4.3, we consider the long-run effect of getting scooped on academic career outcomes. The regression specification is similar to equation 2.2, 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

$$Y_{isp} = \alpha + \beta Scooped_{isp} + \mathbf{X'_{isp}}\delta + \gamma_p + \varepsilon_{isp} \tag{2.3}$$

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

2.3.2 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 2.3. Using the information disclosed by the teams in the PDB, we inspect a variety of observable characteristics that might reasonably be correlated with the probability of treatment or with outcomes. These include the number of authors, the location of the lab, the rank of the university affiliation, and the experience in years of the first and last authors. We also calculate measures of the authors' productivity in PDB-related publications in the five years prior to the racing deposits. These include the number of PDB deposits, publications, and publications in top-ranked journals.[13]

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

Table 2.3 shows the mean values of each covariate for the winning and losing teams, as well as for the teams in the non-racing sample, for reference. We report test statistics for the difference in means between the winning and losing teams, as well as an F-statistic for a test of joint significance of all covariates. We find that many of the covariates are balanced between the winning and losing teams. But winning and losing teams are statistically different in a few notable dimensions. North American and European teams are more likely to win than lose, while Asian teams are more likely to lose than win. Scientists from top-50 ranked universities are more likely to win, as well as first authors with slightly less experience. The prior productivity of these labs is more balanced, with both the first and last authors having almost identical numbers of deposits and publications. We also test whether the scientific results that are being deposited by both teams are similar. Refinement resolution and R-free are two variables reported by the PDB that describe the objective quality of the experimental data and model in each deposit. Resolution describes the degree of precision in the diffraction data produced during crystallography experiments, and R-free measures the goodness-of-fit between the experimental data and the proposed structure model. For both of these measures, smaller values imply better quality. These two measures are very close to balanced between winners and losers, suggesting that the quality of the science or the skill of the scientists is likely not driving our results. Taking the table as a whole, we reject the null hypothesis of balance on the full battery of covariates based on an F-statistic of 3.91.

Unbalanced covariates lead to biased estimates only if they are systematically correlated with the outcome variable. Therefore, to further assess potential selection bias, we visually inspect the difference in expected citations between winners and losers. We estimate a model of citations using a Lasso[14] regression of five-year citation counts on the battery of team covariates. This model is estimated only in the sample of non-racing deposits. We then take the selected variables and estimated coefficients to predict citations in the racing sample in a post-Lasso OLS procedure. The covariates we include are counts of publications, citations, and journal placements in the five years prior to the deposit for the first and last author, as well as the squares of these variables. We

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

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 B1.

Figure 2.4 plots a histogram of the difference in predicted citations between each pair of winning and losing teams (races with three or more teams are omitted here). A perfectly balanced sample would be centered around zero and symmetric. If winners were systematically better-resourced, higher reputation, or more experienced, then the histogram would be skewed to the right. As a benchmark for perfect balance, we compare this distribution to a simulated distribution where we randomly assign one of the paired teams as the winner. We simulate this coin flip 100 times per pair. The true distribution is shifted slightly to the right of the randomly simulated distribution, suggesting that winners are slightly more likely to be high-reputation than would be predicted by chance. But the differences in the distribution are minimal, with an average difference in predicted citations of 0.21 citations (p-value of 0.587). We can also compare the distributions with a Kolmogorov-Smirnov test and calculate a test statistic of 0.040 with a p-value of 0.240. Therefore we fail to reject the hypothesis that the difference between these two histograms is different than zero. While winners and losers of priority races are not identical in observables, their differences appear to have very little systematic effect on our measures of project success.

2.3.3 Control Selection Using Post-double-selection Lasso

In light of potential treatment imbalance, we rely on an identification assumption that treatment is exogenous conditional on observable control variables. There are many potential control variables in our data, so we use a method called post-double-selection Lasso (PDS-Lasso) proposed by Belloni et al. (2014) to optimally select controls variables. Consider a partially linear model similar to equation 2.2

$$Y_{ip} = \alpha + \beta Scooped_{ip} + \mathbf{g}(\mathbf{Z_{ip}}) + \gamma_p + \varepsilon_{ip} \tag{2.4}$$

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

The PDS-Lasso method uses two steps. First, it estimates a Lasso regression of $Scooped_{ip}$ on $\mathbf{Z_{ip}}$ to select a set of regressors that are predictive of treatment. Then it uses a second Lasso regression of Y_{ip} on $\mathbf{Z_{ip}}$ to select regressors that are predictive of the dependent variable. The selected control variables are highly informative of treatment assignment and outcomes, and therefore reduce bias in estimation. The superset of selected regressors from those two regressions are used as the control variables in a post-OLS regression of Y_{ip} on $Scooped_{ip}$. The potential set of regressors we use are the variables in the balance Table 2.3 as well as squares of those variables and university rank binned into 10 school dummies. The protein fixed effects γ_p are included as unpenalized regressors in all steps of the method.

2.4 Results

2.4.1 Short-run Effect on Projects

Table 2.4 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[15], 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

[15] 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).

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 2.2 with no control variables. Panel B adds all controls listed in Table 2.3, and panel C uses controls selected from the PDS-Lasso procedure described in Section 2.3.3. 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. To further test for selection bias on unobservables, we implement a robustness check following Oster (2019) in Appendix Table B2[16]. We will use panel C as the preferred specification to report our estimates throughout the paper.

Scooped projects are 2.5 percentage points less likely to be published off of a baseline publication rate for winning projects of 88 percent. This represents a 3 percent decrease in probability of publishing, or framed differently, a 21 percent increase in the probability of abandoning the project. This modest discouragement rate is likely driven by the low cost of publishing once the project has already been deposited in the PDB (recall that in our sample, all scooped projects have already been deposited in the PDB when they learn that they have been scooped). In many cases, the scooped teams may be well into their submission and revision process at the time of being scooped, and therefore will persist to publication. Even if they are rejected from a journal, there are many lower-ranked outlets that may be more willing to accept scooped papers, a mechanism we explore in Section 2.4.4.

In column 2, we estimate a statistically significant penalty in journal impact factor. Scooped

[16] Adding controls and protein fixed effects increases the R^2 from less than 0.01 to over 0.60 in all regressions, suggesting that most of the variance in the outcome is explained by treatment and observable controls. Implementing the suggested bias adjustment, we conservatively assume a maximum $R^2 = 1$ and $\delta = 1$ (unobservables are equally important for treatment selection as observables), and find that the adjusted coefficients are almost identical to our baseline findings. Further, the δ needed to reduce the estimate to zero is greater than 7 in all specifications, meaning there would need to be an unrealistic degree of selection on unobservables to threaten the robustness of the results.

papers are published in journals with impact factors 0.18 standard deviations below winning papers. In column 3, this translates to a 6 percentage point (18 percent) decrease in the probability of publishing in a top-ten journal. Column 4 shows that scooped papers face a significant citation penalty as well. The winning projects receive 29 citations on average in the first five years. The scooped projects receive 20 percent fewer citations in the same time span. Column 5 suggests that this means scooped projects are 3.5 percentage points (23 percent) less likely to be one of the top 10 percent of papers in that publication year ranked by five-year citations. These results are robust to a variety of cutoffs, including a shorter or longer citation window and different percentiles for the high-citation mark. As a further robustness check, we reproduce this table using a sub-sample of races that have projects with 100 percent similar sequence structure according to the algorithm used by the PDB. Appendix Table B3 shows that the magnitudes are very similar for all outcomes, even if statistical precision is lower due to the smaller sample size.

Taken together, these results suggest that there is a significant penalty for being scooped, both in the likelihood of publication, the journal rank of publication, and the number of citations accrued in the early life cycle. However, these results also indicate that the rewards for priority are not winner-take-all. Losing teams receive a smaller, but still substantial share of the credit as measured by publication and citations. Translating the citation penalty to shares of total citations, losing projects receive approximately 44.5 percent of the total citations accrued to both papers, a much larger share of credit than zero percent for the winner as is typically assumed by classic models of innovation races.[17]

2.4.2 Alternative Measures of Attention

Scooped projects may not only be penalized in terms of journal placement and citations, but also by less formal means of recognition, such as reader downloads, coverage in the scientific press, and mentions on social media. Scientists value these interactions as they build standing and

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

reputation in both the academic community and general public. Table 2.5 shows results of project-level regressions using outcomes sourced from Altmetric.com. In these regressions, we restrict the sample period to 2011-2017 since many of these outcomes are only relevant in the recent internet era. All outcomes are count variables again transformed with the inverse hyperbolic sine function to deal with skewness and facilitate proportional interpretation of the effects. Regression results are again reported with the three different control strategies used in Table 2.4.

Column 1 of Table 2.5 reports the effect of getting scooped on Mendeley readership. Mendeley is a popular citation manager used by many researchers. Downloading a paper on Mendeley can be interpreted as a proxy for popularity of a paper among readers, and especially those readers that might consider citing the paper at some point. Focusing on panel C, getting scooped leads to an approximately 45 percent decline in Mendeley downloads, which is quite a bit larger than the citation penalty reported in Table 2.4. News stories covering the academic articles fall by 11 percent for scooped papers, and Wikipedia citations fall by 3.5 percent. There is no detectable effect on patent citations. Mentions of a paper on Twitter fall by 10 percent, although this estimate is only marginally significant and not robust to all control strategies. Altmetric.com provides a comprehensive score of alternative attention (Huang et al. 2018), which falls by 24 percent for scooped papers. These results suggest that getting scooped has different effects for different audiences. The large effect on readership proxied by Mendeley suggests that scientists who casually interact with the research are more prone to focus on only the race winners. This is likely driven in part by journal placement, where some scientists stay abreast of advances in various fields by only reading papers that appear in the top general interest or field journals. Science reporters in the news tend to be less responsive to priority ordering, suggesting that they might be more likely to cover both papers about a topic instead of just the first paper. Some of the most specialized readers, such as Wikipedia contributors and patent citers seem to be the least responsive, suggesting that they do a much deeper literature search when citing academic papers.

2.4.3 Long-run Effect on Authors

In this section we analyze the long-run consequences of being scooped on the careers of the various authors of scooped papers following equation 2.3. Table 2.6 reports the results of the long-run outcomes regression. Panel A contains results for regressions in the full sample of authors. Panel B restricts to novices only, which are defined as authors who had seven years or less since their first publication at the time of the scooping event.[18] Panel C restricts to veterans, which are all scientists not defined as novices.[19]

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

We find no significant changes to publishing on the intensive margin for novices or veterans. Losing teams have no statistically significant differences in publications or PDB-linked publications in the following years as shown in column 3 and 4, and they are not more or less likely to publish in top-10 journals. However, we do estimate significant penalties in citations for all categories of authors. In the full author sample, the scooped individuals receive 17 percent fewer citations (measured by inverse hyperbolic sine citation-weighted publications) in the next five years, where

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

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

citations are counted up to three years after each paper's publication. This effect falls particularly hard on novices, who receive 32 percent fewer citations, while veterans receive only 13 percent fewer citations. The effect on "hit" papers is reported in column 7 and also suggests that getting scooped decreases attention to future work. The full sample of scientists publish 0.42 fewer hit papers in the five years following a scoop event. The negative effect is lower for novices in levels (0.10 papers versus 0.58 papers for veterans), and not statistically significant for novices. However, if we scale the effect size by the average number of hit papers, the effect is larger for novices (an eight percent decline versus a six percent decline). We also consider outcomes in the following three years in Appendix Table B4 and ten years in Appendix Table B5. 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.

2.4.4 Mechanisms: Role of Scoop Timing in the Publication Process

Scooped projects receive 20 percent fewer citations than their winning counterparts, suggesting that academic researchers pay less attention to the projects that are scooped. In this section, we investigate how the editorial process affects the scoop penalty, and we argue that journal placement is a primary driver of the citation penalty. Further, the size of the penalty is highly correlated with the timing of races. Teams that are scooped early (very shortly after they deposit their findings) receive a much larger penalty than teams that are scooped late (shortly before publication). We provide evidence that top journal editors are unlikely to accept scooped papers, therefore scooped papers consistently fall to lower-ranked journals unless they were deep into the review process at the time they were scooped. These results suggest that editors and reviewers are key policymakers in determining the distribution of academic credit for novel research.

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 2.7 replicates the citation penalty effect from Table 2.4, column 4, but uses a subsample of races where both papers were published in ranked journals. When both papers are published, the citation penalty is 16 percent for scooped papers. In columns 2 and 3, we add controls for journal impact factor, first as a linear term and then as a cubic polynomial. The citation effect falls to 11 percent, but remains statistically significant. Finally, in column 4 we include journal fixed effects to control completely for any direct effect of the publication outlet on citations. The effect falls to five percent. These results suggest that at least two thirds of the citation penalty comes through the channel of the publishing journal. Any remaining effect on citation attention comes through readers differentially citing winning and losing papers in similar journals.

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."[20] 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

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

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 2.6 further shows that this pattern varies greatly by the impact factor of the journal that eventually publishes the scooped paper. For lower ranked journals, such as *PLOS One*, only 60 percent of scooped papers had been received by the journal on the date they were scooped, and just over 20 percent had been accepted. Among the 11 large journals for which we have information about received and accepted dates, there is a positive and statistically significant relationship between the share accepted before the scoop date and the impact factor, with a one standard deviation higher ranked journal being eight percentage points more likely to have already been accepted on the scoop date. Although we cannot directly observe scooped papers being rejected from these journals, we can infer from this pattern that top journals are less willing to accept papers that were scooped before submission or early in the review process. Many of these scooped papers fall to lower ranked general interest journals or highly specialized structural biology journals. Some of these lower-ranked journals, such as *PLOS Biology*, have explicit policies of accepting scooped papers. *PLOS Biology* editors write, “Just as summiting Everest second is still an incredible achievement, so too, we believe, is the scientific research resulting from a group who have (perhaps inadvertently) replicated the important findings of another group. To recognize this, we are formalizing a policy whereby 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.

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 2.5, we plot the difference in outcomes separately for three terciles of races divided by the time between the release dates of the winning and losing projects. The points are placed on the x-axis at the average delay time within the subset of races. The first panel shows the journal impact factor penalty and the second panel shows the citation penalty. Both plots have a strong decreasing trend in the penalty — in other words, the longer the lag between the priority paper and the scooped paper, the less credit the scooped paper receives. The journal impact factor penalty is 0.1 standard deviations in the first three to four months, then drops to 0.3 standard deviations by eight months. Similarly, projects released within one month of each other have no difference in citations. The scoop penalty grows to 50 percent for scooped projects with an eight month delay. In fact, much of the negative effect that we present in Table 2.4 is driven by the tercile of races with the longest delays. An important caveat to these results is that the delay to release after being scooped is potentially endogenous. Teams likely make strategic decisions to rush to publish, revise and delay, or abandon altogether, so the delay times should be viewed as potentially selected on team or project characteristics. These results suggest, however, that the delay time between projects is relevant for editors and readers, perhaps because the community can more clearly attribute priority credit with more time separating similar projects.

2.5 Reputation and the Scoop Penalty

In this section we show that academic recognition is affected not only by priority, but also by the preexisting reputation of winners and losers. When a high-status team scoops a low-status team, they receive 66 percent of the total citations, but when a low-status team scoops a high-status team in a comparable race, they only receive 46 percent of the the total citations. This asymmetry in attention suggests that the distribution of priority rewards is not formulaic and may be affected by the institutions and norms of the academic community. We propose a model of academic attention based on a standard statistical discrimination model (Aigner and Cain, 1977) and present empirical results that support the predictions of the model. Priority rewards are allocated by a decentralized set of actors, including journal editors and readers, in a market for academic attention. Because scientists have limited time for reading and reviewing new papers, it may be difficult to determine the quality of new research. Therefore, editors and readers may rely on signals of ability based on the reputation of the researchers or their institution to supplement their judgement of a paper's quality.

2.5.1 A Model of Academic Attention

Setup

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

$$s = q + u$$

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

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

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

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. See Appendix B.4. The intuition is that if the labs are the same type, there is no differential reputation effect. Therefore, citations are driven solely by the priority effect.

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

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

Proof. See Appendix B.4. 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.

2.5.2 Priority and Academic Reputation

To test our model, we measure the share of total citations received by winning and losing labs, and compare these shares in races where the reputation varies between the two racing teams. More specifically, if lab A and lab B race to write a paper about the same protein, we compute $CitationShare_A = Citations_A/(Citations_A + Citations_B)$. This citation share maps to the probability of citation outlined in the model above.[21]

We proxy for the pre-existing "reputation" of each lab using the Lasso-estimated predicted

[21] 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.

citations from the non-racing data sample as described in Section 3.1.1. Labs with above-median predicted citations correspond to the *H* labs, while teams below median correspond to the *L* labs. In Figure 2.7 we plot the predicted citations of the losers on the x-axis and the predicted citations of the corresponding winners on the y-axis. Each point on this scatter plot represents the observed match between two racing labs. If all labs were equally matched in pre-existing reputation, all points would lie on the dashed 45-degree line. Of course labs are rarely perfectly matched in the data, providing variation in the difference of reputation between the winners and losers.

The median lines in Figure 2.7 conveniently partition the sample into four sub-samples that line up with the four types of "matchups" we discuss in our model. The top right and bottom left corners represent subsamples of closely matched races where both labs were either high-reputation or both low-reputation. The top-left and bottom-right subsamples represent mismatched races where an above-median team scooped a below-median team and vice versa.

In mismatched races, we interpret the difference between citations as being caused by an additive effect of priority and reputation. One potential confounder in that interpretation is that high- and low-reputation teams might produce different quality of scientific outputs for the same structure discovery. If *H* teams produce higher quality or more convincing results, then the additional citations they receive may not only be caused by their high-profile reputation. Although it is difficult to quantify all aspects of paper quality, we examine two important measures of quality reported by the PDB: resolution and R-Free (goodness-of-fit), described in more detail in Section 2.3.2. Appendix Table B6 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 2.8 shows the average citation counts by matchup type, as well as the citation shares. Panel A shows the evenly matched races, which isolates the priority effect. As predicted by the model, the winning labs receive more citations. Moreover, if we look at the *share* received by the winning team, we see that it is identical in the *H* versus *H* matchups and the *L* versus *L* matchups (winning team receives 55 percent of the total citations). This is consistent with the prediction from proposition 1.[22]

Panel B shows the unevenly matched races. When an H lab scoops an L lab, the priority effect and the reputation effect work in the same direction. Here we see that, consistent with proposition 2, the winning team receives an even larger share of the total citations (66 percent). Conversely, when an L lab scoops an H lab, the priority effect and the reputation effect move in opposing directions. In this case, it appears that the reputation effect is the stronger of the two, with the winning team receiving less than half (46 percent) of the total citations. Again, this matches the prediction outlined by proposition 2 of the model.

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

2.6 Benchmarking Magnitudes: Survey Results

We estimate that getting scooped causes a decrease in the probability of publication, leads to publication in lower-impact journals, and reduces citations. However, priority races are not winner-take-all. Our citation estimate suggests that winners get 55 percent of the total citations, a far cry from 100 percent as is often assumed in the theoretical literature. But how does this estimated share of credit compare to scientists' beliefs? In an email survey of structural biologists, we pose a hypothetical situation about a late-stage race to publication. The full text of the questions can be

[22] 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.

found in Appendix B.3. 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 2.8 reports the average responses of the biologists and compares them to the magnitudes estimated in the PDB data. The hypothetical scenario in the survey was designed to match the instances of racing that we have in our data. However, because we tried to pose the survey questions as concretely as possible for clarity, the racing situation does not exactly match the average situation in the PDB. In particular, in the survey the losing team is scooped early in the submission process, and the project is very high-quality, with an expected journal placement in *Science*. Therefore we report estimates in column 2 from a subset of the PDB data where (1) the losing team is scooped soon after they deposit their data,[23] and (2) one of the teams published in one of the three highest impact journals (*Science, Nature,* or *Cell*). These restrictions make some of the PDB estimates smaller or larger, but we still consistently find evidence of pessimism among respondents. Surveyed scientists report a 27 percent chance of being scooped between submission and publication, more than double the 8 percent scoop probability in the comparable PDB sample. Six percent of respondents report that they would abandon the project, but only 70 percent think they would succeed at publishing conditional on submitting, suggesting a 66 percent unconditional probability of publishing. This is much lower than the 86 percent of scooped papers that are actually published in the PDB data, and the 97 percent that are published in the comparable subsample. Scientists are very pessimistic about the potential journal placement of scooped papers, expecting that the best journal they could publish in would be almost three standard deviations below *Science*, which has a standardized impact factor of about three in most years. Finally, we ask about expected

[23] 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.

citation effects. When asked to guess the number of citations they would receive compared to the hypothetical winner's 100 citations, the average guess was only 41 citations, which translates to a 59 percent penalty, or a share of 29 percent of the total citations. The corresponding estimate in the PDB is no more than a 20 percent penalty or a 45 percent share. Ultimately, PDB scientists expect much worse consequences from being scooped than can be found in the data.

Table 2.8 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 2.5. Column 4 reports the average responses for below-median reputation scientists, column 5 reports the average responses for above-median reputation scientists, and the difference with standard errors is reported in column 6. High- and low-reputation respondents predict equal probabilities of being scooped. Low-reputation respondents are more pessimistic however about the probability of publishing conditional on being scooped, with seven percentage points lower probability that they will be able to publish their scooped paper. Perhaps surprisingly, both types of respondents had similar expectations for the types of journals that they would publish in, all expecting that the scooped papers would fall to field journals or middling general interest journals with average impact factor. But they again depart on their expected citations, with high-reputation scientists expecting to get about five more citations (nine percent) than low-reputation scientists. This difference in expectations is consistent with our results about the role of reputation in determining priority rewards. Since both types of authors suggest they would submit to similar journals, it may be that the difference in citations is driven by statistical discrimination of editors, reviewers, and readers as explained in the model in Section 2.5. It appears that although all scientists are pessimistic about the cost of getting scooped, less prominent authors are particularly concerned. Our estimates of significant inequality in citation patterns suggest that these beliefs may be justified.

2.7 Conclusion

Priority races are a common feature of academic science, and credit for priority is considered an important motivator for the generation of new knowledge. Yet, we have little empirical evidence on how these priority rewards are structured. Racing is hard to analyze empirically because proximate research projects are difficult to link in data and many scooped projects are abandoned before entering the scientific record. This paper makes progress on these empirical challenges by focusing on project-level data in a setting that captures the near universe of completed projects in structural biology. By linking adjacent projects using biological measures of similarity, we reconstruct races and compare the outcomes of winners and losers, even in cases where the losing project goes unpublished. We find that losing a priority race decreases the probability of publishing by 2.5 percentage points. Conditional on publishing, the scooped papers are less likely to appear in a top journal and receive 20 percent fewer citations than the winning papers. The effect of getting scooped lingers along some dimensions in the years following the event. We find no effect on exiting academia, but a small increase in the probability of exiting the field of structural biology. We also observe that citations decrease for scooped scientists in subsequent work, particularly for novices. Priority rewards are in part dependent on pre-existing reputation. In cases where a high-reputation team is racing against a low-reputation team, priority rewards are unevenly distributed. High-reputation winners receive much more attention than losers. And in cases where the high-reputation team is scooped, the winning low-reputation team receives no more citations than their high-reputation rival.

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

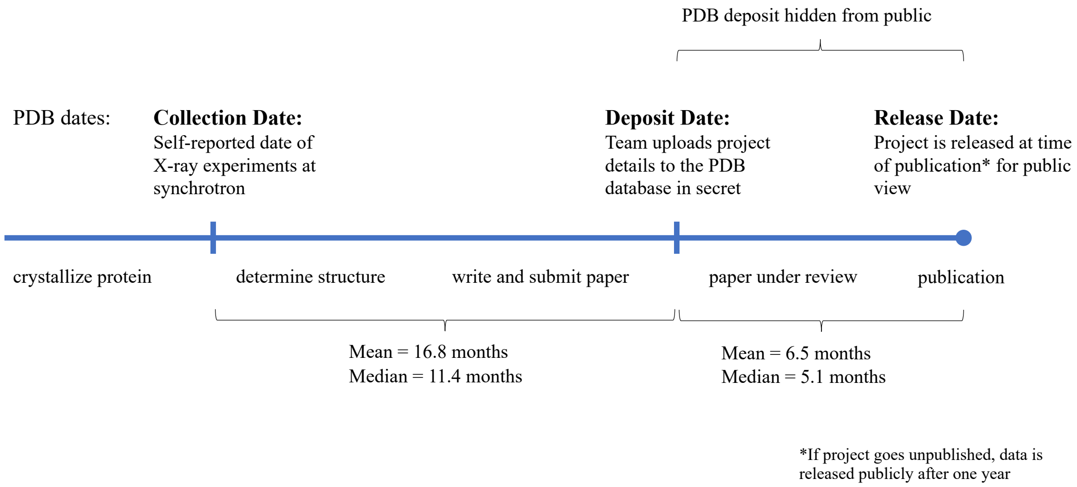
This paper contributes to our understanding of the role of priority and the structure of incentives in basic research. Academic science is an atypical marketplace of productive activity. New ideas are valuable for the world but are not immediately marketable, and are therefore unlikely to be produced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

In this paper, we establish that priority is a relevant incentive in science, but we do not analyze the overall welfare implications of the priority system, or consider alternative systems or policies. An important concern raised in popular and academic writing is the potential "dark side" of priority, where novelty may be pursued at the expense of openness and quality. Racing to complete projects may stimulate effort and hasten the pace of discovery, but it may lead scientists to cut corners on the quality of the results that they disclose. If the incentives for replication are low and the costs of replication are high, science as a whole may suffer as quick and sloppy research becomes the norm. In Hill and Stein (2020a), we analyze objective measures of the quality of crystal diffraction data and corresponding structure models to study how racing in science affects quality outcomes. We find that proteins with high ex-ante potential have more competitors racing to complete the structure, are deposited faster, and are completed with lower quality. This evidence suggests that racing in science does indeed hasten disclosure, but has negative effects on quality. Future work should also focus on how competition affects the openness of science, ease of collaboration, and free transmission of knowledge between scientists. Concerns about the cutthroat nature of racing have led to suggestions of policies that might dampen the strong incentives for novelty. These include allowing a grace period for journal acceptance in a few months after being scooped, providing opportunities to establish priority for early-stage work through pre-prints, or directly incentivizing replication efforts through directed grant funding.

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

Figures and Tables

Figure 2.1: Project Timeline and Key Dates


PDB deposit hidden from public
PDB dates:
Collection Date:
Self-reported date of X-ray experiments at synchrotron
Deposit Date:
Team uploads project details to the PDB database in secret
Release Date:
Project is released at time of publication* for public view
crystallize protein
determine structure
write and submit paper
paper under review
publication
Mean = 16.8 months
Median = 11.4 months
Mean = 6.5 months
Median = 5.1 months
*If project goes unpublished, data is released publicly after one year

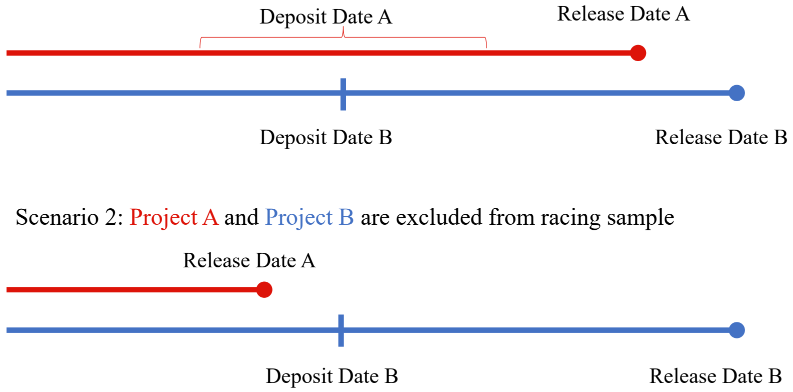
Notes: This figure shows the timeline of a typical PDB project. Dates in bold above the line are observed in our data. Events listed below the timeline are the approximate timing of other project events including the submission and review process. Deposit event and structure data is hidden from public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

Figure 2.2: Defining Priority Races

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

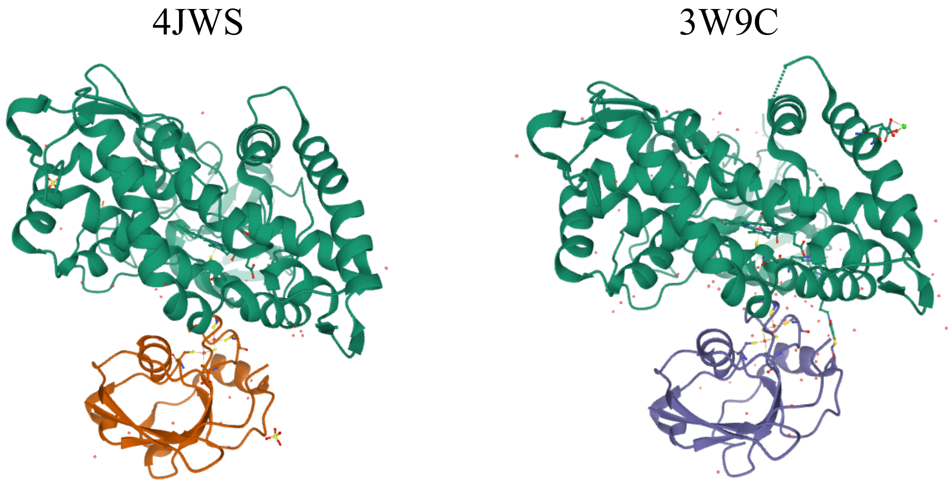
Scenario 1: Project A scoops Project B

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



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

Figure 2.3: Example Priority Race — Pdx-P450cam Complex

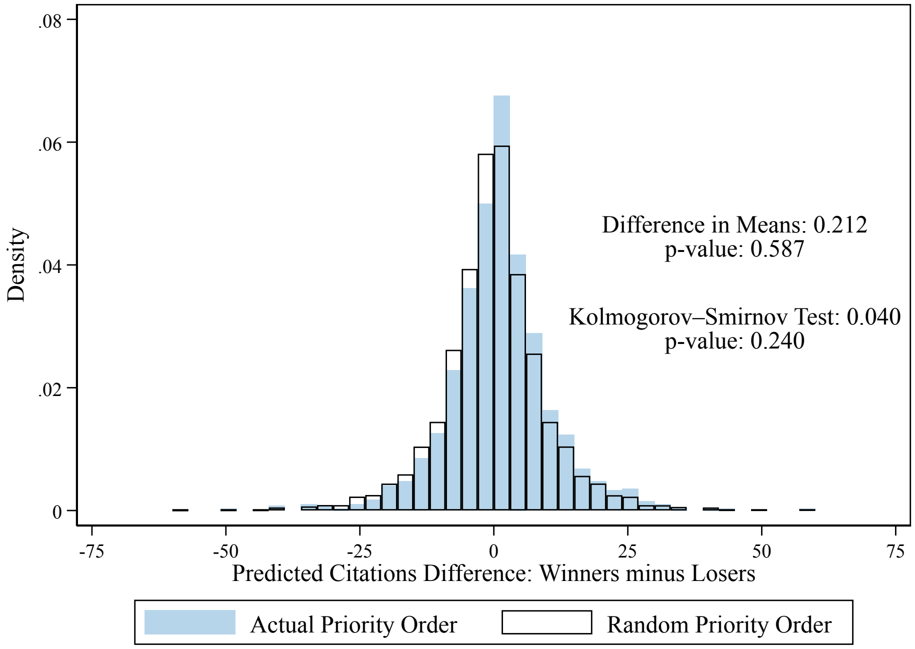


4JWS

3W9C

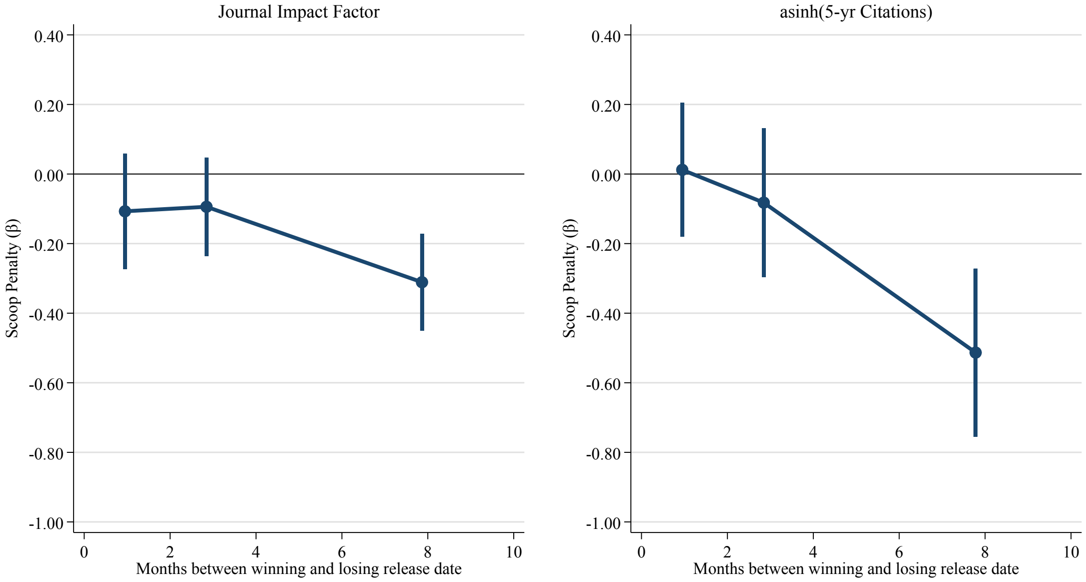
Notes: This figure presents a side-by-side comparison of the biological assembly models of the Pdx–P450cam complex protein deposited by two independent racing teams. According to the scoop definition in Section 2.2.4, structure deposit 4JWS scooped structure deposit 3W9C. See Table 2.1 for more details.

Figure 2.4: Histogram of Team Reputation Difference



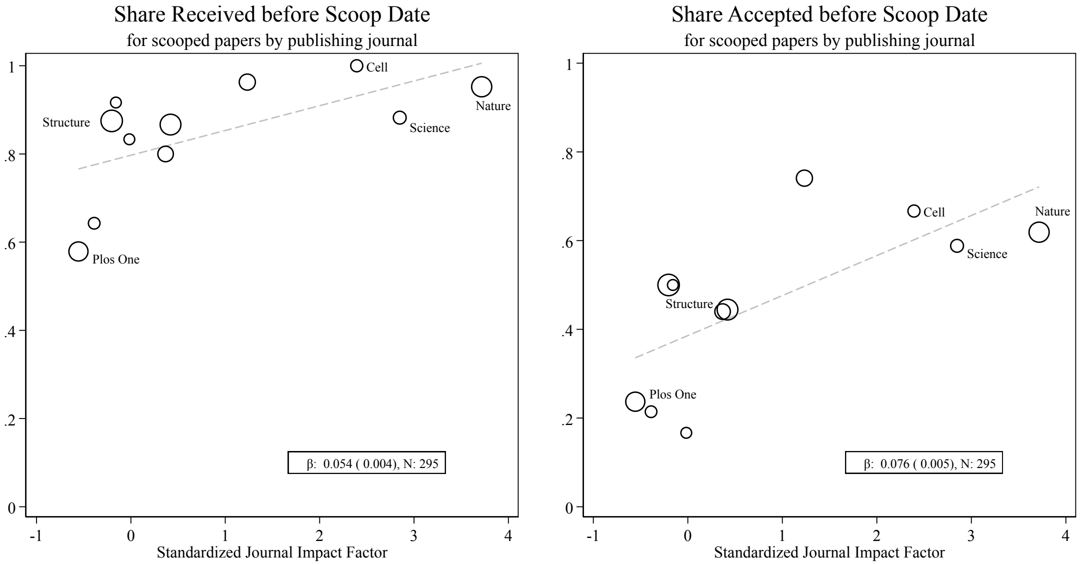
Notes: An observation in this figure is a racing pair. The blue distribution shows the actual difference in predicted citations. Bars the the right of zero represent instances when the winning team had higher predicted citations than the losing team, and bars to the left of zero represent instances when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

Figure 2.5: JIF and Citation Penalty by Scooped Project Release Delay



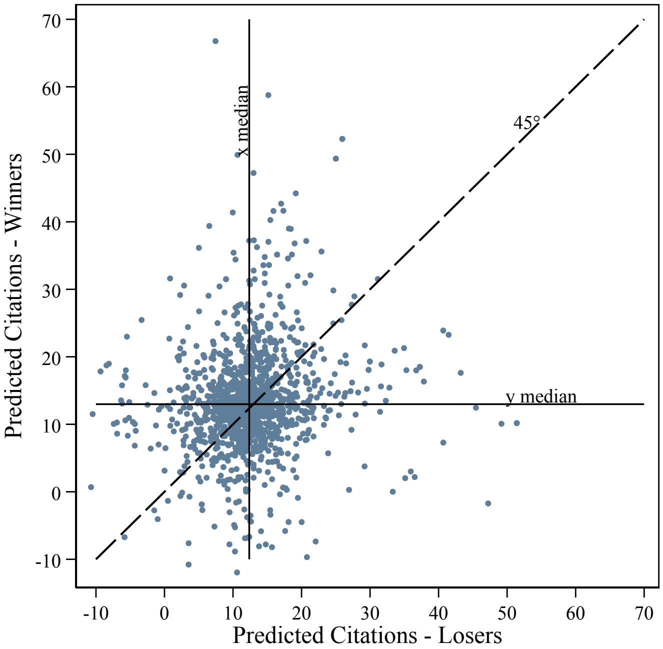
Notes: The sample of races is divided into three terciles along the distribution of time between winning and losing release date. Races are positioned along the x-axis at the average scoop release delay within each group. Projects released in close proximity are to the left, and those with a long delay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

Figure 2.6: Journal Placement and Timing of Scoops



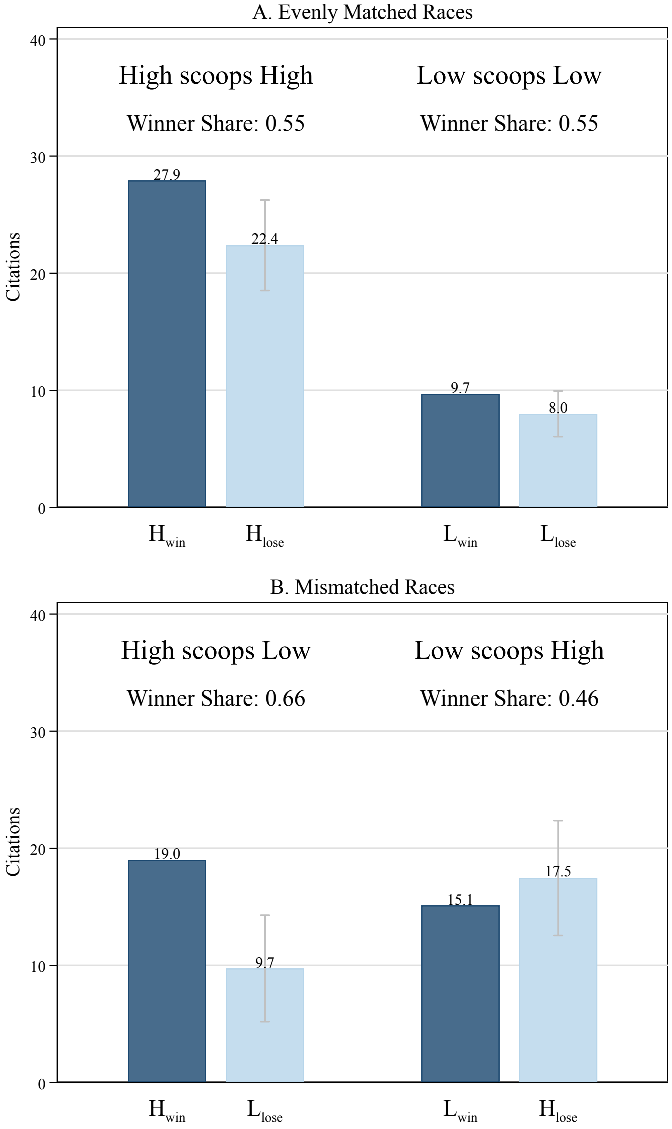
Notes: The figure reports the share of scooped papers that were received and accepted before the scoop date at different journals. Each circle represents one of the eleven largest journals that we collected supplemental data on the editorial timeline. Journals are arranged along the x-axis by their standardized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

Figure 2.7: Scatter Plot of Team Reputation Difference



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

Figure 2.8: Priority Effect by Reputation Match-up



Notes: We divide the sample of races from Figure 2.7 into four quadrants, depending on whether the winners and losers are above- or below-median in expected 3-year citations defined by the Lasso estimation. In each panel, the dark bars represent the actual citations of the winning team and the light bars of the losing team. Panel A reports the comparison between evenly matched races, H scoops H or L scoops L. Panel B reports the comparison between mismatched races, H scoops L or L scoops H. The winner's share of total citations are reported above each set of bars.

Table 2.1: 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 2.2: Summary Statistics for Structure-Level Data

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

Notes: This table presents summary statistics for the racing and non-racing samples. Observations are at the structure level. Column 1 shows the means of the racing sample and column 2 shows the means of the non-racing sample. Column 3 shows the difference between the racing and non-racing projects, and column 4 shows the heteroskedasticity-robust standard error of the difference.
**p <0.1, **p <0.05, ***p <0.01.*

Table 2.3: Covariate Balance Between Winning and Losing Teams

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

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

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

Table 2.4: Effect of Getting Scooped on Project Outcomes

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

Notes: This table presents regression estimates of the scoop penalty, following equation 1 in the text. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. Column 4 regression uses asinh(five-year citations) as the dependent variable, but Winner Y Mean is reported in levels for ease of interpretation.
**p<0.1, **p<0.05, ***p<0.01.*

Table 2.5: Effect of Getting Scooped on Alternative Measures of Attention

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

Notes: Attention outcomes are sourced from Altmetric.com. Sample restricted to years 2011-2017. Each regression contains protein (i.e. race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. All outcomes are cumulative counts of the metrics summed over time between the publication date to August 2019. All counts are transformed with the inverse hyperbolic sine transformation. The Altmetric Attention Score is a composite measure of all metrics used by Altmetric.com.

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

Table 2.6: Effect of Getting Scooped on Five-Year Productivity

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

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

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

Table 2.7: Decomposing Citation and Journal Effect

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

Notes: This table reports the scooped coefficients in regressions with five-year citations as the outcome where we control for journal impact factor. The citation counts are transformed with the inverse hyperbolic sine function in the regression, but the winner Y mean is reported in levels for ease of interpretation. The regression sample is restricted to races where both papers were published in a ranked publication. Column 1 re-estimates the Table 1, column 4 regression in this subsample. Column 2 and 3 add linear and then cubic controls for journal impact factor. Column 4 includes fixed effects for journal. All regressions also include PDS-Lasso selected controls and protein fixed effects.

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

Table 2.8: 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	0.268	0.264	0.004
						(0.016)
Prob (Publication)	0.853	0.976	0.665	0.628	0.703	-0.075***
						(0.022)
Journal impact factor penalty	-0.18	-1.23	-2.92	-2.95	-2.89	-0.055
						(0.084)
Citation penalty	-0.197	-0.150	-0.594	-0.620	-0.568	-0.052**
						(0.024)
Scooped citation share	0.445	0.459	0.257	0.241	0.274	-0.033***
						(0.011)

Notes: This table reports the responses to a survey of 915 structural biologists. The survey asked respondents to estimate the probability and consequences of getting scooped on a hypothetical project. See Appendix C for full survey text. Estimates from the PDB main regressions are reported in column 1. Comparable subsample PDB estimates in column 2 restrict to PDB races where one racer published in *Science, Nature,* or *Cell,* and losing team was scooped early in the process (quarter of sample with the shortest time between loser deposit and winner release). In column 4 and 5, respondents were divided into two groups, high- and low-reputation using the predicted citations measure used for heterogeneity in Section 6 of the text. Column 6 reports the difference in response means between columns 4 and 5 and reports the heteroskedastic-robust standard error in parentheses.
**p<0.1, **p<0.05, ***p<0.01.*

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Are Patent Examiners Gender Neutral?[*]

3.1 Introduction

Roughly 12 percent of all inventors listed on US patents granted in 2016 were women (US Patent and Trademark Office 2019), whereas in that same year women made up over 25 percent of the total science and engineering workforce (National Science Foundation 2018). Even among science and engineering degree holders, women are less likely to patent than men (Hunt et al. 2012). Many different behavioral margins could be relevant in explaining these facts. Do women create fewer inventions than men, either because of differences in the distribution of men and women across technological areas, or because of differential productivity in inventing within technological areas? Conditional on creating a new invention, are women less likely than men to file a patent application? Conditional on filing a patent application, do patent examiners judge applications submitted by female inventors more harshly than they do applications submitted by their male counterparts? Many of these questions are difficult to answer, but a recent study by Jensen et al. (2018) documented one relevant fact: women are more likely to have their patent applications rejected than men, even conditional on the invention's technology class.

In this paper, we investigate the role of gender in the evaluation of patent applications. We

[*] Contact: choi.jane.j@gmail.com, cstein@mit.edu, hlwill@stanford.edu

ask whether the gender of a patent examiner affects the evaluation of patent applications submitted by female inventors relative to male inventors. While as best we are aware we are the first to investigate this question in the context of the patent system, in recent years several related papers have investigated similar questions in other contexts. In general, these past studies have tended to conclude that male and female evaluators either judge female and male candidates similarly, or that the female evaluators more stringently evaluate female candidates. For example, Broder (2003) documents evidence that male applicants' National Science Foundation (NSF) proposals in the field of economics are rated similarly by male and female reviewers, but that female applicants' proposals are lower rated by female reviewers than by male reviewers. Similarly, Bagues and Esteve-Volart (2010) analyze Spanish public examinations for positions in the Corps of the Spanish Judiciary, which involve evaluation by committees, and document evidence that a female candidate is less likely to be hired if her committee (randomly) has a greater share of female evaluators. In contrast, similar analyses of committees evaluating candidates for associate and full professorships in Italy and Spain documented evidence that female evaluators are not more or less favorable toward female candidates (Bagues et al. 2017). Relatedly, Sarsons (2019) finds evidence that while physicians become more pessimistic about a female surgeon's ability relative to a male surgeon's ability following a patient death, that shift in pessimism does not seem to depend on the gender of the referring physician. Consistent with this evidence, Card et al. (2020) document evidence that male referees for papers submitted for publication at economics research journals are not differentially biased against manuscripts submitted by female authors.

Our motivation to focus on patent examination as an empirical setting is the fact that the evaluation of scientific ideas through the patent examination process is economically important, and that data on the key aspects of the patent evaluation process are publicly available. In 2015, over 600,000 patent applications were filed with the USPTO (US Patent and Trademark Office 2016). The agency employed around 8,000 patent examiners in the same year (US Patent and Trademark Office 2017). Each patent examiner is responsible for determining whether a given patent application qualifies for patentability, in the sense of being patent-eligible, novel, non-obvious, useful, and the text of the patent application satisfying the disclosure requirement (Williams 2017). Because questions such as whether any given invention is “novel” are by construction subjective, patent examiners hold an enormous amount of discretion when evaluating the patentability of submitted applications.

We analyze data on the universe of published patent applications submitted to the US Patent & Trademark Office (USPTO) between 2000 and 2013. These published patent applications list inventors’ first and last names, and – by matching on the patent application numbers – we merge on patent examiners’ first and last names as reported in a second USPTO administrative dataset. We probabilistically assign gender to both inventors and examiners based on the previously developed methodology of Jensen et al. (2018), who combine data on the gender distributions of first names from the US Social Security Administration with two supplementary commercial databases (GenderAPI and genderize.io). This process allows us to assign gender to 67 percent of inventor names and 74 percent of examiner names. Descriptively, around 8 percent of the inventors with an identified gender are female, compared to around 24 percent of examiners. Because patents may have multiple inventors, our measure of inventor gender is a continuous variable, representing the share of inventors coded as female. As a robustness check, we also restrict to single-inventor applications, where we can code gender as an indicator variable.

Our empirical analysis leverages these data to assess whether the gender of a patent examiner affects the evaluation of patent applications submitted by female inventors relative to male inventors. We condition our comparison on a patent application’s year of application as well as various variables that proxy for the type of technology being reviewed (Art Unit, technology class, and technology subclass). Past qualitative (Cockburn et al. 2003; Lemley and Sampat 2010, 2012) and quantitative (Sampat and Williams 2019; Gaulé 2018; Feng and Jaravel 2020; Farre-Mensa et al. 2020) evidence has suggested that patent applications are quasi-randomly assigned to patent examiners, within Art Units and years, at least in some sub-samples of applications. We document evidence supporting the quasi-random assignment of patent applications to patent examiners by documenting empirically that – conditional on application year, Art Unit, technology class, and technology subclass – the patent applications reviewed by male and female examiners appear to be balanced based on application-level covariates fixed at the time the patent application is filed. The fact that patent assignment appears to be as good as random, and is therefore orthogonal to examiner gender, allows us to estimate how examiner gender affects patent application outcomes.

Our estimates are economically small and meaningfully precise: although female patent examiners are more stringent overall relative to male examiners, they are not differentially stringent on female inventors compared to male examiners. Our preferred estimate suggests that the probability of an initial allowance of a patent application – a decision completely in the hands of the patent examiner – is around 0.1 percentage point lower for female inventors when they are reviewed by female patent examiners. That estimate is statistically indistinguishable from zero, and is arguably economically small relative to the mean initial allowance rate of 12 percent.

3.2 Data

3.2.1 USPTO data on published patent applications

We analyze the census of patent applications – both accepted and rejected applications – published by the US Patent & Trademark Office (USPTO) and filed between November 29, 2000 and December 31, 2013. The start date of this time period is determined by the American Inventor's Protection Act (AIPA), which included a provision requiring the publication of patent applications regardless of acceptance or rejection, with some exceptions, and was effective for patent applications submitted on or after November 29, 2000. Lemley and Sampat (2008, 2010, 2012) and Jensen et al. (2018) employ a similar restriction, and the USPTO has also documented a significant decline in coverage of published patents in public-use USPTO datasets of patent applications submitted prior to that date (Graham et al. 2015). The published patent applications list inventors' first and last names, as well as patent application numbers and other variables, but do not list patent examiner names.

For each published patent application, we merge in data on the name of the patent examiner who reviewed the application from the USPTO PAIR (Patent Application Information Retrieval) database. The USPTO PAIR data also records a patent application's Art Unit, technological class, and technological subclass – three different variables which each provide some information on the technological area to which the invention is relevant.

Because "continuations" of previously filed patent applications (referred to as "children" of the previously filed "parent" applications) are known to be assigned non-randomly to patent examiners – specifically, these child applications are automatically assigned to the same patent examiner who reviewed the parent patent application – we exclude child applications from our analysis. This is a substantial restriction: roughly 50 percent of patent applications in our data are child applications.

3.2.2 Probabilistic name-based assignment of inventor and examiner gender

We probabilistically assign gender to both inventors and examiners based on the previously developed methodology of Jensen et al. (2018). Jensen et al. (2018) determine the probability of an inventor or examiner being female by using the gender distributions of first names provided by the US Social Security Administration and two supplementary commercial databases: GenderAPI and genderize.io. The two commercial databases record the gender of users on social media and other sites to calculate the gender frequency of first names. The probability of a name being a female name is based on how frequently a name is associated with a male or female.

We use a 95 percent probability cutoff to assign gender. That is, we code a given inventor or examiner as female if 95 percent or more individuals with the same first name are female. For example, some names with high probabilities of being female are Crystal, Linda, and Pamela. Likewise, if 95 percent or more individuals with the same first name are male, we code the inventor or examiner as male. Some names with high probabilities of being male are Jonathan, Robert, and Stephen. We exclude inventors and examiners whose first names lie within 5 percent and 95 percent of being female, such as Akira (14 percent female), Robin (59 percent female), and Dominique (69 percent female).

Using this 95 percent cutoff rule as in Jensen et al. (2018), we are able to assign a gender to 67 percent of inventors' names and 74 percent of examiners' names in our sample, as documented in

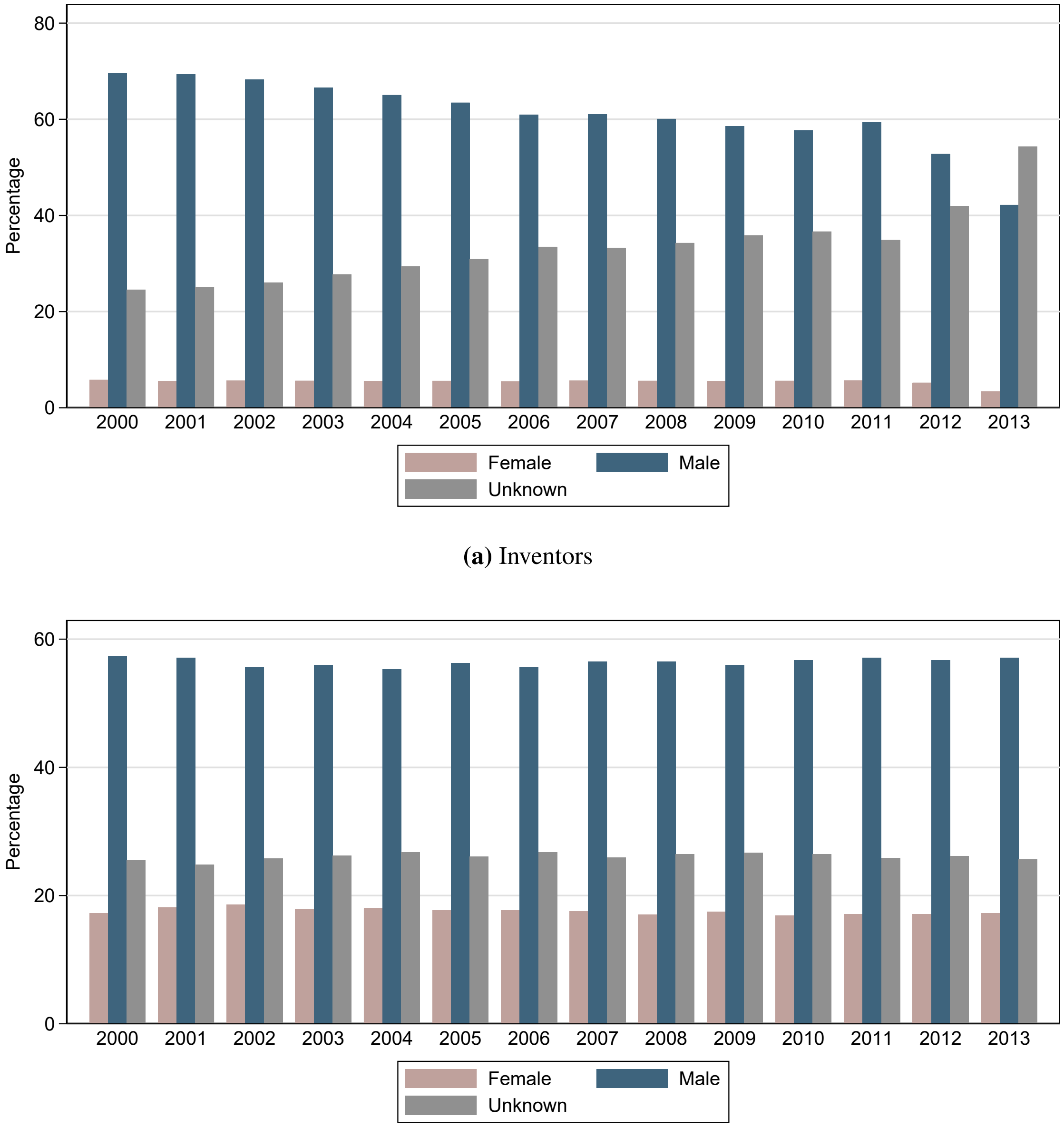
Table 3.1. As one point of comparison, 94.1 percent of all names covered in the Social Security Administration data are associated with only one gender at least 95 percent of the time. Of the inventor names that are assigned a gender, 8.3 percent are female. In their sample Jensen et al. (2018) estimate that 8.8 percent of inventor names are female.

Because a patent application often lists more than one inventor, for each patent application we calculate the percentage of listed inventors who are almost certainly female (i.e., have a 95 percent or greater probability of having a female name). In our sample, there are a total of 4,322,418 unique application-inventor combinations and 1,668,259 patent applications. As an example, consider an application with eight inventors, with one classified as female, four classified as male, and three classified as unknown. This application will have a female percentage of $\frac{1}{8} = 12.5 percent$.

Jensen et al. (2018) perform a similar calculation to determine the percentage of an application's inventors that are female, but they drop any application that has an inventor with an unknown gender (i.e., with a less than 95 percent probability of being either male or female). This restriction disproportionately excludes more applications with a large number of inventors, because the probability of assigning a gender to *all* inventors' names decreases as the number of inventors increases. We opt to instead include all published patent applications and calculate the percentage of inventors whose names are very likely female, with male and unknown inventors as the other category. This calculation will result in the same percentage calculated by Jensen et al. (2018) for the subset of applications that have all inventors' names assigned a gender, while also allowing us to include all applications. As a robustness exercise, we also separately analyze applications with only a female or only a male inventor listed.

Figures and Tables

Figure 3.1: Probabilistic Name-Based Assignment of Inventor and Examiner Gender, by Year



(a) Inventors

(b) Examiners

Notes: This figure documents the distribution of our probabilistic name-based assignment of inventor and examiner gender, separately by the year in which the patent application is filed. Gender assignment is based on the frequency of a name being female or male in US Social Security Administration data, GenderAPI, and genderize.io as in Jensen et al. (2018). Names that are female 95% or more of the time are assigned as female, and names that are male 95% or more of the time are assigned as male; all others are assigned to the "unknown" category.

Table 3.1: Probabilistic Name-Based Assignment of Inventor and Examiner Gender

	Inventors		Examiners	
	Count	Percent	Count	Percent
Male	2,651,310	61.3%	2,433,576	56.3%
Female	239,238	5.5%	758,468	17.5%
Unknown	1,431,870	33.1%	1,130,374	26.2%
Total	4,322,418	100.0%	4,322,418	100.0%

Notes: This table documents the distribution of our probabilistic name-based assignment of inventor and examiner gender. Gender assignment is based on the frequency of a name being female or male in US Social Security Administration data, GenderAPI, and genderize.io as in Jensen et al. (2018). Names that are female 95% or more of the time are assigned as female, and names that are male 95% or more of the time are assigned as male; all others are assigned to the "unknown" category.

Table 3.2: Distribution of Inventors per Patent Application

Number of Inventors	Number of Applications	Percent of Applications
1	580,833	34.8%
2	434,463	26.0%
3	264,797	15.9%
4	182,052	10.9%
5	81,157	4.9%
6	58,446	3.5%
7	20,355	1.2%
8	21,573	1.3%
9	5,364	0.3%
10	9,021	0.5%
11+	10,198	0.6%
Total	1,668,259	100.0%

(a) Number of Inventors

Percent of Female Inventors	Number of Applications	Percent of Applications
0%	1,475,527	88.4%
1-25%	70,278	4.2%
26-49%	45,114	2.7%
50%	42,434	2.5%
51-74%	5,655	0.3%
75-99%	177	0.0%
100%	29,033	1.7%
Total	1,668,218	100.0%

(b) Percentage of Female Inventors

Notes: Panel (a) documents the distribution of number of inventors per patent application. Panel (b) documents the distribution of percentage of inventors identified as female per application. The percentage of female inventors on an application is calculated by dividing the number of inventor names identified as female by the total number of inventor names identified as either male or unknown.

Table 3.3: Classes with Highest Percentage of Female Inventors and Female Examiners

USPC Class	Class Title	Average Percentage of Female Inventors	Percentage Examined by Female Examiners
450	Foundation garments	36.0%	97.9%
150	Purses, wallets, and protective covers	27.2%	50.3%
054	Harness for working animal	23.0%	19.1%
132	Toilet	21.5%	87.4%
289	Knots and knot tying	19.4%	0.0%
063	Jewelry	19.2%	31.7%
281	Books, strips, and leaves	19.1%	25.5%
002	Apparel	14.9%	57.4%
168	Farriery	12.8%	17.0%
119	Animal husbandry	12.3%	49.5%

(a) Classes with Highest Percentage of Female Inventors

USPC Class	Class Title	Percentage Examined by Female Examiners	Average Percentage of Female Inventors
999	Miscellaneous	98.2%	1.3%
450	Foundation garments	97.9%	36.0%
028	Textiles: manufacturing	89.8%	3.8%
132	Toilet	87.4%	21.5%
159	Concentrating evaporators	87.0%	0.0%
026	Textiles: cloth finishing	84.8%	4.3%
201	Distillation: processes, thermolytic	81.5%	0.0%
203	Distillation: processes, separatory	80.4%	1.7%
202	Distillation: apparatus	79.6%	0.6%
534	Organic compounds: part of the class 532-570 series	78.4%	1.8%

(b) Classes with Highest Percentage of Female Examiners

Notes: This table shows the USPC classes with the highest average percentage of female inventors per application and the USPC classes with the highest percentage of applications assigned to female examiners. The percentage of female inventors on an application is calculated by dividing the number of inventor names identified as female by the total number of inventor names identified as either male or unknown. Only classes with at least five applications in our sample period 2000-2013 are included. Organic compounds that are part of the class 532-570 series are miscellaneous organic carbon compounds (an example is Patent 8,933,208, for "Photo-responsive liquid crystalline compound and its applications."

Table 3.4: Balance tests: Applications Reviewed by Male vs. Female Examiners

Variable	(1) Male Examiners	(2) Female Examiners	(3) Difference	(4) Observations
Number of Claims	19.068	19.010	-0.058 (0.044) [0.188]	903,452
Number of Countries Filed	2.583	2.576	-0.006 (0.006) [0.280]	903,452
Fraction Women Inventors	0.051	0.051	0.001 (0.001) [0.189]	903,452
Clustered SEs?	N	N	N	N

Notes: This table tests for balance between applications assigned to male and female examiners by regressing the listed covariate on an indicator variable for female examiner with year-Art Unit-class-subclass fixed effects. Column (1) documents the mean for male examiners. Column (2) documents the mean for male examiners plus the regression estimate of the difference. Column (3) documents the regression estimate of the difference. Robust standard errors are in parentheses, p-values are in brackets.

Table 3.5: Examiner Gender and Patent Allowance decisions

	(1)	(2)	(3)	(4)	(5)
			Initial Allowance or Allowance After Rejection Within		
Dependent variable:	Initial Allowance	Any Allowance	90 Days	180 Days	365 Days
% Female inventor	-0.0036	-0.0422***	-0.0023	-0.0022	-0.0156***
robust se	(0.0025)	(0.0039)	(0.0026)	(0.0035)	(0.0038)
clustered se	(0.0027)	(0.0043)	(0.0029)	(0.0037)	(0.0041)
Female examiner	-0.0144***	-0.0113***	-0.0146***	-0.0257***	-0.0226***
robust se	(0.0012)	(0.0018)	(0.0012)	(0.0016)	(0.0018)
clustered se	(0.0036)	(0.0044)	(0.0038)	(0.0058)	(0.0056)
% Female inventor x female examiner	-0.0010	-0.0132**	-0.0014	-0.0083	-0.0075
robust se	(0.0036)	(0.0059)	(0.0037)	(0.0050)	(0.0056)
clustered se	(0.0042)	(0.0065)	(0.0044)	(0.0057)	(0.0063)
Year-Art Unit-Class-Subclass FE	Y	Y	Y	Y	Y
Observations	672,047	693,072	693,072	693,072	693,072
Adjusted R^2	0.132	0.233	0.138	0.226	0.274
Mean of dep. var. (male examiner, male inventors)	0.1203	0.6436	0.1355	0.3079	0.4702

Notes: ***p<0.001, **p<0.05, *p<0.1. This table reports the results of regressing various patent application allowance outcomes on the percentage of an application's inventors who are female, an indicator variable for female examiner, and the interaction between these two terms. Application year-Art Unit-class-subclass fixed effects are included. We show both robust standard errors and standard errors clustered on examiner. Statistical significance is denoted based on robust standard errors.

Table 3.6: Examiner Gender and Patent Allowance Decisions, for Patent Applications with a Single Inventor

	(1)	(2)	(3)	(4)	(5)
			Initial Allowance or Allowance After Rejection Within		
Dependent variable:	Initial Allowance	Any Allowance	90 Days	180 Days	365 Days
Female inventor	-0.0010	-0.0447***	0.0003	-0.0018	-0.0142**
robust se	(0.0044)	(0.0070)	(0.0046)	(0.0060)	(0.0066)
clustered se	(0.0048)	(0.0074)	(0.0049)	(0.0064)	(0.0068)
Female examiner	-0.0112***	-0.0086**	-0.0120***	-0.0188***	-0.0135***
robust se	(0.0026)	(0.0038)	(0.0027)	(0.0035)	(0.0038)
clustered se	(0.0046)	(0.0058)	(0.0048)	(0.0069)	(0.0070)
Female inventor x female examiner	0.0002	-0.0162	-0.0017	-0.0097	-0.0131
robust se	(0.0079)	(0.0125)	(0.0081)	(0.0105)	(0.0117)
clustered se	(0.0086)	(0.0138)	(0.0086)	(0.0112)	(0.0127)
Year-Art Unit-Class-Subclass FE	Y	Y	Y	Y	Y
Observations	153,138	156,910	156,910	156,910	156,910
Adjusted R^2	0.144	0.228	0.152	0.239	0.277
Mean of dep. var. (male examiner, male inventor)	0.1147	0.6403	0.1317	0.2982	0.4638

*Notes: ***p<0.001, **p<0.05, *p<0.1.* This table reports the results of regressing various patent application allowance outcomes on an indicator variable for female inventor, an indicator variable for female examiner, and the interaction between these two terms. Only applications with a single inventor are included. Application year-Art Unit-class-subclass fixed effects are included. We show both robust standard errors and standard errors clustered on examiner. Statistical significance is denoted based on robust standard errors.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Appendix for Chapter 1

A.1 Theory Appendix

A.1.1 Proofs of Propositions

Proof of Proposition 1.

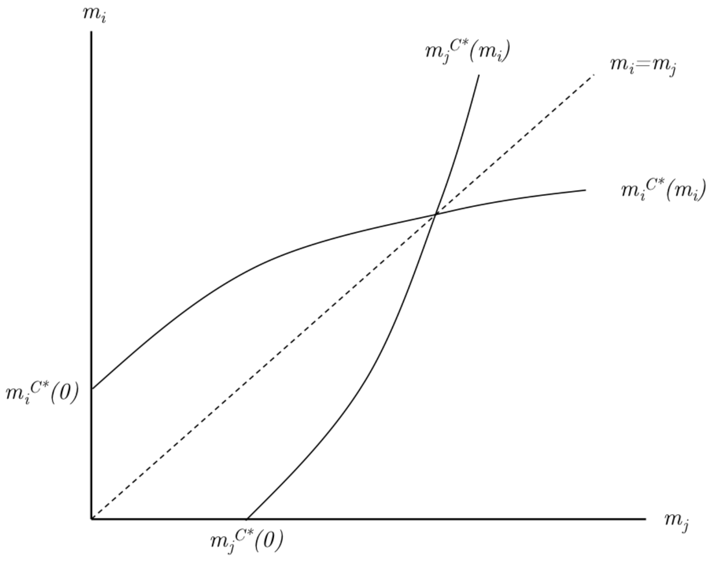
First, we will expand on how we derive the first-order condition for $m_i^{C^*}$ (Equation 1.7). Taking the derivative of Equation 1.6 with respect to m_i and setting it equal to zero yields:

$$\frac{Q'(m_i^{C^*})}{Q(m_i^{C^*})} = r - \frac{\frac{\partial \pi}{dm_i}(\overline{\theta} - \underline{\theta})}{\pi(m_i, m_j)\overline{\theta} + (1 - \pi(m_i, m_j))\underline{\theta}}. \tag{A.1}$$

Next, we note that $\pi(m_i, m_j) = (1-g) + g(\frac{1}{2} + \frac{m_j - m_i}{2\Delta})$ and therefore $\frac{\partial \pi}{\partial m_i} = -\frac{g}{2\Delta}$ *if m_i is close enough to m_j*. We will assume this is the case for the moment, and plugging these values into Equation A.1 above yields Equation 1.7 in the text. However, if m_i is much larger than m_j (i.e., if $m_i > m_j + \Delta$), then $\frac{\partial \pi}{\partial m_i} = 0$ and Equation A.1 collapses to the no-competition case, i.e., Equation 1.4. We will return to this caveat, but for now we will assume m_i is close to m_j.

Equation 1.7 implicitly defines $m_i^{C^*}(m_j)$ as a function of m_j and parameters. If we can show that (i) $m_i^{C^*}(0) > 0$ and (ii) $\frac{dm_i^{C^*}}{dm_j} \in (0,1)$, then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $m_i^{C^*}(m_j)$ and $m_j^{C^*}(m_i)$ will only cross the $m_i = m_j$ line once.

Figure A1: Maturation Best Response Functions



To show (i), plug $m_j = 0$ into Equation 1.7. This results in an equation that implicitly defines a unique $m_i^{C^*}(0) > 0$. To show (ii), we can totally differentiate equation 1.7 with respect to m_j. For notational ease, define $\zeta \equiv \Delta\left(\frac{2\overline{\theta}-g(\overline{\theta}-\underline{\theta})}{g(\overline{\theta}-\underline{\theta})}\right)$, and note that $\zeta > 0$. Gathering terms and rearranging, we have that

$$\frac{dm_i^{C^*}}{dm_j} = \left[\underbrace{\left(\frac{-Q(m_i^{C^*})Q''(m_i^{C^*}) + Q'(m_i^{C^*})^2}{Q(m_i^{C^*})^2}\right)\left(\zeta + m_j - m_i^{C^*}\right)^2}_{>0} + 1\right]^{-1} \in (0,1). \tag{A.2}$$

Next, we confirm that the second-order conditions hold. Differentiating the objective function

(Equation 1.6) twice with respect to m_i and evaluating at $m_i = m_j = m^{C^*}$ yields

$$Pe^{-rm_i}\left[Q''(m^{C^*}) - Q'(m^{C^*})\left(r + \frac{1}{\zeta}\right)\right] < 0. \tag{A.3}$$

Therefore, $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ is a local optimum. Plugging m^{C^*} in for both m_i and m_j (and assuming that $I_i = I_j = I^{C^*}$) in Equation 1.7 yields the expression in Proposition 1.

However, as a final check, we need to confirm that this is also a global optimum. Note that Equation 1.8 tells us that as $\Delta \to 0$, $m_i^{C^*} \to 0$. This will yield a payoff of zero for researcher i. This cannot be researcher i's best response, because there is always a $1-g$ probability that her competitor did not enter. Therefore, she would be better off selecting $m_i = m^{NC^*}$ and hoping that her competitor fails to enter the project. To map this intuition to the math, note that we are now considering a case where $m_i > m_j + \Delta$, and so we the relevant first-order condition is now Equation 1.4.

More generally, in order to ensure that $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ is a global optimum we need the payoff from playing $m_i = m^{C^*}$ to be larger than the payoff to playing $m_i = m^{NC^*}$:

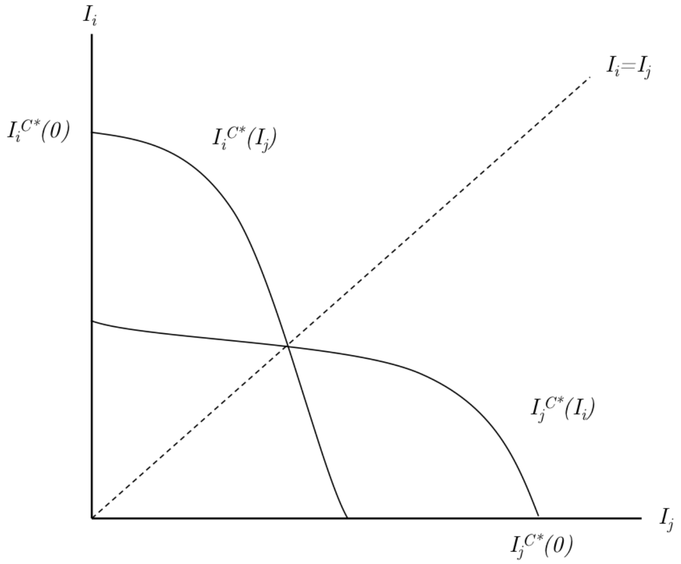
$$e^{-rm^{C^*}}PQ(m^{C^*})\left[(1-\frac{g}{2})\overline{\theta} + \frac{g}{2}\underline{\theta}\right] > e^{-rm_i^{NC}}PQ(m_i^{NC^*})\left((1-g)\,\overline{\theta} + g\underline{\theta}\right). \tag{A.4}$$

Because m^{C^*} is increasing in Δ, this defines a lower bound on Δ such that this equation will hold. Therefore, $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ is a symmetric pure strategy Nash equilibrium as long as Δ is sufficiently large. Moreover, this is the only possible pure strategy Nash equilibrium. To see this, note that if $|m_i - m_j| < \Delta$, then the first-order condition in Equation 1.7 applies and we have the equilibrium defined by $m_i^{C^*} = m_j^{C^*} = m^{C^*}$. Alternatively, if $|m_i - m_j| \geq \Delta$, then the first-order condition defined by Equation 1.4 applies. But this implies that $m_i^* = m_j^* = m^{NC^*}$, which violates the assumption that $|m_i - m_j| \geq \Delta$. Therefore, if Δ is below some threshold, the Nash equilibrium must be mixed. We will focus on the pure strategy case throughout the remainder of the paper.

Proof of Proposition 2.

Equation 1.10 implicitly defines $I_i^{C^*}(I_j)$ as a function of I_j, $m_i^{C^*}$ (which depends on I_j), and parameters. If we can show that (i) $I_i^{C^*}(0) > 0$ and (ii) $\frac{dI_i^{C^*}}{dI_j} < 0$ then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $I_i^{C^*}(I_j)$ and $I_j^{C^*}(I_i)$ will only cross the $I_i = I_j$ line once.

Figure A2: Investment Best Response Functions



To show (i), imagine that j invests zero. Then i should surely invest some positive amount, because the marginal return will be be proportional to $g'(I_i)$. Due to the Inada conditions assumption on $g(\cdot)$, $g'(I_i)$ will be quite large for small values of I_i. To show (ii), we can totally differentiate Equation 1.10 with respect to I_j. Gathering terms and rearranging, we have that

$$\frac{dI_i^{C^*}}{dI_j} = \frac{e^{-rm_i^{C^*}} P \left[\left(rQ(m_i^{C^*}) - Q'(m_i^{C^*}) \right) \frac{dm_i^{C^*}}{dI_j} + Q(m_i^{C^*}) g'(I_j)(\overline{\theta} - \underline{\theta}) \right]}{g''(I_j) \left[e^{-rm_i^{C^*}} PQ(m_i^{C^*}) \left(\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta}) \right) \right]^2} < 0 \qquad \text{(A.5)}$$

where we can sign this expression by noting that $rQ(m_i^{C^*}) - Q'(m_i^{C^*}) < 0$ (due to Equation 1) and $\frac{dm_i^{C^*}}{dI_j} < 0$ and applying assumptions about the function $g(I)$. Therefore, $I_i^{C^*} = I_j^{C^*} = I^{C^*}$ is a unique, pure strategy Nash equilibrium. Plugging in I^{C^*} for both I_i and I_j, and plugging in m^{C^*} for m_i and m_j yields the expression in Proposition 2. This also confirms our assumption that $I_i = I_j = I^{C^*}$ in Proposition 1.

Proof of Proposition 3.

Looking at Equation 1.7, the left hand side is decreasing in m^{C^*}. Looking at the right hand side, we see it is increasing in $g(I^{C^*})$. For the equality to hold as $g(I^{C^*})$ increases, it must be the case that m^{C^*} decreases, i.e., that $\frac{dm^{C^*}}{dg(I^{C^*})} < 0$. Because $Q(m)$ is increasing, this also implies that $\frac{dQ(m^{C^*})}{dg(I^{C^*})} < 0$.

Proof of Proposition 4.

Suppose this were not the case. In particular, consider two projects with P_1 and P_2, and further suppose that $P_1 > P_2$. If Proposition 4 is not true, investment for project 1 would be lower than for project 2, i.e., $I^{C^*,1} \leq I^{C^*,2}$. From Proposition 3, we then know that then $m^{C^*,1} > m^{C^*,2}$ and $Q(m^{C^*,1}) > Q(m^{C^*,2})$. The expected PDV of successfully entering an arbitrary project is given by

$$e^{-rm^{C^*}} PQ(m^{C^*}) \left[\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta}) \right]. \tag{A.6}$$

It is clear that this value is unambiguously higher for project 1 than for project 2. Therefore, a researcher would want to invest more to enter project 1 than project 2 (see Equation 2 to confirm this intuition). Therefore, we have a contradiction. This implies that $I^{C^*,1} > I^{C^*,2}$ for any arbitrary pair of projects where $P_1 > P_2$. This implies that $\frac{dg(I^{C^*})}{dP} > 0$.

Proof of Proposition 5.

See main text.

Proof of Lemma 1.

Let $\Delta Q = Q(m^{IMP^*}) - Q(m^{C^*})$ denote the realized quality improvement. The derivative of the present discounted value of a project improvement (Equation 1.16) with respect to project potential P is given by:

$$-re^{-rm^{IMP^*}} \frac{dm^{IMP^*}}{dP} P\Delta Q + e^{-rm^{IMP^*}} \Delta Q + e^{-rm^{IMP^*}} P \frac{d\Delta Q}{dP}. \tag{A.7}$$

The first term represents the change in discounting due to the effect of P on m^{IMP^*}, the second term represents the direct effect of shifting P, and the final term represents the change in the quality improvement, via the effect of P on m^{IMP^*} and m^{C^*}. Totally differentiating Equation 1.18 with respect to P and rearranging yields:

$$\frac{dm^{IMP^*}}{dP} = \frac{rQ'(m^{C^*})\frac{dm^{C^*}}{dP}}{rQ'(m^{IMP^*}) - Q''(m^{IMP^*})} < 0 \tag{A.8}$$

where we can sign the expression by noting that $\frac{dm^{C^*}}{dP}$ is negative, as shown in Proposition 5. Next, we can re-write Equation 1.18 as

$$\Delta Q = \frac{Q'(m^{IMP^*})}{r}.$$

Taking the derivative of this equation with respect to P yields

$$\frac{d\Delta Q}{dP} = \frac{Q''(m^{IMP^*})}{r} \cdot \frac{dm^{IMP^*}}{dP} > 0 \tag{A.9}$$

due to the concavity of $Q(\cdot)$. Together, these two derivatives allow us to unambiguously show that the expression in Equation A.7 is positive.

Proof of Proposition 6.

See main text.

Proof of Proposition 7.

Taking the derivative of Equation 1.21 with respect to P yields

$$\frac{d\overline{Q}_{max}}{dP} = \frac{dQ(m^{C^*})}{dP} + g'(I^{IMP^*})\frac{dI^{IMP^*}}{dP}\Delta Q + g(I^{IMP^*})\frac{d\Delta Q}{dP}. \tag{A.10}$$

Because we have already shown that $\frac{dI^{IMP^*}}{dP} > 0$ (Proposition 6) and $\frac{d\Delta Q}{dP} > 0$ (see the proof of Lemma 1), we know that $\frac{d\overline{Q}_{max}}{dP} > \frac{dQ(m^{C^*})}{dP}$.

Proof of Lemma 2.

Plugging $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ into Equation 1.7, we recover Equation 1.4, which defines both the no-competition maturation period and the social planner's optimal maturation period. Plugging $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ and $m = m^{SP^*}$ into Equation 1.10, we have

$$g'(I^{C^*}) = \frac{1}{e^{-rm^{SP^*}}PQ(m^{SP^*})(V/2)}.$$

Comparing this to Equation 1.25, we see that as long as k is sufficiently large (in this case, as long as $k > \frac{V/2}{(1-g(I^{SP^*}))}$), then $I^{SP^*} > I^{C^*}$.

Proof of Proposition 8.

We start by writing out $\frac{dm^{C^*}}{d\overline{\theta}}$ and $\frac{dI^{C^*}}{d\overline{\theta}}$ using the chain rule. We then apply the implicit function theorem to Equations 1 and 2 (after substituting $\underline{\theta} = V - \overline{\theta}$ in both equations) to sign all the partial derivatives. This leaves us with the following:

$$\frac{dm^{C^*}}{d\overline{\theta}} = \underbrace{\frac{\partial m^{C^*}}{\partial \overline{\theta}}}_{<0} + \underbrace{\frac{\partial m^{C^*}}{\partial I^{C^*}}}_{\leq 0} \cdot \frac{dI^{C^*}}{d\overline{\theta}}$$

and

$$\frac{dI^{C^*}}{d\overline{\theta}} = \underbrace{\frac{\partial I^{C^*}}{\partial \overline{\theta}}}_{>0} + \underbrace{\frac{\partial I^{C^*}}{\partial m^{C^*}}}_{\geq 0} \cdot \frac{dm^{C^*}}{d\overline{\theta}}.$$

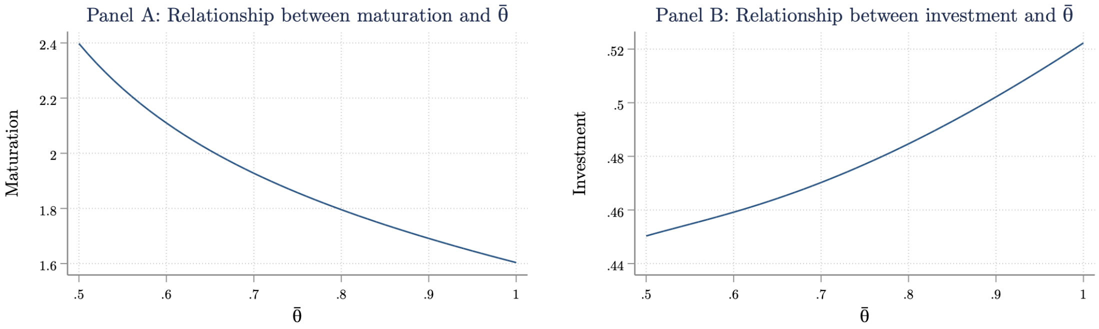
We can immediately note that $\frac{dm^{C^*}}{d\overline{\theta}} < 0$ (to see this, assume $\frac{dm^{C^*}}{d\overline{\theta}} \geq 0$ and arrive at a contradiction). The sign of $\frac{dI^{C^*}}{d\overline{\theta}}$ is ambiguous, and depends on whether the direct effect ($\frac{\partial I^{C^*}}{\partial \overline{\theta}}$) dominates or whether the indirect effect via m ($\frac{\partial I^{C^*}}{\partial m^{C^*}} \cdot \frac{dm^{C^*}}{d\overline{\theta}}$) dominates.

At this point, it is helpful to construct an example. Suppose we have the following parameter values and expressions for $Q(m)$ and $g(I)$:

- $r = 0.1$, $P = 4$, $\Delta = 2$, $k = 2$, $V = 1$
- $Q(m) = 1 - e^{-m}$
- $g(I) = 1 - e^{-1.2I}$

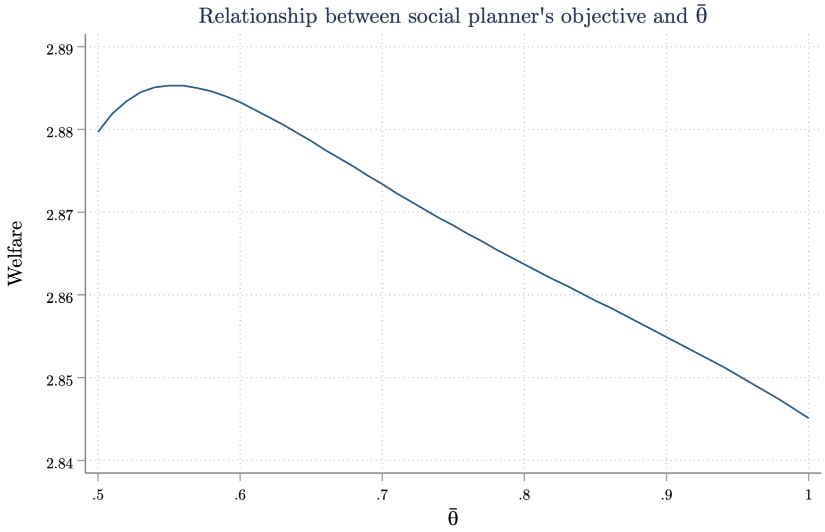
Then, we can numerically compute $\frac{dm^{C^*}}{d\overline{\theta}}$ and $\frac{dI^{C^*}}{d\overline{\theta}}$. We show these below. This results in $\frac{dm^{C^*}}{d\overline{\theta}} < 0$ and $\frac{dI^{C^*}}{d\overline{\theta}} > 0$.

Figure A3: Numerically calculated $\frac{dm^{C^*}}{d\overline{\theta}}$ and $\frac{dI^{C^*}}{d\overline{\theta}}$



In this particular example, this means that as we increase $\overline{\theta}$ from $\frac{V}{2} = \frac{1}{2}$ toward 1, m^{C^*} falls from the socially optimal value, but I^{C^*} increases toward the socially optimal value. In this example, this results in an optimal choice of $\overline{\theta}^*$ that is between $\frac{V}{2} = \frac{1}{2}$ and 1, as shown in the figure below.

Figure A4: Welfare as a function of $\overline{\theta}$



Proof of Proposition 9.

As long as $\overline{\theta} = \underline{\theta}$, then $m^{C^*} = m^{SP^*}$, as shown in the proof of Proposition 8. To achieve $I^{C^*} = I^{SP^*}$, we plug $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ and $m = m^{SP^*}$ into Equation 2, and equate this with Equation 1.25:

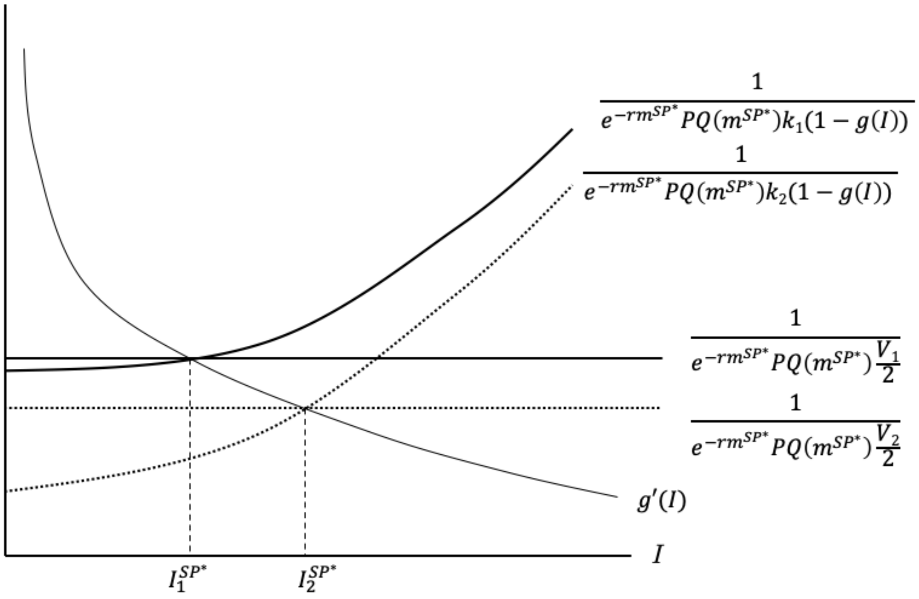
$$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V/2)} = \frac{1}{e^{-rm^{SP^*}} kPQ(m^{SP^*})(1 - g(I^{SP^*}))}.$$

Here, we treat V as a free variable. Re-arranging, we arrive at

$$V = 2k(1 - g(I^{SP^*})).$$

So we can recover the first best if $\overline{\theta} = \underline{\theta} = k(1-g(I^{SP^*}))$. Figure A5 below helps illustrate that $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ is increasing in k. Suppose $k = k_1$. To achieve $I^{C^*} = I^{SP^*}$, we need $\frac{1}{e^{-rm^{SP^*}} k_1 PQ(m^{SP^*})(1-g(I))}$ to intersect both $g'(I)$ and $\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V_1/2)}$, which occurs at $I = I_1^{SP}$ in Figure A5. However, if we increase k from k_1 to k_2, then $\frac{1}{e^{-rm^{SP^*}} k_2 PQ(m^{SP^*})(1-g(I))}$ shifts down (shown by a dotted line). To maintain this intersection, then $\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V_2/2)}$ must also shift down (again shown by a dotted line), which implies that $V_2 > V_1$.

Figure A5: Achieving Optimal Investment


$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*}) k_1 (1 - g(I))}$
$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*}) k_2 (1 - g(I))}$
$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*}) \frac{V_1}{2}}$
$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*}) \frac{V_2}{2}}$
$g'(I)$
I
$I_1^{SP^*}$
$I_2^{SP^*}$

A.2 Data Appendix

A.2.1 Description of the Protein Data Bank Data

The first iteration of the Protein Data Bank (PDB) started in 1971. Today, a non-profit organization called the World Wide Protein Data Bank (wwPDB) curates and manages the database. The wwPDB is a collaboration of four existing data banks from around the world: 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.[1]

We access the data directly from the RCSB Custom Report Web Service.[2] 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, experimental technique, classification, macromolecule type, molecular weight, residue count, and atom site count.
- Citation: PubMed ID, publication year, and journal name.
- Cluster Entity: entity ID, chain ID, UniPROT accession number, taxonomy, gene name, BLAST sequence 100 percent similarity clusters.
- Data Collection Details: collection date (the self-reported date the scientists generated diffraction data at a major synchrotron or in a home lab).
- Refinement Details: r-free and refinement resolution.

Data about Ramachandran outliers, one of the quality metrics, was not available through RCSB custom reports. Instead, we accessed validation reports data from the PDBe REST API[3] provided by the European Bioinformatics Institute (EMBL-EPI). Data for this study was downloaded on October 25, 2019 and merged using the standard PDB structure identifiers.

Many of the variables we use in the analysis, such as predicted citations, are calculated at the paper level. However 20 percent of PDB-linked papers have more than one structure, with

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

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

[3] `https://www.ebi.ac.uk/pdbe/api/doc/validation.html`

an average of 1.5 structures per paper. Because each linked structure has a unique set of quality metrics, it is difficult to ascribe paper-level characteristics to any one of the individual structures. Our main analysis sample therefore drops all structures linked to multi-structure papers. Since about 30% of deposits are never published, we make a similar restriction for groups of structure deposits that appear to have been part of the same unpublished project. We group unpublished structures into the same "project" if the deposits have the same first and last PDB structure author and share the same release date. Unpublished projects with more than one structure are dropped to mirror the single-structure paper restriction.

A further complication of the PDB data is that cluster groupings are defined at a level of granularity that is smaller than the structure or 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 particular, our main analysis sample includes only "priority" structure deposits, meaning that the PDB entry was the first to produce a structure for a given entity. In practice, we keep any structure that has at least one entity that is the first deposit among all other entities that are 100 percent similar according to the BLAST algorithm. This means that in some cases, only one part of the structure is truly a novel discovery, but these deposits still represent important contributions for which scientists often compete to publish first.

Some relevant protein characteristics are assigned at the entity, rather than the structure level. For example, we use gene-protein linkages as an input to the predicted citation LASSO model described in Section 1.4.1. The PDB data assigns gene linkages at the entity level, meaning some proteins (9.4 percent) have multiple gene linkages. To simplify the citation prediction model, we assign a single gene-linkage to the full protein by taking the modal gene name amongst the protein entities and breaking ties alphabetically. Similarly, some structures are complexes of entities from different organisms (e.g. a human protein bound to a virus), so we assign the modal taxonomy to the 5.9 percent of proteins with multiple taxonomies.

A.2.2 Description of the Web of Science Data

Citation data is sourced from the Web of Science produced by Clarivate Analytics and accessed through a license with Stanford University. Our version of the dataset includes digitized academic references through the end of 2018 and is linked to the PDB data using PubMed identifiers. The citation data is restricted to citations between papers linked to PubMed IDs,[4] and self-citations are excluded. Citations are aggregated for each cited paper by publication year of the citing paper. When we report three-year citations, it represents the total number of citations in the publishing year and the subsequent three calendar years.

A.2.3 Description of the UniPROT Knowledgebase Data

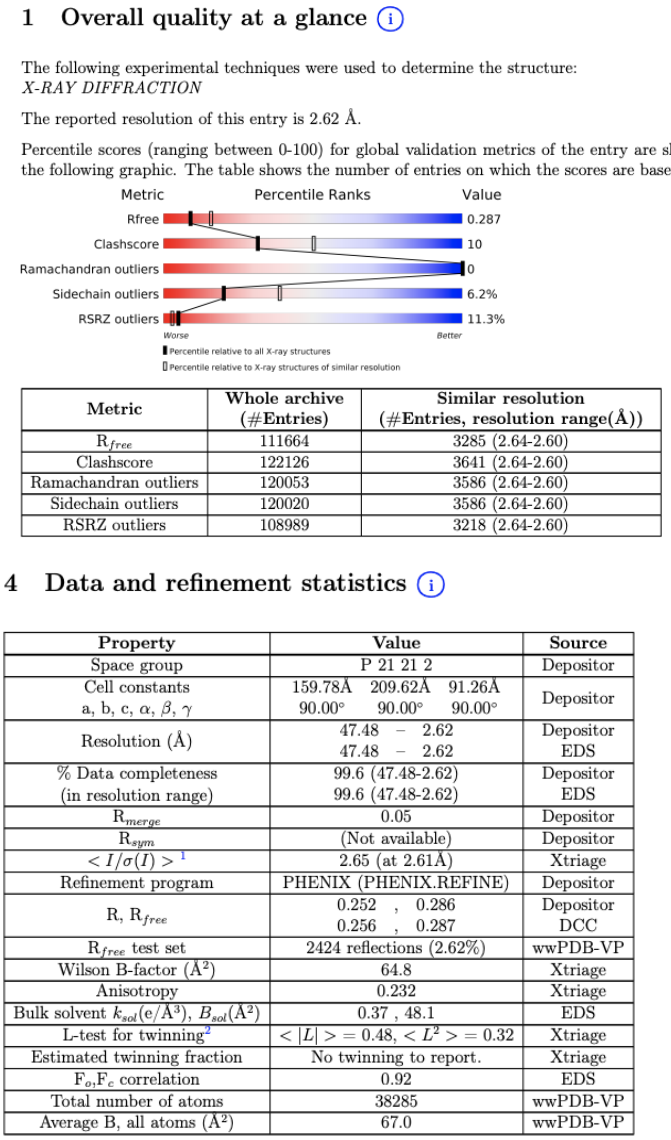
The UniPROT Knowledgebase is a comprehensive, curated database of the biological and functional details of most known proteins. Importantly for our purposes, each protein entry contains a linkage to PDB identifiers of associated structure discoveries. It also contains an annotated bibliography of all associated scientific articles, both structure papers and others, such as articles describing protein function. We count the number of PubMed-linked articles that were published before the first structure discovery as a measure of "potential" or ex-ante demand for a structure model. We only include papers that had been manually reviewed (Swiss-Prot) and exclude those that had only been annotated automatically (TrEMBL). Raw data was accessed on August 26, 2018.[5]

[4] 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.

[5] Downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.xml.gz

A.3 Appendix Figures and Tables

Figure A6: Validation Report for PDB ID 4CMP — Crystal Structure of S. pyogenes Cas9



1 Overall quality at a glance ⓘ

The following experimental techniques were used to determine the structure:
X-RAY DIFFRACTION

The reported resolution of this entry is 2.62 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

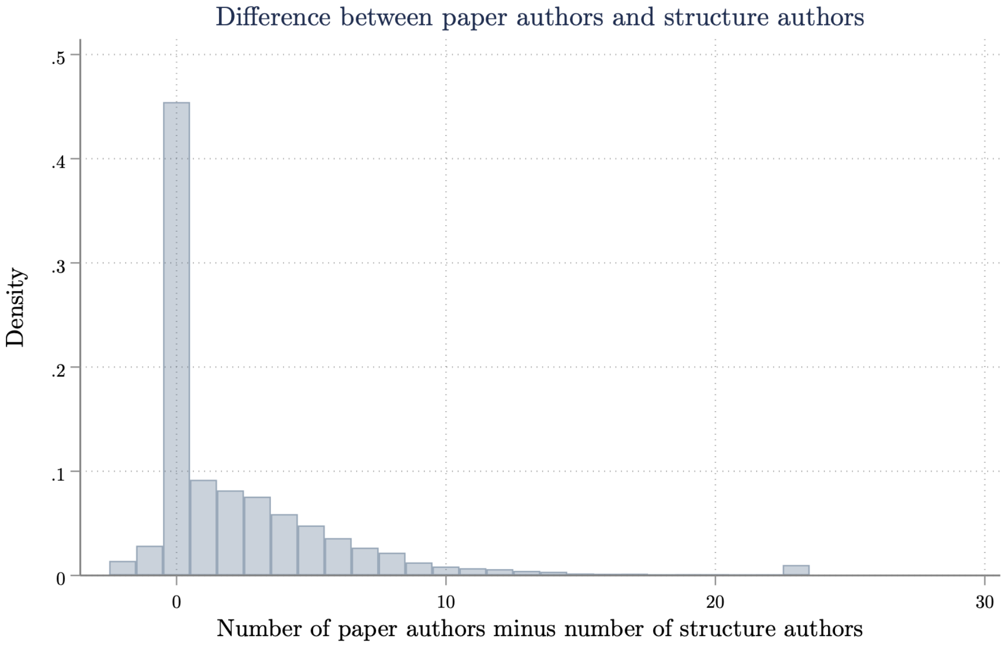
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	3285 (2.64-2.60)
Clashscore	122126	3641 (2.64-2.60)
Ramachandran outliers	120053	3586 (2.64-2.60)
Sidechain outliers	120020	3586 (2.64-2.60)
RSRZ outliers	108989	3218 (2.64-2.60)

4 Data and refinement statistics ⓘ

Property	Value	Source
Space group	P 21 21 2	Depositor
Cell constants a, b, c, α, β, γ	159.78Å 209.62Å 91.26Å 90.00° 90.00° 90.00°	Depositor
Resolution (Å)	47.48 – 2.62 47.48 – 2.62	Depositor EDS
% Data completeness (in resolution range)	99.6 (47.48-2.62) 99.6 (47.48-2.62)	Depositor EDS
R_{merge}	0.05	Depositor
R_{sym}	(Not available)	Depositor
$< I/\sigma(I) >$ [1]	2.65 (at 2.61Å)	Xtriage
Refinement program	PHENIX (PHENIX.REFINE)	Depositor
R, R_{free}	0.252 , 0.286 0.256 , 0.287	Depositor DCC
R_{free} test set	2424 reflections (2.62%)	wwPDB-VP
Wilson B-factor ($Å^2$)	64.8	Xtriage
Anisotropy	0.232	Xtriage
Bulk solvent k_{sol}(e/$Å^3$), B_{sol}($Å^2$)	0.37 , 48.1	EDS
L-test for twinning[2]	$< \lvert L \rvert > = 0.48$, $< L^2 > = 0.32$	Xtriage
Estimated twinning fraction	No twinning to report.	Xtriage
F_o,F_c correlation	0.92	EDS
Total number of atoms	38285	wwPDB-VP
Average B, all atoms ($Å^2$)	67.0	wwPDB-VP

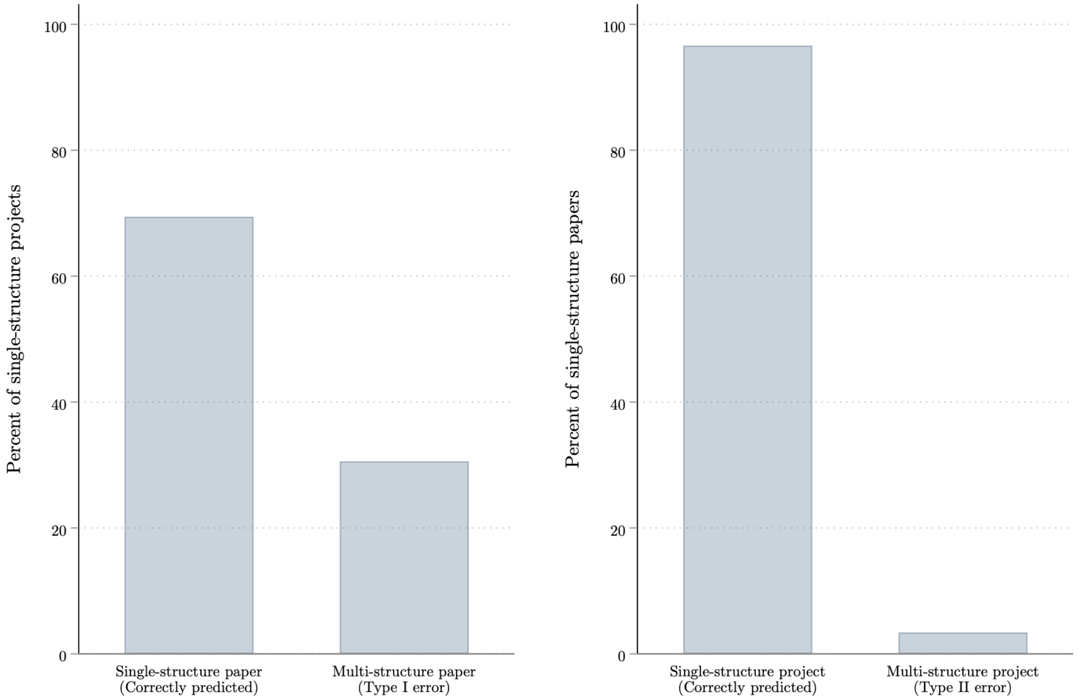
Notes: This figure presents some snapshots from the PDB x-ray structure validation report for PDB ID 4CMP. The "Source" column describes the software package (if applicable) that calculated the quality measure / property.

Figure A7: Difference between Number of Structure Authors versus Number of Paper Authors



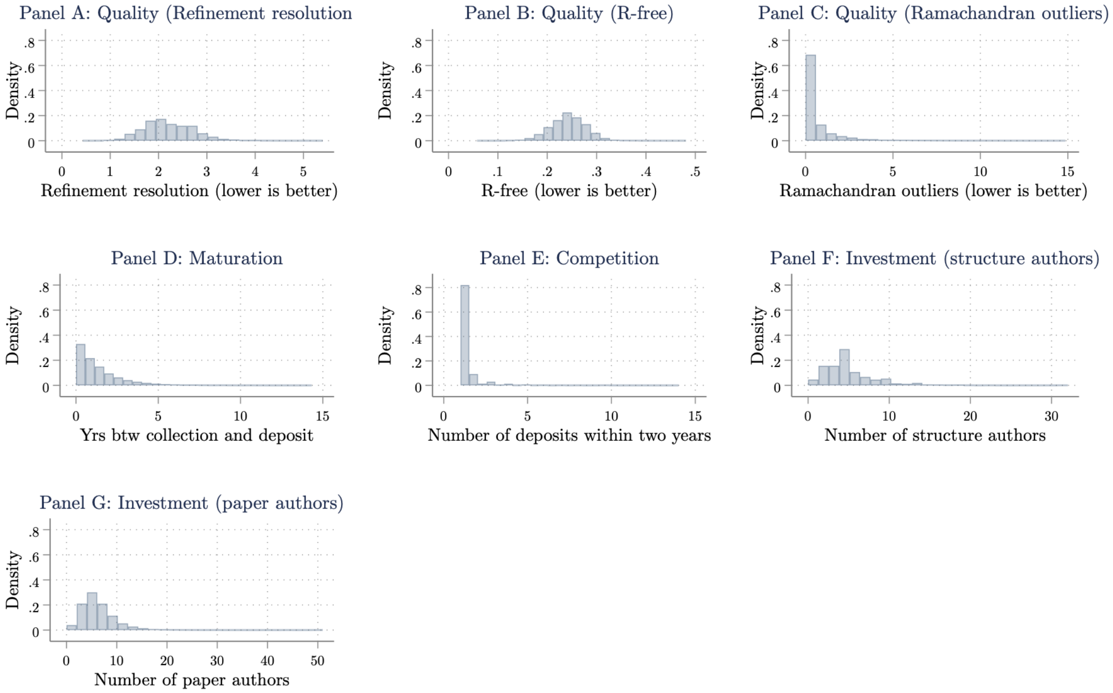
Notes: This figure the difference between the number of paper authors and the number of structure authors. The difference variable has been winsorized at the 1^{st} and 99^{th} percentile. The sample is the full analysis sample, excluding unpublished papers (which lack a paper author count).

Figure A8: Predicting Single-Structure Projects



Notes: This figure assesses how well we predict whether a structure will be the only structure in a paper. Panel A looks at the set of structures we predict will fall in single-structure papers ("single structure projects"). About 70 percent of these are indeed single-structure papers, implying a 30 percent false positive (Type I) error rate. Panel B looks at the set of structures that actually fall in single-structure papers. We predict that 95 percent of these are "single structure projects," implying a 5 percent false negative (Type II) error rate.

Figure A9: Distributions of Key Outcome Variables



Notes: This figure provides histograms of the distributions of our key outcome variables. All variables have winsorized at the 99.9^{th} percentile to make the figures easier to read. The sample is the full analysis sample.

Table A1: Correlation Between Quality Outcomes

	Resolution	R-free	Rama. Outliers
Resolution	1.00		
R-free	0.66	1.00	
Rama. Outliers	0.41	0.43	1.00

Notes: This table shows the correlation between our three quality outcomes. A given cell shows the correlation between the two variables on the x and y-axis.

Table A2: LASSO-Selected Covariates

LASSO-selected variables	Post-LASSO OLS coefficients	LASSO-selected variables	Post-LASSO OLS coefficients
Molecule classification		*Other*	
Isomerase	-12.45	UniProt citations (prior to PDB)	0.085
Lyase	-11.87		
Other	7.43	*Publication Year*	
Oxioreductase	-5.33	1996	25.62
Oxioreductase (CHOH(D)-NAD+(A))	-2.40	1997	20.89
RNA binding protein / RNA	19.07	1998	18.15
Serine esterase	-7.98	1999	17.39
Transferase	-5.03	2000	15.28
Transport Protein	11.10	2001	13.31
Unknown function	-15.81	2002	9.58
		2003	8.62
Macromolecule Type		2015	-3.82
Protein-RNA complex	9.77		
		Constant	46.93
Taxonomy		R-squared	0.17
Homo sapiens	7.46	Observations	13,284
Mycobacterium avium	1.50		
Sapporo virus	1.99		
Gene			
BETVIA	1.68		
BSHA	7.01		
CUL2	5.41		
DESI1	1.90		
INAD	1.08		
ISIB	-13.47		
LINA	13.51		
MAP3K5	7.08		
Missing	-10.61		
MOXF	15.46		
NAGZ	1.99		
NUTF2	1.23		
Other	-3.23		
PEPT	-7.76		
RRM2	-0.47		
THYX	6.93		
TPSAB1	-8.40		
VP40	-0.21		
YWLE	1.90		

Notes: This table presents results from a LASSO regression of cumulative three-year citations (excluding self-citations, transformed to percentiles) on observable protein characteristics. Estimated coefficients are from a post-LASSO OLS regression on the selected characteristics. The coefficients span two sets of columns for readability.

Table A3: The Effect of Potential on Investment and Competition, Bootstrapped Standard Errors

	Investment		Competition
	Number of structure authors	Number of paper authors	Log number of deposits within two years
Dependent variable	(1)	(2)	(3)
Panel A. Without complexity controls			
Potential	0.008	0.030	0.009
OLS SE	(0.0023)	(0.0032)	(0.0004)
Bootstrapped SE	(0.0025)	(0.0040)	(0.0010)
Panel B. With complexity controls			
Potential	0.007	0.033	0.008
OLS SE	(0.0022)	(0.0033)	(0.0004)
Bootstrapped SE	(0.0024)	(0.0041)	(0.0010)

Notes: This table compares the OLS standard errors from Table 2 to the bootstrapped standard errors, which account for the use of generated regressors. Our bootstrapping procedure comprises two steps. First, we randomly draw from our sample with replacement, creating a new sample with the same number of observations as the original sample. We use this new sample to re-generate our potential variable, allowing LASSO to re-select the model. We then use these generated potential measures and the same sample to estimate the OLS relationship between potential and our dependent variable. We repeat this procedure 200 times. The standard deviation in the sample of 200 coefficient estimates is our bootstrapped standard error.

Table A4: The Effect of Potential on Alternative Competition Measures

Dependent variable	Log number of deposits within one year (1)	Log number of deposits (ever) (2)	Priority race (3)
Panel A. Without complexity controls			
Potential	0.006***	0.037***	0.001***
	(0.000)	(0.001)	(0.000)
R-squared	0.036	0.136	0.009
Panel B. With complexity controls			
Potential	0.006***	0.035***	0.001***
	(0.000)	(0.001)	(0.000)
R-squared	0.064	0.173	0.010
Mean of dependent variable	0.143	0.655	0.072
Observations	17,688	17,688	17,688

Notes: This table shows the relationship between additional measures of competition and potential, testing Proposition 4 of the model and estimating regression equation (12) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include molecular weight, residue count, and atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table A5: The Effect of Potential on Maturation and Quality, Bootstrapped Standard Errors

	Maturation	Quality			
		Std.	Std.	Std. Rama.	Std. quality
	Years	resolution	R-free	outliers	index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	-0.005	-0.021	-0.019	-0.012	-0.021
OLS SE	(0.0014)	(0.0008)	(0.0007)	(0.0009)	(0.0007)
Bootstrapped SE	(0.0015)	(0.0009)	(0.0009)	(0.0010)	(0.0010)
Panel B. With complexity controls					
Potential	-0.005	-0.018	-0.018	-0.009	-0.018
OLS SE	(0.0014)	(0.0007)	(0.0007)	(0.0009)	(0.0008)
Bootstrapped SE	(0.0015)	(0.0009)	(0.0010)	(0.0011)	(0.0010)

Notes: This table compares the OLS standard errors from Table 3 to the bootstrapped standard errors, which account for the use of generated regressors. Our bootstrapping procedure comprises two steps. First, we randomly draw from our sample with replacement, creating a new sample with the same number of observations as the original sample. We use this new sample to re-generate our potential variable, allowing LASSO to re-select the model. We then use these generated potential measures and the same sample to estimate the OLS relationship between potential and our dependent variable. We repeat this procedure 200 times. The standard deviation in the sample of 200 coefficient estimates is our bootstrapped standard error.

Appendix B

Appendix for Chapter 2

B.1 Data Appendix

B.1.1 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.[1]

We access the data directly from the RCSB Custom Report Web Service.[2] The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

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

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

- 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.2 below)
- Data Collection Details: collection date (the self-reported date the scientists generated diffraction data at a major synchrotron or in a home lab).

Additional data on cluster entities was accessed through a separate raw file archive at RCSB[3] on December 14, 2018. These files provided additional cluster groupings for the BLAST algorithm at 50 percent and 70 percent sequence similarity.

B.1.2 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.[4] We link the Journal Citations Reports using the journal name listed in the PDB. Each journal has an impact factor in each year and is calculated as the average number of citations per paper in the preceding two years. We standardize 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,[5] 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.

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

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

[5] 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.

B.1.3 Altmetric.com Data

We use data from Altmetric.com to measure alternative forms of attention for academic research.[6] One limitation of the Altmetric data extract we use is that it only reports cumulative counts from the time of publication to the present (date of access: August 2nd, 2019). We account for the fact that scooped papers are published later and have less time to accumulate attention scores, using information about the change in score in recent time periods. The Altmetric.com data reports the change in attention in the past week, month, etc. We can therefore restrict the regression sample to races in which both teams had not accrued any additional attention in the amount of time that had passed between publications. For example, if paper A was released two months before paper B, we do not include this race in the analysis if paper A or paper B had accrued any additional attention in the most recent two months. This allows paper B to have the same window of time to accrue attention despite starting two months late. Because races in our sample end across a wide range of years, the regression coefficients are interpreted as the percent difference in outcomes for papers of an average vintage.

B.1.4 Editorial Dates

We access the received, accepted, and published dates from the websites of publications of Science, Nature Journals, Cell Press, and Public Library of Science. These data are used to compare the scoop date to the timeline of the journal review process as reported in Section 2.4.4.

We also use these data to look at the correspondence between the journal publication date and the release date. Appendix Figure B2 reports the correspondence between the PDB release date and the publication date for the 625 articles in the racing sample for which they are available. This correspondence is not exact for a few reasons. First, according to PDB policy, scientists are allowed to release their findings immediately after deposit, which could potentially come before the publication date. In typical practice, the scientists prefer to wait until publication so that other

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

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 102 races where we have journal publication dates for the winner and loser, the priority ordering as defined by deposit corresponds with the priority ordering as defined by publication 82 percent of the time. To the degree that this is interpreted as measurement error, the scooped estimate will be somewhat attenuated.

B.1.5 Affiliations and University Rankings

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

We link the university affiliations to the QS Top Universities Ranking for Life Sciences and Medicines.[7] 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.

B.1.6 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 2.3 and 2.4) and to predict citations as a measure of team reputation (Figures 2.7 and 2.8). Second, we use a panel of publications to construct long-run outcomes in the years following a scoop event (Table 2.6). The PDB does not explicitly link authors between deposits, and neither PubMed nor Web of Science have author identifiers across publications. A further challenge is that many PDB deposits are not linked to a publication, so constructing control variables of past productivity is difficult using only publication data. We therefore use two separate approaches for constructing author-level panel variables: 1) Link PDB deposits by simple author name matching for control variables, 2) Use name disambiguation clustering from the Author-ity project (Torvik et al. 2005; Torvik and Smalheiser 2009) to count future publications and citations for long-run outcomes.

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

Simple Author Name Matching in PDB

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

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

Scientists usually identify themselves on publications with a consistent last name, but are sometimes inconsistent with their use of first and last initials, or first names and nicknames.[8] 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

[8] 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.

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.

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

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

1. Names that do not match to any Author-ity cluster (11 percent of racing sample authors). These are individuals who deposit at least once in the PDB, but never publish a paper (e.g. a graduate student that does not pursue academia).

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

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 2.3, 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.2 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.2.1 Sequence Similarity Algorithm

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

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

B.2.2 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.

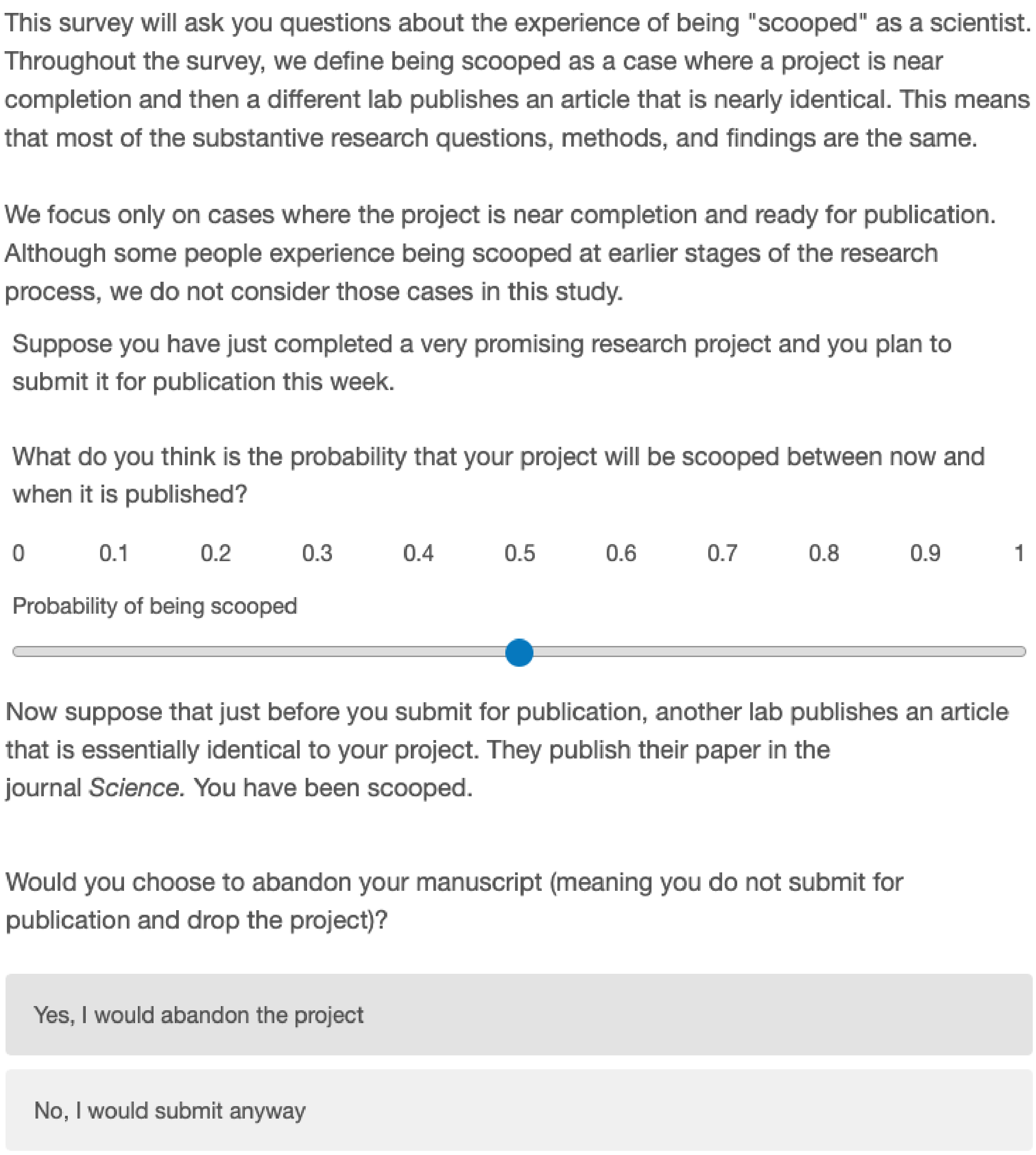
3. Compare the list of structure authors on the priority deposit with the list of authors on all subsequent deposits. Drop any follow-on deposits with one or more author names that were also on the priority deposit.[9]

[9] 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.

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 2.2.3 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 B3, we include robustness results of our main effects for races defined at the 100 percent level, and show that the results are comparable.

B.3 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

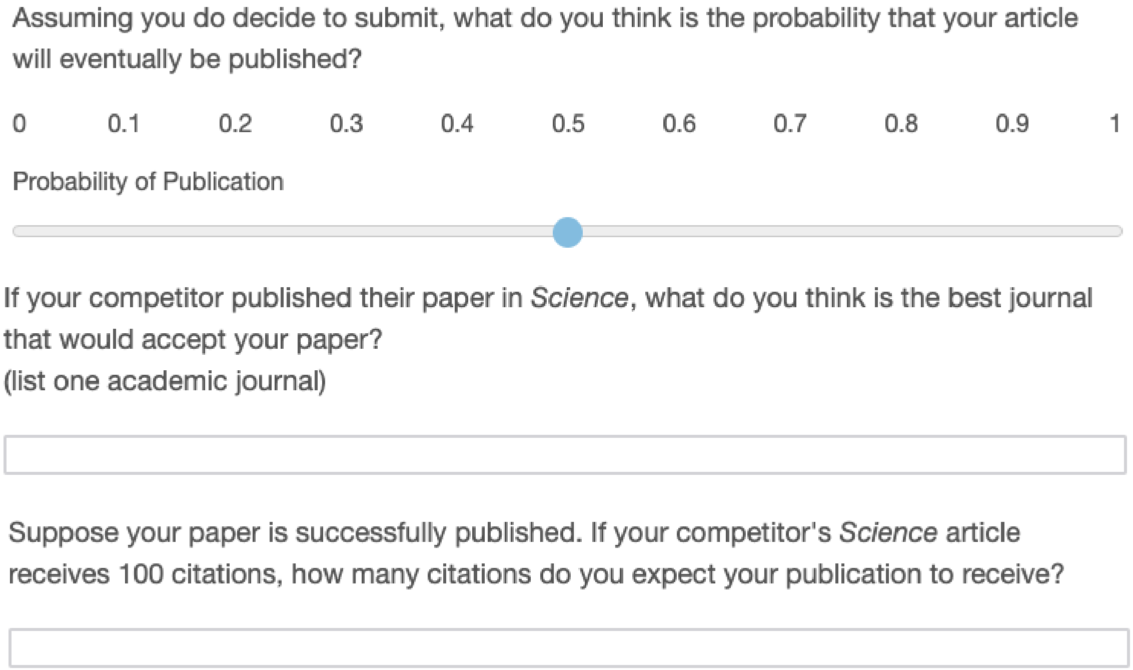
Probability of being scooped

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?

B.4 Proofs of Propositions

Proof of Proposition 1.

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

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

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 $\Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) > \frac{1}{2}$.

Proof of Proposition 2.

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

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

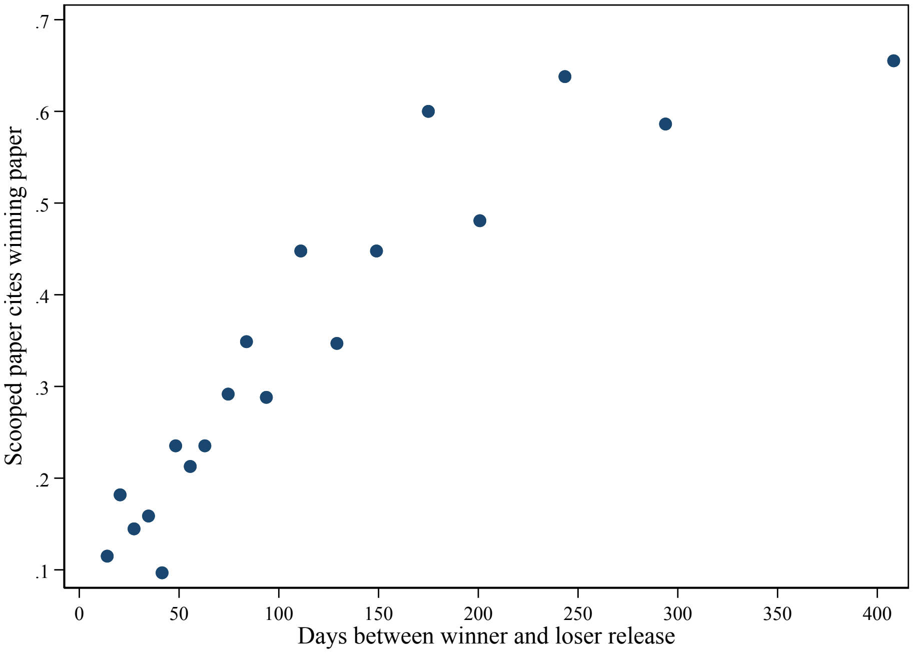
again using the fact that $(u_2 - u_1) \sim N\left(0, 2\sigma_u^2\right)$ and $(1-\lambda) > 0$, $\alpha_H > \alpha_L$. Similarly, consider a low-reputation lab and a high-reputation lab, L_1 and H_2. L_1 publishes before H_2. The probability that L_1 is cited is:

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

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

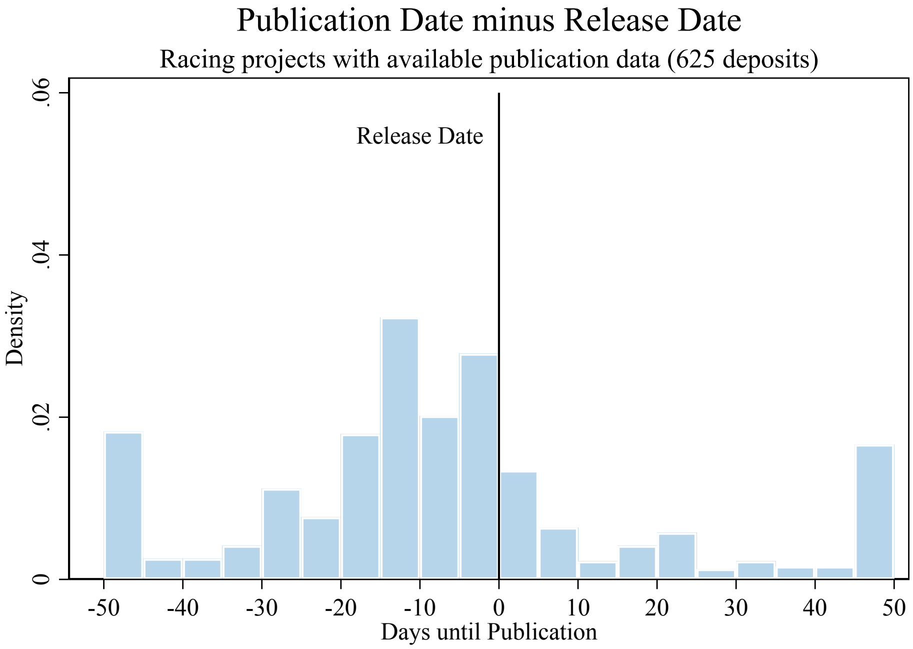
B.5 Appendix Figures and Tables

Figure B1: Probability that Scooped Paper Cites Winning Paper by Release Date Gap



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

Figure B2: 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 B1: Lasso-selected Variables and Coefficients for Predicted Citations

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

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

Table B2: Effect of Getting Scooped on Project Outcomes - Oster (2019) Robustness Check

		Std. journal	Top-ten	Five-year	Top-10% five year
	Published	impact factor	journal	citations	citations
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. No controls, no FE					
Scooped	-0.027**	-0.187***	-0.064***	-0.237***	-0.034***
	(0.011)	(0.031)	(0.014)	(0.050)	(0.010)
	[0.002]	[0.008]	[0.005]	[0.005]	[0.003]
Panel B. Base controls, protein FE					
Scooped	-0.026**	-0.176***	-0.062***	-0.208***	-0.028**
	(0.013)	(0.044)	(0.020)	(0.063)	(0.014)
	[0.704]	[0.675]	[0.604]	[0.762]	[0.725]
Oster (2019) Bias-adjusted β	-0.026	-0.170	-0.061	-0.197	-0.025
Selection ratio (δ) needed for $\beta = 0$	19.7	14.1	13.3	13.4	7.6

Notes: This table presents regression estimates of the scoop penalty following equation 1 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) need 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 B3: Effect of Getting Scooped on Project Outcomes - 100 Percent Sequence Similarity

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

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

Table B4: Effect of Getting Scooped on Three-Year Productivity

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

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

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

Table B5: Effect of Getting Scooped on Ten-Year Productivity

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

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 2 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 4 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 5 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with less than eight years of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.
**p<0.1, **p<0.05, ***p<0.01.*

Table B6: 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	2.586	2.507	0.078	(0.216)	672
Low scoops Low	2.340	2.227	0.113	(0.128)	467
High scoops Low	2.188	2.205	-0.017	(0.074)	498
Low scoops High	2.158	2.155	0.003	(0.053)	652
Panel B. R-free goodness-of-fit					
High scoops High	0.256	0.249	0.007	(0.004) **	649
Low scoops Low	0.245	0.242	0.002	(0.004)	462
High scoops Low	0.242	0.245	-0.003	(0.004)	490
Low scoops High	0.240	0.239	0.002	(0.004)	650

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

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

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix C

Bibliography

Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers, "Competition, Imitation and Growth with Step-by-Step Innovation," *Review of Economic Studies*, 2001, *68*, 467–492.

Aigner, Dennis J. and Glen G. Cain, "Statistical Theories of Discrimination in Labor Markets," *ILR Review*, 1977, *30* (2), 175–187.

Alberts, Bruce, Marc W. Kirschner, Shirly Tilghman, and Harold Varmus, "Rescuing US Biomedical Research from its Systemic Flaws," *Proceedings of the National Academy of Sciences*, 2014, *111* (16), 5773–5777.

Altman, Lawrence K., "U.S. and France End Rift on AIDS," *The New York Times*, 1987.

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.

Anderson, Amy C., "The Process of Structure-Based Drug Design," *Chemistry & Biology*, 2003, *10* (9), 787–797.

Anderson, Melissa S., Emily A. Ronning, Raymond De Vries, and Brian C. Martinson, "The Perverse Effects of Competition on Scientists' Work and Relationships," *Science and Engineering Ethics*, 2007, *13*, 437–461.

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.

Bagues, Manuel F. and Berta Esteve-Volart, "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *The Review of Economic Studies*, 2010, *77*, 1301–1328.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva, "Does the Gender Composition of Scientific Committees Matter?," *American Economic Review*, 2017, *107* (4), 1207–1238.

Bai, Xiah-Chen, Greg McMullan, and Sjors H.W. Scheres, "How Cryo-EM is Revolutionizing Structural Biology," *Trends in Biochemical Sciences*, 2015, *40* (1), 49–57.

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 and Victor Chernozhukov, "High Dimensional Sparse Econommetric Models: An Introduction," 2011.

_ , _ , and Christian Hansen, "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 2014, *81* (2), 608–650.

Berman, Helen, Kim Henrick, Haruki Nakamura, and John L Markley, "The Worldwide Protein Data Bank (wwPDB): Ensuring a Single, Uniform Archive of PDB Data," *Nucleic Acids Research*, 2006, *35*, D301–D303.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T.N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, January 2000, *28* (1), 235–242.

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

Bikard, Michaël, "Simultaneous Discoveries as a Research Tool: Method and Promise," *SSRN Working Paper*, 2013.

_ , "Idea Twins: Simultaneous Discoveries as a Research Tool," *Strategic Management Journal*, 2020, *41* (8), 1528–1543.

Bloom, Floyd E., "Policy Change," *Science*, 1998, *281* (5374).

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.

Broder, Ivy E., “Review of NSF Economics Proposals: Gender and Institutional Patterns,” *American Economic Review*, September 2003, *83* (4), 964–970.

Brown, Eric N. and S. Ramaswamy, “Quality of Protein Crystal Structures,” *Acta Crystallographica Section D*, 2007, *63*, 941–950.

Brünger, Axel T., “Free R Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures,” *Nature*, 1992, *355* (6359), 472–475.

Budish, Eric, Benjamin N. Roin, and Heidi Williams, “Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials,” *American Economic Review*, 2015, *105* (7), 2044–2085.

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., Andrzej Joachimiak, Gaetano T. Montelione, and Ian A. Wilson, “Contributions to the NIH-NIGMS Protein Structure Initiative from PSI Production Centers,” *Structure*, January 2008, *16.*

_ , Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M. Duarte, Shuchismita Dutta et al., “RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy,” *Nucleic Acids Research*, January 2019, *47* (D1), D464–D474.

Campbell, Philip, “New Policy for Structural Data,” *Nature*, July 1998, *394* (6689), 105.

Card, David and Stefano DellaVigna, “What Do Editors Maximize? Evidence from Four Economics Journals,” *The Review of Economics and Statistics*, 2019, *forthcoming.*

_ , _ , Patricia Funk, and Nagore Iriberri, “Are Referees and Editors in Economics Gender Neutral?,” *Quarterly Journal of Economics*, 2020, *135* (1), 269–327.

Carpenter, Elisabeth P., Konstantinos Beis, Alexander D. Cameron, and So Iwata, “Overcoming the Challenges of Membrane Protein Crystallography,” *Current Opinion on Structural Biology*, 2008, *18* (5), 581–586.

Cengiz, Doruk, Arindrajit Dube, Atilla Lindner, and Ben Zipperer, “The Effect of Minimum Wages on the Total Number of Jobs: Evidence from the United States Using a Bunching Estimator,” Working Paper 25434, National Bureau of Economic Research 2019.

Chayen, Naomi E. and Emmanuel Saridakis, “Protein Crystallization: From Purified Protein to Diffraction-Quality Crystal,” *Nature Methods*, 2008, *5*, 147–153.

Cockburn, Iain M., Samuel Kortum, and Scott Stern, "Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes," in Wesley M. Cohen and Stephen A. Merrill, eds., *Patents in the Knowledge-Based Economy*, The National Academies Press, 2003.

Cockburn, Ian and Rebecca Henderson, "Racing to Invest? The Dynamics of Competition in Ethical Drug Discovery," *Journal of Economics & Management Strategy*, 1994, *3* (3), 481–519.

Corum, Jonathan and Carl Zimmer, "Bad News Wrapped in Protein: Inside the Coronavirus Genome," *The New York Times*, 2020.

Cudney, Bob, "Protein Crystallization and Dumb Luck," *The Rigaku Journal*, 1999, *16* (1).

Darwin, Charles, *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*, Vol. 1, John Murray, 1887.

Dasgupta, Partha and Eric Maskin, "The Simple Economics of Research Portfolios," *The Economic Journal*, 581-595 1987, *97.*

_ **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.

Diamond, Arthur M., "What Is a Citation Worth?," *Journal of Human Resources*, 1986, *21* (2), 200–215.

Ellison, Glenn, "How Does the Market Use Citation Data? The Hirsch Index in Economics," *American Economic Journal: Applied Economics*, July 2013, *5* (3), 63–90.

Fang, Ferric C. and Arturo Casadevall, "Competitive Science: Is Competition Ruining Science?," *Infection and Immunity*, 2015, *83* (4452).

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery'," *Journal of Finance*, 2020, *75* (2), 639–682.

Feng, Josh and Xavier Jaravel, "Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 140–181.

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.

Gaulé, Patrick, "Patents and the Success of Venture-Capital Backed Startups: Using Examiner Assignment to Estimate Causal Effects," *Journal of Industrial Economics*, June 2018, *66* (2), 350–376.

Goodsell, David S., "Guide to Understanding PDB Data," Technical Report, Protein Data Bank: PDB-101 2019.

_ , "Methods for Determining Atomic Structures," Technical Report, Protein Data Bank: PDB-101 2019.

Grabowski, Marek, Ewa Niedzialkowska, Matthew D. Zimmerman, and Wladek Minor, "The Impact of Structural Genomics: The First Quindecennial," *Journal of Structural Functional Genomics*, 2016, *17* (1), 1–16.

Graham, Stuart, Alan Marco, and Richard Miller, "The USPTO Patent Examination Research Dataset: A Window on the Process of Patent Examination," 2015. November 30, 2015.

Green, Jerry R. and Suzanne Scotchmer, "On the Division of Profit in Sequential Innovation," *RAND Journal of Economics*, 1995, *26* (1), 20–33.

Grossman, Gene and Elhanan Helpman, "Quality Ladders in the Theory of Growth," *Review of Economic Studies*, 1991, *58* (1), 43–61.

Hagstrom, Warren O., *The Scientific Community*, Basic Books, 1965.

_ , "Competition in Science," *American Sociological Review*, February 1974, *39* (1), 1–18.

Hamermesh, Daniel and Gerard Pfann, "Reputation and Earnings: The Roles of Quality and Quantity in Academe," *Economic Inquiry*, January 2012, *50* (1), 1–16.

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

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

Hengel, Erin, "Publishing While Female," *Working Paper*, 2018.

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

_ and Carolyn Stein, “Race to the Bottom: Competition and Quality in Science,” *Working Paper*, 2020.

_ and _ , “Scooped! Estimating Rewards for Priority in Science,” *Working Paper*, 2020.

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

Holmstrom, Bengt, “The Firm as a Subeconomy,” *Journal of Law, Economics, & Organization*, 1999, *15* (1), 74–102.

Hong, Wei and John P. Walsh, “For Money or For Glory? Commercialization, Competition, and Secrecy in the Entrepreneurial University,” *The Sociological Quarterly*, 2009, *50*, 145–171.

Hopenhayn, Hugo and Francesco Squintani, “Patent Rights and Innovation Disclosure,” *Review of Economic Studies*, 2016, *83* (199-230).

Horstmann, Ignatius, Glenn M. MacDonald, and Alan Slivinski, “Patents as Information Transfer Mechanisms: To Patent or (Maybe) Not to Patent,” *Journal of Political Economy*, 1985, *93* (5), 837–858.

Huang, Wenya, Peiling Wang, and Qiang Wu, “A Correlation Comparison Between Altmetric Attention Scores and Citations for Six PLOS Journals,” *PloS One*, 2018, *13* (4).

Hunt, Jennifer, Jean-Philippe Garant, Hannah Herman, and David J. Munroe, “Why Don’t Women Patent?,” *NBER Working Paper 17888*, 2012.

Jacob, Brian and Lars Lefgren, “The Impact of NIH Postdoctoral Training Grants on Scientific Productivity,” *Research Policy*, 2011, *40* (6), 864–874.

Jardim, Ekaterina, Mark C. Long, Robert Plotnick, Emma van Inwegen, Jacob Vigdor, and Hilary Wething, “Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle,” Working Paper 23532, National Bureau of Economic Research 2018.

Jensen, Kyle, Balazs Kovacs, and Olav Sorensen, “Gender Differences in Obtaining and Maintaining Patent Rights,” *Nature Biotechnology*, April 2018, *36* (4), 307–309.

Justman, Quincey, “Scooping Hurts Science and Scientists,” *Cell Systems*, 2018, *7* (469-470).

Lattman, Eaton E., “No Crystals No Grant,” *Proteins: Structure, Function, and Genetics*, 1996, *26.*

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

Lemley, Mark A. and Bhaven N. Sampat, "Is the Patent Office a Rubber Stamp?," *Emory Law Journal*, 2008, *58* (1), 415–427.

_ and _ , "Examining Patent Examination," *Stanford Technology Law Review*, 2010, *2010.*

_ and _ , "Examiner Characteristics and Patent Office Outcomes," *Review of Economics and Statistics*, August 2012, *94* (3), 817–827.

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.*

Martz, Eric and Eran Hodis, "Free R," 2013.

_ , 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.

_ , "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science," *Proceedings of the American Philosophical Society*, October 1961, *105* (5), 470–486.

_ , "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.

Minor, Wladek, Zbigniew Dauter, and Mariusz Jaskolski, "A Young Person's Guide to the PDB," *Postepy Biochem*, 2016, *62* (3), 242–249.

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

Murphy, Kevin M. and Robert H. Topel, "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, 1985, *3* (4), 370–379.

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

National Science Foundation, "Science and Engineering Labor Force," *Science & Engineering Indicators*, 2018.

Nature Editors, "Must Try Harder," *Nature*, 2012, *483* (7391), 509.

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.

Ouellette, Lisa L., "Pierson, Peer Review, and Patent Law," *Vanderbilt Law Review*, 2019, *69* (6), 1825–1848.

Pagan, Adrian, "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 1984, *25* (1), 221–247.

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).

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of Polypeptide Chain Configurations," *Journal of Molecular Biology*, 1963, *7* (1), 95–99.

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

Read, Randy J., Paul D. Adams, W. Bryan Arendall III, and Peter H. Zwart, "A New Generation of Crystallographic Validation Tools for the Protein Data Bank," *Structure*, 2011, *19* (10), 1395–1412.

Rhodes, Gail, *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, Elsevier Science and Technology, 2006.

Sampat, Bhaven and Heidi L. Williams, "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, January 2019, *109* (1), 203–236.

Sarsons, Heather, "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," *Working Paper*, 2019.

Scotchmer, Suzanne and Jerry R. Green, "Novelty and Disclosure in Patent Law," *RAND Journal of Economics*, 1990, *21* (131-146).

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.

Sibley, Charles G., "The Electrophoretic Patterns of Avian Egg-White Proteins as Taxonomic Characters," *Ibis*, 1960, *102*, 215–284.

Stephan, Paula E., "The Economics of Science," *Journal of Economic Literature*, 1996, *34* (3), 1199–1235.

_ , *How Economics Shapes Science*, Harvard University Press, 2012.

Strasser, Bruno J., *Collecting Experiments*, The University of Chicago Press, 2019.

Sullivan, Kevin P., Peggy Brennan-Tonetta, and Lucas J. Marxen, "Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank," Technical Report, Office of Research Analytics, Rutgers 2017.

Sussman, Joel L., "What's New at the PDB," *Quarterly Newsletter published by Brookhaven National Laboratory Protein Data Bank*, April 1998, *84*, 1.

The PLOS Biology Staff Editors, "The Importance of Being Second," *PLOS Biology*, 2018, *16* (1).

The UniProt Consortium, "UniProt: A Worldwide Hub of Protein Knowledge," *Nucleic Acids Research*, 2019, *47* (D1), D506–D515.

Thompson, Neil C. and Jeffrey M. Kuhn, "Does Winning a Patent Race Lead to More Follow-on Innovation?," *SSRN Working Paper*, January 2017.

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

Tiokhin, Leonid and Maxime Derex, "Competition for Novelty Reduces Information Sampling in a Research Game - A Registered Report," *Royal Society Open Science*, 2019, *6*.

_ , Minhua Yan, and Thomas Morgan, "Competition for Priority and the Cultural Evolution of Research Strategies," *MetaArXiv Preprints*, 2020.

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.

Tuckman, Howard and Jack Leahey, "What Is an Article Worth?," *Journal of Political Economy*, 1975, *83* (5), 951–967.

US Patent and Trademark Office, "U.S. Patent Statistics Chart: Calendar Years 1963-2015," 2016. Retrieved from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm.

_ , *Performance and Accountability Report FY17* 2017.

_ , "Progress and Potential: A Profile of Women Inventors on U.S. Patents," 2019. February 2019.

Vale, Ronald D. and Anthony A. Hyman, "Priority of Discovery in the Life Sciences," *eLife*, 2016, *5*.

Walsh, John P. and Wei Hong, "Secrecy is Increasing in Step with Competition," *Nature*, 2003, *422* (6934), 801.

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

Westbrook, John D. and Stephen K. Burley, "How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals," *Structure*, 2018, *27*, 1–7.

Williams, Heidi, "How Do Patents Affect Research Investments?," *Annual Review of Economics*, 2017, *9* (1), 441–469.

Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson, "DrugBank 5.0: A Major Update to the DrugBank Database for 2018," *Nucleic Acids Research*, 2018, *46* (D1), 1074–1082.

Wlodawer, Alexander and Jiri Vondrasek, "Inhibitors of HIV-1 Protease: A Major Success of Structure-Assisted Drug Design," *Annual Review of Biophysics and Biomolecular Structure*, 1998, *27*, 249–284.

_ , 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.

Worldwide Protein Data Bank, "wwPDB 2013 News," 2013.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan, "Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation," *Science*, 2020, *367* (6483), 1260–1263.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi, "The Increasing Dominance of Teams in Produciton of Knowledge," *Science*, 2007, *316*, 1036–1039.

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.