RACE TO THE BOTTOM: COMPETITION AND QUALITY IN SCIENCE*

RYAN HILL AND CAROLYN STEIN

This article investigates how competition to publish first and thereby establish priority affects 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 they should let their projects mature, scientists trade off the marginal benefit of higher-quality research against the marginal risk of being preempted. Projects with the highest scientific potential 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. We estimate that the costs associated with improving these low-quality structures are between $1.5 and $8.8 billion since the PDB's founding in 1971.
JEL codes: D82, I11, I23, O31, O36.

*We are deeply grateful to our advisers, 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, Glenn Ellison, Chishio Furukawa, Matthew Gentzkow, Colin Gray, Sam Hanson, Ariella Kahn-Lang, Layne Kirshon, Sam Kortum, Matt Notowidigdo, Tamar Oostrom, Jonathan Roth, Adrienne Sabety, Bhaven Sampat, Michael Stepner, Jeremy Stein, Sean Wang, Michael Wong, and numerous seminar participants for thoughtful comments. This material is based on 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.

The Quarterly Journal of Economics (2025), 1–75. https://doi.org/10.1093/qje/qjaf010.
Advance Access publication on February 3, 2025.

I. 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; Dasgupta and David 1994; Stephan 1996). As described by Merton (1957), credit comes at least in part from disclosing one's findings first, thereby establishing priority. Thus, it is not surprising that scientists compete intensely to publish important findings first. Indeed, scientific history has been punctuated with cutthroat races and fierce disputes over priority (Merton 1961; Bikard 2020).[1] This competition and fear of preemption or "getting scooped" 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 one-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 forthcoming) suggests that preemption remains common and the threat of being scooped continues to be perceived as a serious concern.

Competition for priority has potential benefits and costs for science. Winner-take-all (or winner-take-most) compensation schemes can induce researchers to exert costly effort, as emphasized by the tournament literature (Lazear and Rosen 1981; Nalebuff and Stiglitz 1983). This can hasten the pace of discovery and incentivize timely disclosure. However, this same competition may have a dark side if, as highlighted by Dasgupta and David (1994), it induces researchers to engage in "deviant" patterns of behavior. These behaviors can take many forms, from secrecy and incomplete disclosure (Walsh and Hong 2003),[2] to more

1. To name 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 that bore an uncanny resemblance to Darwin's (then 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). For more examples, see chapter 10 of Lamb and Easton (1984).

2. Indeed, Dasgupta and David (1994) even highlight structural biology as an example of when researchers would intentionally delay sharing their experimen-

extreme behaviors, such as the intentional sabotage of competitors (Anderson et al. 2007). In this article, we focus on a particular behavior: the pressure to publish quickly and preempt competitors may lead to "quick and dirty experiments" rather than "careful, methodical work" (Anderson et al. 2007; Yong 2018).[3] In other words, the faster pace of research may lead to lower-quality science. The goal of this article is to assess 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. Moreover, because important projects tend to be the most competitive, we find that important projects are also lower quality. We present a range of evidence that supports a causal relationship between competition and lower-quality research rather than a spurious relationship driven by omitted factors.

We begin by developing a model where researchers race to publish their findings in a secretive field. The winner of the race receives a larger reward than does the runner-up. Because researchers cannot observe each others' progress (in other words, there is no learning until the race is over), this model is similar in spirit to the memoryless patent race models developed by Loury (1979), Lee and Wilde (1980), Dasgupta and Stiglitz (1980), and Reinganum (1981), where past research and development (R&D) spending does not affect the probability of success.[4] No team can visibly pull ahead, and therefore researchers compete vigorously. In our model, researchers work to develop ideas that arise exogenously. However, ideas alone cannot be published.

tal data, citing this as a form of incomplete disclosure. Since 1999, most scientific journals now require structural biologists to deposit their data at the time of publication. Another important type of incomplete disclosure is the failure to disclose "dead ends," as emphasized by Akcigit and Liu (2016), which can lead to inefficient duplication of effort.

3. Scientists have long voiced this concern. 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).

4. In these memoryless patent race models, breakthroughs are drawn from exponential distributions. R&D spending affects the rate parameter of the distribution. Thus, past R&D spending does not affect the current probability of success in these models, so players do not learn or update their strategies over time. This is similar to our one-shot model. See Reinganum (1989) for a review of the patent-race literature and memoryless patent races specifically.

Similar to Hopenhayn and Squintani (2016) and particularly Bobtcheff, Bolte, and Mariotti 2017, ideas must be developed. The longer ideas are allowed to mature, the better the quality of the resulting research. Thus, researchers face a tension between letting the projects mature for longer (improving the quality of the research) and publishing quickly (to minimize the risk of being preempted). As a result, the threat of competition leads to lower-quality projects than if the scientist knew she was working in isolation.[5]

In a departure from Bobtcheff, Bolte, and Mariotti 2017, 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 leads scientists to prematurely publish their findings. The key prediction of the model is that high-potential projects—those tackling questions the scientific community has deemed the most important—are the projects that will also be executed with the lowest quality.

Although the model provides a helpful framework, the primary contribution of this study is to provide empirical evidence for the theoretical forces that it describes. Compared with the rich theoretical literature, far fewer papers have studied innovation races empirically; Cockburn and Henderson (1994), Lerner (1997), and Thompson and Kuhn (2020) are notable exceptions. To test the predictions of our specific model, we require a setting

5. Tiokhin, Yan, and Morgan (2021) 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 hypothesis in a lab experiment.

that satisfies four demanding criteria. First, we need a field where discoveries are discrete, self-contained, and comparable. Second, we must be able to measure projects' distance from one another in idea space, to construct project-level measures of scientific competition. Third, we need a way to score projects in terms of their ex ante potential. This is critical for testing the core predictions of our model. Last, we require measures of the quality of scientific work. By quality, we mean quality of execution—not a measure of the paper's interest or importance.[6]

We make progress on all of these challenges in structural 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 in an effort to publicize this information and promote the use of these structures for follow-up work (Berman et al. 2000; Strasser 2019). This rich setting satisfies the four criteria. First, in structural biology, research projects center around consistent experimental methods to deduce the three-dimensional structure of known proteins. Thus, individual projects are well defined and comparable, satisfying our first criteria. Second, projects are grouped together by structure similarity, and progress is timestamped. This allows us to identify competitive proteins: structures that are identical and being worked on contemporaneously.[7] Third, the PDB provides a rich array of characteristics about each protein, such as the protein type, the protein's organism, the gene–protein linkage, and the prior number of papers written about the protein. These are all characteristics the researcher would observe before starting a project and would inform her view of its potential. We construct a measure of potential by using these characteristics to predict the number of citations the structure will ultimately receive. Last, every macromolecular structure is scored on a variety of quality metrics. At a high level, structural biologists are

6. Some studies (Hengel 2022) have used text analysis to measure a paper's readability as a proxy for 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, Stuart, and Wang 2014)—a distinction that is critical for our research question.

7. This is a more context-specific application of Bikard (2020)'s concept of simultaneous discoveries or "idea twins."

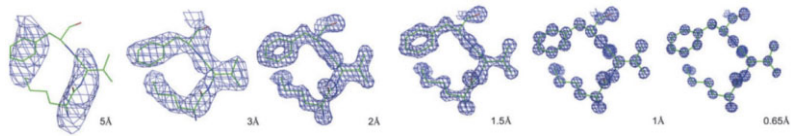

FIGURE I

Protein Structure at Different Refinement Resolutions

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. The figure is taken from Wlodawer et al. (2008), "Protein Crystallography for Non-Crystallographers, or How to Get the Best (But Not More) from Published Macroemolecular Structures," *FEBS Journal*, 275 (2008), 1–21. Copyright © (2007) John Wiley and Sons. Reprinted with permission from John Wiley and Sons.

concerned with fitting three-dimensional structure models to experimental data, so these quality metrics are measures of goodness of fit. They allow us to compare quality across different projects in an objective, scientific way. 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. Figure I shows the same protein structure solved at different refinement resolutions to illustrate these quality differences.

We use our computed values of potential to test the key predictions of the model. Comparing structures in the 90th versus 10th percentile of the potential distribution, we find that high-potential projects induce meaningfully more competition. High-potential structures are 4 percentage points (60%) more likely to be involved in a priority race. This suggests that more researchers are pursuing the most important (and highest citation-generating) structures. We look at how project potential affects maturation and quality. We find that high-potential structures are completed more than two months faster and have quality measures that are about 0.7 standard deviations lower than low-potential structures. These results echo findings by 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.[8]

8. Brown and Ramaswamy propose multiple reasons this might be the case. One is increased competition among structures published in top journals. Another

A concern when interpreting these results is that 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 attempt to control for complexity directly, which has a minimal effect on the magnitude of our estimates. Second, we leverage another source of variation: 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 whom face the priority incentives described above to publish early. 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 making future structure discovery easier (The Structural Genomics Consortium 2020; Zhou 2023). Qualitative evidence suggests that these groups are less focused on publication and priority, which is consistent with the fact that only about 20% of SG structures ever appear in journal publications, compared with more than 80% 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 with non-SG structures. If complexity is correlated with potential, this should be the case for both the SG and non-SG structures. Intuitively, by comparing the slopes across both groups, we can "net out" the potential omitted-variables bias. Consistent with competition acting as the causal channel, we find a more negative relationship between potential and quality among non-SG (i.e., more competitive) structures.

Last, we design a survey experiment that tests the predictions of our model directly. By randomly varying perceived potential, we find that researchers indeed expect high-potential projects to be more competitive. By randomly varying perceived competition, we confirm that researchers work more quickly and

is that top journals tend to be more general interest and less specialized, and therefore reviewers may not be as able to evaluate structure quality. However, we find a quantitatively similar negative relationship between potential and quality within journal, which suggests that the latter explanation cannot fully explain the authors' findings.

perform fewer quality control checks when they believe their project is very competitive. This provides additional evidence that the competition channel explains the negative relationship between potential and quality. Because we surveyed researchers across multiple fields of science, this analysis provides evidence that the phenomenon we identify is not unique to structural biology.

We conclude the article by turning to the welfare costs of this racing behavior. Ideally, we would like to compare the behavior of individual scientists (who care about priority) to a benevolent social planner (who only cares about knowledge generation, not *who* generates it). In practice, the SG researchers represent a reasonable approximation of this social planner. By comparing the behavior of lab-based researchers with that of their SG counterparts, we can estimate the welfare costs that arise from racing. We already know that non-SG researchers do lower-quality work than SG researchers when working on high-potential structures. Yet if we look at follow-on work (new deposits of the same structure), we find that most of the quality is eventually recovered. Thus, low quality does not seem to be the main cost in the long run. However, given the experimental nature of this work, it is difficult to improve protein structures. The vast majority of the time, improving a protein structure requires an entirely new experiment. This model of an initial low-quality structure followed by a subsequent improvement is inefficient—it would be less costly for the first team to slow down and do a careful job the first time. Given estimates that it costs $120,000 to replicate a typical protein structure, we calculate that researchers have spent between $2.3 and $6.6 billion in an effort to improve low-quality structures generated by racing behavior since the PDB's founding in 1971. This interval widens if we consider a range of cost estimates.

The remainder of the article proceeds as follows. Section II presents the model, and Section III describes our setting and data. Section IV tests the predictions of the model in the observational data; Section V tests the predictions of the model using a survey experiment. Section VI considers the welfare implications. Section VII concludes.

II. 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 perhaps intuitive. Our goal here is to develop a model that formalizes this insight and generates additional testable predictions. Scientists in our model are rational agents, seeking to maximize the total credit or recognition they receive for their work.[9] We allow projects to differ in terms of expected payoffs. Scientists must decide whether to start a project and, conditional on starting, how long to spend on it. More time spent working on a project translates to higher-quality work. The threat of competition induces scientists to spend less time working on a project. This threat is particularly acute for high-payoff projects, because more scientists choose to start these projects. We walk through the basic framework of the model below and direct interested readers to a more formal treatment in Online Appendix A.

II.A. Preliminaries

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

2. *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. The scientist must pay an upfront cost 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 $g(I_i) \in [0, 1]$ 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

9. 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).

generate preliminary results. In our setting, a natural interpretation is that I represents the time and resources spent trying to grow a protein crystal.

The second stage occurs if the scientist successfully starts the project.[10] 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 complete at $t_i^F = t_i^S + m_i$.

3. *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 the researcher achieved. The total value of the project is thus 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. 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

$$\overline{\theta} e^{-rm} PQ(m). \tag{1}$$

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

$$\underline{\theta} e^{-rm} PQ(m). \tag{2}$$

10. Note that before the second stage, the scientist learns about her own entry success. However, no information about her opponent is revealed. Thus, there are no subgames in this model and therefore no notion of subgame perfection.

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

4. *Information Structure.* The competing scientists have limited information about their competitor's progress in the race. Scientist i does not observe I_j, 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. We assume that she believes 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 $\Delta > 0$.[11] Online Appendix Figure A1 summarizes the model setup.

II.B. The Maturation Decision

We work backward, solving the second-stage problem of the optimal maturation delay, taking both teams' first-stage investment decision as given. Let $\pi(m_i, m_j, I_j)$ denote the probability that scientist i wins the race, conditional on successfully entering. We write this as simply π for convenience. This probability will depend on the likelihood that j is in the race (otherwise i wins by default) and each player's choice of maturation. Then scientist i's

11. Researcher i's beliefs about j's start time being identically distributed around her own, no matter her value of t_i^S, implies that there is no notion of starting "early" or "late." This simplifies the model, because it means that the optimal maturation choice does not depend on t. Note that the uniformity assumption is not critical—it merely simplifies some expressions. One way to microfound such a model is to assume that t_i^S and t_j^S are random variables, but there is uncertainty about the support of the distribution from which they are drawn. Thus, a single draw is not informative about whether the player is early or late, so players cannot infer their relative position (Abreu and Brunnermeier 2003).

best response to scientist j is given by:

$$m_i^*(m_j) \in \arg\max_{m_i} \left\{ \underbrace{e^{-rm_i} PQ(m_i)}_{\text{full PDV of project}} \underbrace{\left[\pi\overline{\theta} + (1-\pi)\underline{\theta}\right]}_{\text{expected credit share}} \right\}. \tag{3}$$

We show in Online Appendix A that under mild assumptions, there is a unique and symmetric pure-strategy Nash equilibrium, where both researchers select the same m^*. Moreover, this choice of maturation is shorter when (i) the difference between $\overline{\theta}$ and $\underline{\theta}$ is large (priority rewards are more lopsided), (ii) Δ is small (competitors start projects close together on average, so the "flow risk" of getting scooped is high), or when g is close to one (the entry of a competitor is likely).

II.C. The Entry Decision

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. Recall that $g(I_i)$ is the probability that she is successful conditional on a given level of investment. Scientist i's best response to j's investment choice is given by

$$I_i^*(I_j) \in \arg\max_{I_i} \left\{ \underbrace{g(I_i)}_{\text{prob. of successful entry}} \underbrace{e^{-rm_i^*} PQ(m_i^*)}_{\text{full PDV of project}} \underbrace{\left[\pi\overline{\theta} + (1-\pi)\underline{\theta}\right]}_{\text{expected credit share}} - \underbrace{I_i}_{\text{investment cost}} \right\}. \tag{4}$$

We show in Online Appendix A that there is a unique and symmetric pure-strategy Nash equilibrium for investment.

II.D. Model Predictions

So far, we have defined the optimal investment level and maturation period when entry into projects is endogenous. This allows us to prove three key results.

PROPOSITION 1. 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, $\frac{dm^*}{dg} < 0$ and $\frac{dQ(m^*)}{dg} < 0$.

Proof. See Online Appendix A. Scientist i selects m_i^* 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 2. Higher-potential projects generate more investment and are therefore more competitive. In other words, $\frac{dI^*}{dP} > 0$ and $\frac{dg(I^*)}{dP} > 0$.

Proof. See Online Appendix A. 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 3. Higher-potential projects are completed more quickly and are therefore of lower quality. In other words, $\frac{dm^*}{dP} < 0$ and $\frac{dQ(m^*)}{dP} < 0$.

Proof. This comes immediately from Propositions 1 and 2, by applying the chain rule. □

These are the three predictions that we take to the data in Section IV. Proposition 1 predicts that there should be a negative correlation between a project's level of competition and its maturation period and quality. Proposition 2 predicts that there should be a positive correlation between a project's potential and its level of competition. Proposition 3 predicts that there should be a negative correlation between a project's potential and its maturation period and quality. The next section explains how we are able to measure these theoretical quantities using data from structural biology, allowing us to test these predictions empirically.

III. STRUCTURAL BIOLOGY AND THE PDB

We provide some scientific background on structural biology and describe 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 a clean link between our theoretical model and

our empirical setting. Online Appendix B provides additional detail on our data sources and construction.

III.A. Structural Biology

Structural biology is the study of the three-dimensional structure of biological macromolecules, including deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and most commonly, proteins. Understanding how macromolecules perform their functions in cells is one of the key themes in molecular biology. Structural biologists shed light on these questions by determining the 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.[12] 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, and 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 disease, Alzheimer's disease, and Huntington's disease—are the direct result of protein misfolding. Protein structures play a critical role in drug development and vaccine design (Westbrook and Burley 2019).[13] Over a dozen Nobel Prizes have been awarded for advances in the field (Martz et al. 2019).

1. *Why Structural Biology.* Our empirical work focuses on the field of structural biology for several reasons. First, projects in this field are well-defined and comparable—they aim to solve

12. Recent developments in the field of cryo-electron microscopy now allow scientists to observe larger structures directly (Bai, McMullan, and Scheres 2015). Despite the recent growth in this technique, less than 5% of PDB structures deposited since 2015 have used this method in our sample.

13. 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).

the structure of a known protein. This makes cross-project comparisons sensible. Second, we can use the amino acid sequence of proteins to determine how close two proteins are to each other in idea space. Projects include timestamps indicating when particular milestones were reached in the process. The combination of these features allows us to identify proteins that are (i) identical or nearly identical and (ii) being solved contemporaneously. We define these proteins as being involved in a competitive priority race.

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 III.E.

Finally, and most importantly, structural biology has unique measures of objective project quality. Scientists deposit their structural models in the PDB, and there are several measures of how precise and correct their solutions are. We discuss these measures in later sections; here 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.

2. *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%) of structures in our time frame are solved using this method.

X-ray crystallography broadly consists of three steps (see Figure II). Individual proteins are too small to analyze or observe directly. 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

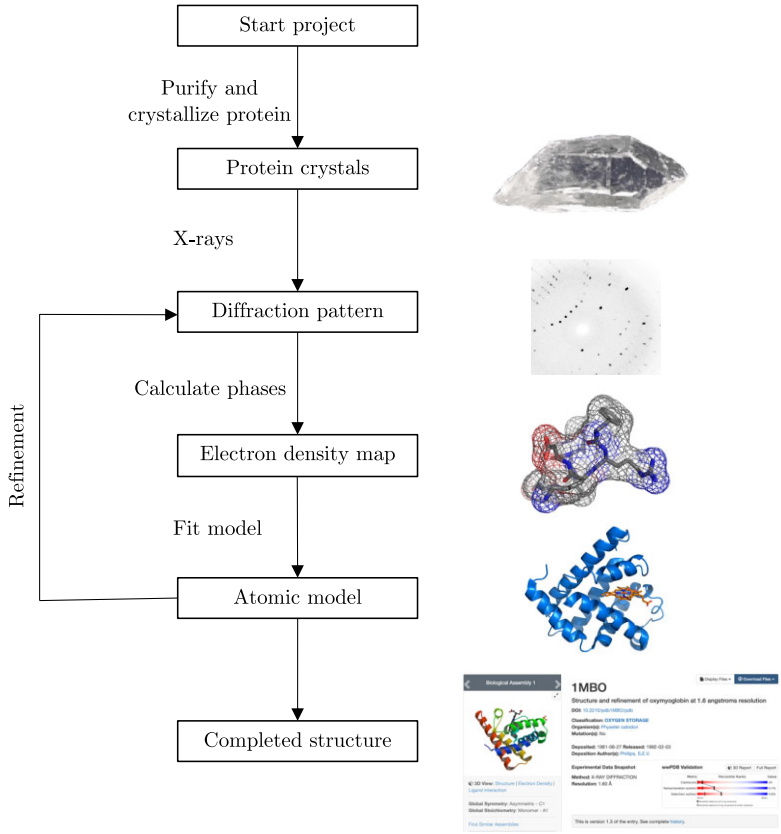

FIGURE II

Summary of the X-Ray Crystallography Process

This figure summarizes the process of solving a protein structure via X-ray crystallography. The images in this figure were taken from Thomas Splettstoesser (https://www.scistyle.com) and rendered with PyMol based on PDB ID 1MBO. Reuse of the figure is licensed under a Creative Commons Attribution-Share Alike 3.0 Unported license.

art than science" (Rhodes 2006) or at times simply "dumb luck" (Cudney 1999). Success typically comes from trial and error.[14]

14. 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

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 the 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 compares simulated diffraction data from her model to her actual experimental data and adjusts 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, Dauter, and Jaskolski 2016, 3). In other words, 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.

III.B. The Protein Data Bank

Our primary data source is the PDB, a worldwide repository of biological macromolecules, 95% of which are proteins.[15] It was established in 1971 with just 7 entries, and by the end of our sample period it contained nearly 150,000 structures. Its goal is to promote the dissemination and further use of protein struc-

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, 1).

15. Because the vast majority of structures deposited to the PDB are proteins, we use the terms "structure" and "protein" interchangeably.

tures by structural biologists and scientists in other fields.[16] 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. 2016, 2000; Strasser 2019). Therefore, the PDB represents a near-universe of macromolecule structure discoveries. 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.[17]

1. *Measuring Quality.* The PDB provides several measures intended to assess quality. These measures were developed by the X-Ray Validation Task Force 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 structure errors, increasing overall user confidence, and it makes the PDB more useful and accessible for scientists without the specialized knowledge to critically evaluate structure quality. Below we describe the three measures used in our empirical analysis. We selected these three because they are scientifically distinct and have good coverage in our data. We combine these measures into a single quality index. Together, these measures map closely 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 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 before 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 histor-

16. Indeed, the PDB is a great example of the importance of scientific institutions in cumulative research, as highlighted by (Furman and Stern 2011) in the context of biological resource centers, and more recently by Thompson and Zyontz (2017) in the context of plasmid repositories.

17. Some structures are composed of multiple "entities," and some variables are indexed at the entity level. We discuss this in more detail in Online Appendix B.

ical coverage for these quality measures. Online Appendix Figure E1 provides a snapshot from one of these reports.

i. 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 are more detailed. This allows for better electron density maps, as shown in Figure I. At resolutions less than 1.5 Å, individual atoms can be resolved and structures have almost no errors. At resolutions greater than 4 Å, individual atomic coordinates are meaningless, and only secondary structures can be determined. Scientists can improve resolution by spending time improving the quality of the protein crystals and fine-tuning the experimental conditions during X-ray exposure. In the main analysis, we standardize refinement resolution so that the units are in standard deviations and higher values represent better quality.

ii. 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 10% 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 (Brünger 1992; Goodsell 2019). Most crystallographers agree that 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 2 Å or better should have an R-free of $(\frac{\text{resolution}}{10} + 0.05)$ or better. In other words, if the resolution is 2 Å, 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

the main analysis, we standardize R-free so that the units are in standard deviations and higher values represent better quality.

iii. Ramachandran Outliers. Ramachandran outliers are a 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, Ramakrishnan, and Sasisekharan 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 the main analysis, we will standardize Ramachandran outliers so that the units are in standard deviations and higher values represent better quality.

iv. Quality Index. Finally, we combine the measures into a single quality index. All three measures are correlated, with correlation coefficients in the 0.4–0.6 range (see Online Appendix Table E1). We create the index by adding all three standardized quality measures and standardizing the sum. Throughout our analysis, this index is our primary measure of quality. However, all of our results are robust to each of the individual quality measures, which we report in the Online Appendix.

2. *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, that is, 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 provide unique insights about project timelines.

As shown in Figure III, the PDB collects two key dates that allow us to infer the maturation period: the collection date and the deposition date. The collection date is self-reported and 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 (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 the paper for peer re-

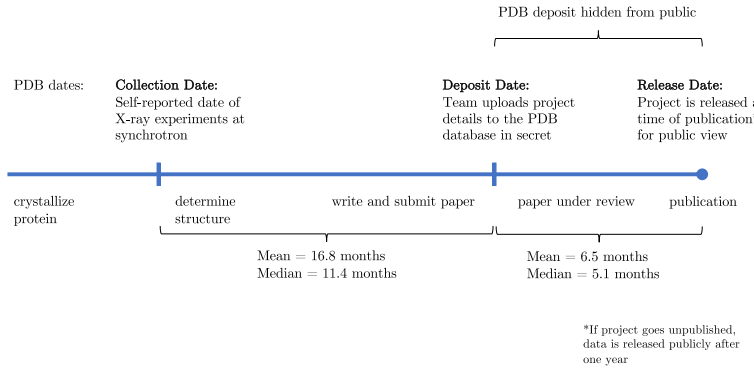

FIGURE III

PDB Timeline

This figure shows the PDB dates we observe in timeline form. Means and medians are from the full PDB sample. This figure is identical to Hill and Stein (forthcoming), figure 1.

view.[18] The time span between these 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 II). In other words, it is the time spent determining the protein's structure, refining the structure, and writing the paper.

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 II) can be the most time-consuming step. At least part of this process is devoted to improving the crystal quality, which directly influences the structure quality, and should therefore be considered part of the maturation process. Since we do not observe the date the scientist began attempting to crystallize the protein, we cannot measure this part of the process. Therefore our maturation variable does

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

not capture the full interval of time spent working on a given project. We assume the maturation that we measure is positively correlated with the true maturation time, but for this reason we interpret our maturation results more cautiously than other results.[19]

3. *Measuring Investment.* There is no clear way to measure the total resources 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 a measure of resources invested in a project. In addition, we can 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 those who directly contributed to solving the protein structure. 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.

4. *Measuring Competition.* Measuring competition directly in our data is challenging. We would ideally like to observe g, the equilibrium probability that a competitor has also started the project. Because we cannot directly measure the ex ante probability of competition, we instead measure ex post realized competition. We use an indicator for whether the protein was involved in a race for publication. We can measure this due to two features of the PDB. First, the PDB assigns each protein to a "similarity cluster" based on the protein's amino acid sequence. Two identi-

19. To be more precise, we can call unobserved time devoted to improving the crystal m_1 and the observed time spent building the model m_2. We would like to measure $m = m_1 + m_2$ but we only observe m_2. We might think that a scientist who wants to move quickly makes both m_1 and m_2 shorter—this would imply that m is certainly positively correlated with m_2. However, if spending more time improving the crystal makes it easier to subsequently build the model, then it is possible that m_1 is negatively correlated with m_2. If this negative correlation is strong enough, then m and m_2 could be negatively correlated. This possibility is why we are more cautious in interpreting the maturation results.

cal or near-identical proteins will belong to the same similarity cluster.[20] Second, the timeline measures shown in Figure III allow us to focus on proteins that are not only near-identical but are also being worked on concurrently. Following the procedure in Hill and Stein (forthcoming), we define a priority race as an instance where the winning team releases first, but the losing team had already deposited their structure at the time of release. Thus, both teams were working on the structure concurrently. This somewhat narrow definition restricts us to late-stage races. However, because the PDB releases all deposited structures by default a year after deposition, this definition ensures we do not miss any priority races due to strategic abandonment by the losing team—even if the second team abandons at this late stage, we will still see their structures.

Our priority-race proxy is a noisy estimate of g—the researcher's perceived competition—which is the relevant variable for dictating researchers' decision making and behavior. In regressions where we use this as a dependent variable—for instance, estimating the effect of potential on competition, as in Proposition 2—this measurement error does not pose an issue. However, if we want to use this competition as an independent variable—for example, estimating the effect of competition on quality—then we run into issues of attenuation bias due to measurement error. We discuss how we handle this in Section IV.E.

5. *Complexity Covariates.* In some of our regressions, we want to control for the complexity or difficulty of solving a given protein. This can help us rule out alternative explanations for why potential is correlated with lower quality, such as high-potential proteins being more difficult to solve. 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 (number of atoms in the structure), and residue count (number of amino acids the structure contains). Because these variables are heavily right-skewed,

20. More specifically, there are different "levels" of sequence similarity clusters. Two proteins belonging to the same 100% similarity cluster share 100% of their amino acids in an identical order. Two proteins belonging to the same 90% similarity cluster share 90% of their amino acids in an identical order. We use all clusters at the 50% level and higher, consistent with the scientific literature. For more detail, see Hill and Stein (forthcoming).

we take their logs. We include these three variables and their squares as complexity controls. Our results show that these size measures—while uncorrelated with potential—are strong predictor's of a protein's quality, suggesting that we are able to account for complexity quite well.[21]

6. *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.), 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 III.E.

III.C. Other Data Sources

1. *Web of Science.* The Web of Science links more than 70 million scientific publications to their respective citations.[22] Our version of these data start in 1990 and end in 2018. Broadly, we 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 US National Library of Medicine. The PDB manually links all structures to the published paper that "debuts" the structure and includes 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 use these linked data to compute citation counts for PDB-linked papers. We compute citations by counting citations in the three years following publication and we exclude any self-

21. Membrane proteins are a key exception to the discussion. 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 (Carpenter et al. 2008; Rhodes 2006). This has made membrane-protein structures a rarity in the PDB—although membrane proteins make up nearly 25% of all proteins (and an even higher share of drug targets), they make up just 1.5% of PDB structures. We drop membrane proteins from our sample, although their inclusion or exclusion do not meaningfully affect our results.

22. The Web of Science has been owned by Clarivate Analytics since 2016.

citations.[23] 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, others may have multiple affiliated structures. We discuss how we handle multiple matching of structures to a single paper in Section III.D.

2. *UniPROT Knowledgebase.* The UniPROT Knowledgebase is a database of more than 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 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 publication of the protein's structure. We view this as a measure of ex ante demand for the protein's structure.[24] 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 to proxy for a protein's importance, a topic discussed in more detail in Section III.E.

3. *DrugBank.* DrugBank is a comprehensive database containing information on drugs, their mechanisms, their interactions, and their protein targets. It is widely used by researchers, physicians, and the pharmaceutical industry (Wishart et al. 2018). The current release contains more than

23. 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.

24. This is very similar to the strategy Williams (2013) uses to measure the importance of genes.

11,000 drugs, including about 2,600 approved drugs (approved by the FDA, Health Canada, EMA, etc.), 6,000 experimental (i.e., preclinical) drugs, and about 4,000 investigational drugs (in Phase I/II/III human trials).[25] 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.

III.D. 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. From here, we make a series of sample restrictions to construct our final analysis sample. Following (Hill and Stein forthcoming), we drop a few hundred exceptionally large proteins (structures with 15 or more substructures, known as entities).[26] This leaves us with 128,270 structures. Key variables in our data are indexed at two distinct levels: the structure level and the paper level. We start by restricting to publications with just one structure. This leaves us with 35,538 structures linked to 35,538 papers (or "projects" in the case of structures without an associated publication).[27] The resulting data have a one-to-one mapping between a paper and structure. This restriction allows us to assign paper-level characteristics, such as expected citations, directly to individual structure deposits in the PDB.

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

25. Some drugs fall into more than one category.

26. Some variables are defined at the entity level, rather than at the structure level. We discuss how we aggregate entity-level variables up to the structure level in detail in Online Appendix B. These aggregation choices in some cases become more difficult when the entity count is very high, so we drop the less than 1% of structures with more than 15 entities.

27. For structures without an associated publication, we attempt to predict whether the structure would have been the only structure in a paper had it been published. See Online Appendix B for details. Online Appendix Figure E2 suggests that we are able to correctly classify these structures the majority of the time.

us with 22,127 structures. Finally, we drop structures that are missing any of our three quality measures. We also drop membrane proteins.[28] This leaves us with a final sample of 20,434 structures.

Table I provides summary statistics for the full sample and our analysis sample. Panel A presents structure-level statistics, and Panel B presents paper-level statistics. Although our analysis sample is a small subset of the total structures, it appears fairly representative of the full sample in terms of quality, publication rates, and citations. However, the maturation period is shorter in the analysis sample, likely because we focus on the first deposit of a given protein, so racing is more likely. Competition (as measured by priority racing) is more common in the analysis sample, for the same reason. Complexity is slightly lower. Finally, the number of UniPROT papers (i.e., papers published before the first structure discovery) is lower in the analysis sample, though this is somewhat mechanical, because there are more UniProt papers in more crowded clusters, and the analysis sample (by definition) only includes one structure per cluster.[29] For more detail on the full distributions of our key outcome variables, see the histograms in Online Appendix Figure E3.

III.E. Defining Project Potential

The final empirical object we must define is an analog to the project-potential variable in our model. Project potential captures the notion that ex ante, some proteins are likely to be highly cited. Scientists are usually aware of which projects, if successfully completed, will publish well and be heavily cited. This information guides their choices over which projects to pursue. For example, the COVID-19 pandemic 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

28. We drop membrane proteins because they are exceptionally difficult to purify and crystallize (Rhodes 2006; Carpenter et al. 2008). This exclusion only drops 369 structures and does not meaningfully affect our results.

29. For example, in a cluster with 100 deposits we drop 99 deposits from the analysis sample, while in a cluster with 2 deposits, we only drop 1. If the 100-deposit cluster has more UniProt papers, it will be underrepresented in the analysis sample.

TABLE I

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	0.5	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	0.11	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	1.7	0.0	30.9	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												
Priority-race indicator	0.03	0.00	0.16	0.00	1.00	0.0	0.07	0.00	0.25	0.00	1.00	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	29.9
Complexity measures												
Number of entities	1.5	1.0	3.0	1.0	91.0	0.0	1.3	1.0	0.8	1.0	14.0	0.0
Molecular weight (1,000s of Daltons)	107.1	51.9	600.1	0.3	97,730.5	0.0	95.7	55.3	421.3	1.9	47,370.7	0.0
Residue count (1,000s of amino acids)	0.8	0.5	1.5	0.0	89.2	0.0	0.7	0.5	0.9	0.0	33.1	0.0
Atom site count (1,000s of atoms)	6.5	3.4	16.4	0.0	717.8	0.0	5.5	3.6	6.7	0.1	261.5	0.0
UniProt papers	9.5	4.0	16.9	0.0	199.0	0.0	5.7	2.0	10.7	0.0	196.0	0.0
Deposition year	2009.1	2010.0	6.2	1972.0	2018.0	0.0	2008.6	2009.0	5.5	1993.0	2018.0	0.0
Total number of structures	128,876						20,434					

TABLE I
CONTINUED

	All X-ray crystallography sample						Analysis sample					
	Mean	Median	Std. dev.	Min	Max	% Missing	Mean	Median	Std. dev.	Min	Max	% Missing
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	0.70	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	16.9	9.0	29.2	0.0	811.0	39.8
Total number of papers/projects	63,806						20,434					

Notes. This table shows summary statistics for the structure-level and paper-/project-level data. We present summary statistics for 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, we impute which structures were part of the same project (see the text and Online Appendix B for details). Complexity variables (molecular weight, residue count, and atom site count) are divided by 1,000 for ease of interpretation.

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 10,000 citations in the five years since its publication (Wrapp et al. 2020; also see PDB ID 6VSB). Although 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.

Project potential is a key variable in our model, but it cannot be observed directly in the data. We estimate it among the proteins in our analysis sample. We use the 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 as "Homo sapiens" (human) or "Mus musculus" (mouse)), and a link to the gene that codes for the protein (over 2,500 different genes). We take advantage of the UniPROT prior-paper measure as an additional predictor.

We do not predict total citation counts. Instead, 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 use these detailed data to predict these citation percentiles for each structure. These predicted percentiles are the empirical analog of project potential.

In this context, the number of predictors is large (over 4,000 variables) relative to the number of observations. 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 toward zero. To remove this bias, we reestimate

an OLS regression using the LASSO-selected covariates (Belloni and Chernozhukov 2011). We use the post-LASSO coefficients to generate predicted citations.[30]

In our analysis sample of 20,434 structures, 8,128 (about 40%) do not have a three-year citation count. This happens because either the associated paper was published after 2015 (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 nonmissing citation counts.

Online Appendix Figure E4 compares actual versus predicted citation percentiles, to help assess the prediction quality. Panel A shows a histogram of actual versus predicted percentiles. Although 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 that our prediction exercise is reasonably successful. Online Appendix Table E2 shows the LASSO-selected covariates and the post-LASSO OLS coefficients. While many of the coefficients are difficult to interpret, it is reassuring to see some commonsense coefficients—for example, human proteins, along with proteins that had more prior papers written before the structure discovery, tend to be more highly cited. The R^2 from the post-LASSO OLS regression suggests that we are able to capture about 18% of the variation in actual citation percentile with our predictions.

30. In Section IV.C we discuss how structure complexity might affect our results and discuss strategies to account for it. However, it is also possible that excluding complexity controls in our LASSO prediction biases the coefficients of our citation predictors, and thus biases our predicted citations measure. To check for this, we implement an additional prediction exercise, where we include the complexity controls as unpenalized regressors in our LASSO model, to strip the other coefficients of this bias. We predict the citation percentile, but exclude the complexity variables from the prediction. The predicted values are nearly identical to the ones that we estimate in our original approach ($Corr = 0.99$) and thus, our results are virtually identical no matter which measure we use.

IV. Testing the Model: Empirical Strategy and Results

We test the predictions laid out by the model in Section II. We start by focusing on Propositions 2 and 3, which rely on cross-sectional variation in potential. Proposition 2 states that high-potential projects should generate more investment and therefore more competition. Proposition 3 states that high-potential projects should be more rushed and lower quality. We provide a variety of evidence that points to increased competition and rushing—rather than other omitted factors—as the primary channel.

Finally, we return to Proposition 1, 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 previously, we only measure ex post realized competition. This noise will lead to attenuation bias in our estimates. Therefore, we use an instrumental-variables strategy—instrumenting for competition with protein characteristics—to estimate the effect of competition on maturation and quality.

IV.A. The Relationship between Potential and Competition

Proposition 2 predicts that high-potential projects will be more competitive because researchers invest more in starting these projects. We proxy for competition using our priority-race variable, and we measure investment using the number of structure authors and paper authors.

Figure IV shows the relationship between competition and potential. We illustrate the relationship using a binned scatterplot. As Figure IV demonstrates, high-potential projects are more likely to be involved in a priority race. The highest-potential structures are in priority races over 10% of the time on average, and the lowest-potential structures are in priority races less than 6% of the time.

Table II, column (1) formalizes this relationship. For structure i deposited in year t, we estimate:

$$Y_{it} = \alpha + \beta P_i + X_i'\gamma + \tau_t + \varepsilon_{it}, \tag{5}$$

where Y is our outcome of interest (in this case, 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,

TABLE II

THE EFFECT OF POTENTIAL ON COMPETITION, MATURATION, AND QUALITY

	Competition	Maturation		Quality	
Dependent variable	Priority race (1)	Years (2)	Years (3)	Std. index (4)	Std. index (5)
Panel A: Without complexity controls					
Potential	0.0012***	−0.0063***	−0.0039	−0.0208***	−0.0153***
	(0.0002)	(0.0013)	(0.0025)	(0.0008)	(0.0015)
Principal investigator fixed effects?	No	No	Yes	No	Yes
R-squared	0.010	0.017	0.493	0.068	0.480
Panel B: With complexity controls					
Potential	0.0012***	−0.0060***	−0.0034	−0.0190***	−0.0141***
	(0.0002)	(0.0014)	(0.0026)	(0.0008)	(0.0014)
Principal investigator fixed effects?	No	No	Yes	No	Yes
R-squared	0.010	0.019	0.494	0.210	0.553
Mean of dependent variable	0.077	1.746	1.723	−0.069	−0.118
Observations	16,215	14,638	12,088	16,215	13,505

Notes. This table shows the relationship between competition/maturation quality and potential, estimating equation (5). The level of observation is a structure-paper. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include log molecular weight, log residue count, and log 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 (2) is lower because maturation is missing for a subset of observations. The number of observations in columns (3) and (5) are lower because we drop singleton-PI observations when adding PI fixed effects. 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.

* $p < .1$, ** $p < .05$, *** $p < .01$.

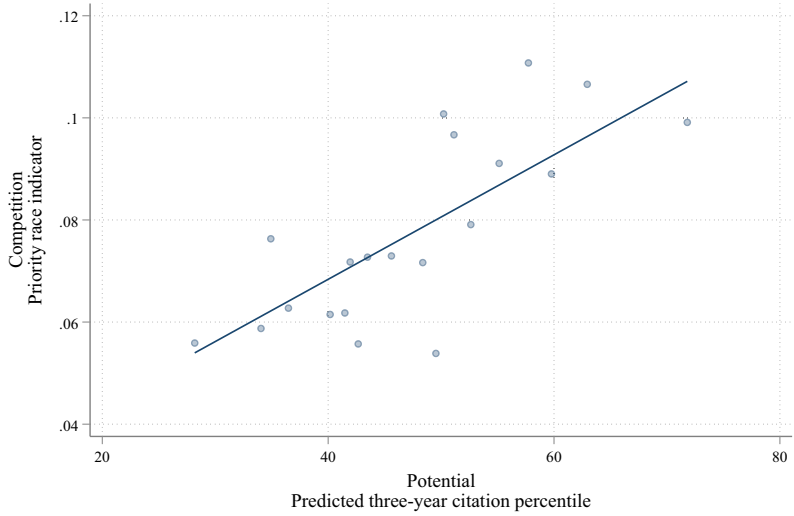

FIGURE IV

The Effect of Potential on Competition

This figure plots the relationship between potential and competition, testing Proposition 2. Potential is measured as the predicted three-year citation percentile. Competition is measured as an indicator for whether the structure was involved in a priority race. The plot is presented as a binned scatterplot (Stepner 2014). To construct this binned scatterplot, we residualize potential and competition with respect to a set of deposition-year indicators. We 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.

and ε is the idiosyncratic error term. β is the coefficient of interest because it describes the relationship between potential and our outcome of interest.[31]

31. We report heteroskedasticity-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 Online Appendix Table E3, we recompute 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 regenerate the potential variable, allowing LASSO to reselect the model. Second, we use these generated potential measures and the same sample to estimate the OLS relationship between potential and the 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,

Panel A presents the estimates of β with deposition-year fixed effects, which corresponds to the plot shown in Figure IV. In the remainder of this article, we find it convenient to benchmark effect sizes by comparing structures in the 90th percentile of the potential distribution (corresponding to structures predicted to fall in the 63rd percentile of the citation distribution, as shown in Online Appendix Figure E4, Panel A) to structures in the 10th percentile of the potential distribution (corresponding to structures predicted to fall in the 31st percentile of the citation distribution). We call these "high-potential structures" and "low-potential structures" respectively. The coefficient of 0.0012 in column (1) implies that high-potential structures have a 3.8 percentage point higher probability of being involved in a priority race.[32] Since the typical low-potential structure has a mean of around 6%, this represents a more than 60% increase. This effect is significant at the 1% level.

We also see evidence that researchers invest more in high-potential structures. Online Appendix Figure E5 is similar to Figure IV but shows the relationship between investment (as proxied by author count) and potential. The highest-potential structures have about 4.75 structure authors and 7.5 paper authors on average, and the lowest-potential structures have 4.5 structure authors and 6.5 paper authors.

Collectively, these results suggest that researchers are interested in maximizing their citations and rationally choose which projects to invest in and to 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 citations or acclaim. This provides credibility for the setup of our model, where we assume that researchers are behaving as strategic citation maximizers.

IV.B. The Relationship between Potential and Quality

Here we turn to the core predictions from our model. The first part of Proposition 3 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 3 predicts that this decrease in matura-

the bootstrapped standard errors do not differ meaningfully from those reported in the main text.

32. We calculate this by taking $0.0012 \times (63 - 31) = 0.038$.

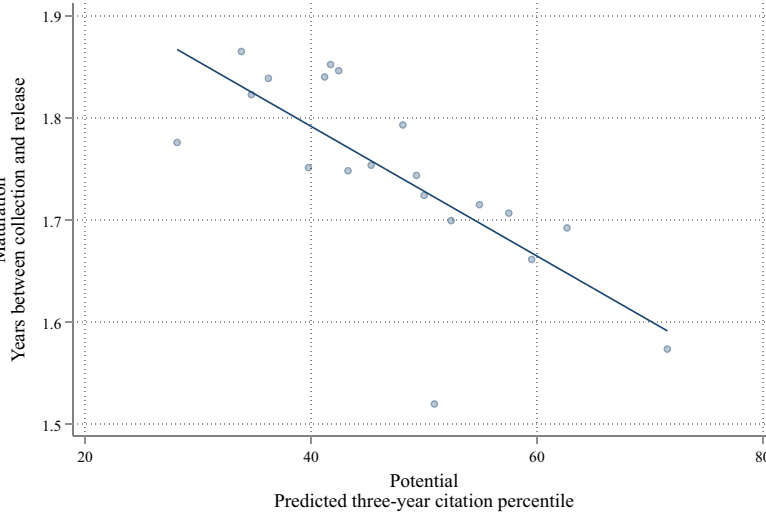

FIGURE V

The Effect of Potential on Maturation

This figure plots the relationship between potential and maturation, testing Proposition 3. 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, constructed as described in Figure IV. The sample is the full analysis sample as defined in the text, excluding SG deposits and deposits where the maturation variable is missing.

tion will lead to lower quality among the high-potential projects. Figure V shows the relationship between our maturation measure and potential, controlling for deposition year. The highest-potential projects have maturation periods of about 1.6 years, and the lowest-potential projects have maturation periods of almost 1.9 years—a difference of two to three months. Although our maturation measure is imperfect, we view this result as being consistent with our model. It suggests that at the very least, high potential is correlated with a shortening of part of the project life span.

Figure VI illustrates the relationship between potential and quality, using our quality index. We see that higher potential is associated with lower quality and that the magnitude of these correlations is notable. The highest-potential projects have resolution measures that are nearly a full standard deviation lower than the lowest-potential projects. In

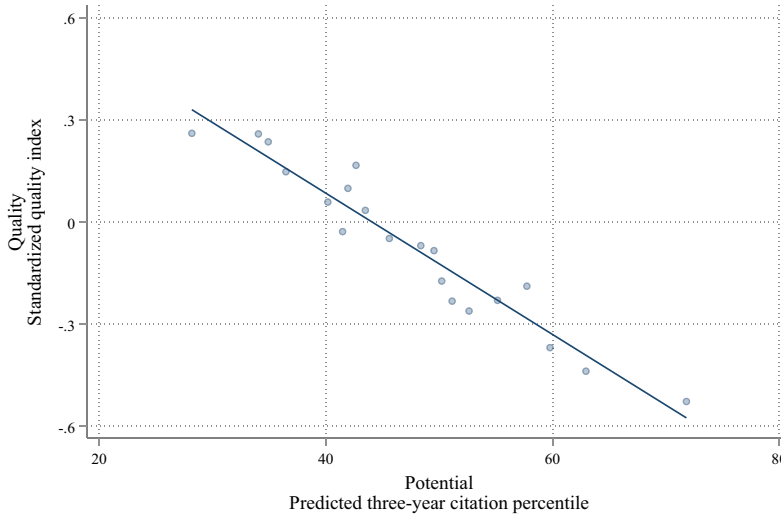

FIGURE VI

The Effect of Potential on Quality

This figure plots the relationship between potential and quality, testing Proposition 3. Potential is measured as the predicted three-year citation percentile. Quality is measured by our standardized quality index described in detail in Section III.B.1. The plot is presented as a binned scatterplot, constructed as described in Figure IV. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Online Appendix Figure E6 we show that these trends are consistent across each of the individual quality measures, with very similar magnitudes.

Table II, columns (2)–(5) presents these relationships in regression form. We estimate the same regression as in equation (5), 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 (2) shows that higher-potential projects have shorter maturation periods. The coefficient of -0.0063 implies that high-potential structures are completed about 0.20 years (about two and a half months) faster than low-potential structures. Since the typical low-potential structure has a maturation period of about 1.8 years, this represents a decline of about 11%. This effect is statistically significant at the 1% level.

Column (4) measures the effect of potential on quality. Again looking at Panel A, the coefficient of −0.0208 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 individual quality measures (see Online Appendix Table E4), and all the coefficients are statistically significant at the 1% level.

One mechanism could be that researchers who are willing to cut corners on quality sort into high-potential projects. To assess this, we add principal-investigator fixed effects to our regressions.[33] Columns (3) and (5) report the results from these regressions for maturation and quality, respectively. The signs and magnitudes are broadly similar, though the coefficient on maturation becomes statistically insignificant. Researcher sorting does not appear to explain our results. Instead, we find that the same researcher—within her portfolio of projects—executes high-potential projects more quickly and with lower quality. We also show that our results hold within journal (Online Appendix Table E5), suggesting that our results are not driven by strategic submission to journals with different quality standards.

Together, these results provide support for our model of researchers rushing in an effort to publish first. However, 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 concern motivates our work in the next two sections.

IV.C. 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 more 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 next section, we provide two distinct pieces of evidence that suggest

33. In the sciences, the last author is usually the principal investigator, so we actually use last-author fixed effects as a proxy for principal-investigator fixed effects.

that complexity alone cannot explain the negative relationship that we observe.

In general, our estimates of β in equation (5) 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. We focus on the correlation between complexity and potential because it is the most likely violation of this conditional-independence assumption.

However, a qualitative reading of the scientific literature suggests this form of omitted-variables bias may not be a critical concern. Proteins can be difficult to solve because (i) they are hard to crystallize, and (ii) once crystallized, they are hard to model. 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. The fact that crystallization is not easily predictable is reassuring, because it suggests that ease of crystallization is not correlated with easily observable protein characteristics, which in turn makes it less likely to be correlated with a protein's potential—a variable we construct from a protein's observable characteristics. Yet as a general rule, larger and "floppier" proteins are more difficult to crystallize than their smaller and more rigid counterparts (Rhodes 2006). 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.

Our next strategy is to include controls for structure complexity in an effort to achieve conditional independence. These controls, which are outlined in Section III, proxy for the size of the protein structure. Table II, Panel B illustrates the effect of adding these complexity controls in equation (5). To start, we note that these controls are powerful predictors of project quality. The R^2 dramatically increases in columns (4) and (5) with the inclusion of these controls. For example, in column (4), the R^2 increases by a factor of three (going from 0.068 in Panel A to 0.210 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 II, we observe that the coefficients remain stable. In particular, looking at our quality-index outcome in col-

umn (4), we see that complexity controls reduce the magnitude of our estimate by just 10%. Across all four quality outcomes, the coefficients remain negative and statistically significant at the 1% level (see Online Appendix Table E4).[34]

These results suggest that scientific complexity is not the main driver of the negative correlation between project potential and project quality. They reinforce our assertion that complexity, while predictive of quality, is largely uncorrelated with project potential. Instead, it appears that competition and rushing play a significant role. In a further effort to cleanly isolate the effect of competition alone, we take advantage of the fact that different researchers face different competitive incentives.

IV.D. Investigating Structural-Genomics Groups

We contrast structures deposited by SG groups and those deposited by other researchers to separate the effect of researcher rushing from other omitted factors (such as project complexity). As we will discuss, researchers in SG groups are less focused on competing for priority. These researchers will choose longer maturation periods and higher quality when working on competitive structures compared with their non-SG counterparts. This implies that the relationship between potential and quality should be flatter for SG researchers.[35] 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 the SG and non-SG samples. Therefore, the difference in slopes between the two samples is not driven by complexity but by differing levels of concern over competition.

1. *Background on Structural-Genomics Consortia.* We focus on SG groups because we argue that researchers in these groups face different competitive incentives than those of the typical aca-

34. In addition, we estimate specifications where we control for complexity nonparametrically, binning the three complexity measures into five bins and fully interacting these bins. These results are reported in Online Appendix Table E6 but are nearly identical to those reported in the main text.

35. 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, Roin, and Williams (2015).

demic lab. Since the early 2000s, SG consortia around the world have focused their efforts on solving and depositing protein structures in the PDB. In the United States, these efforts were coordinated through the Protein Structure Initiative funded by the National Institutes of Health. Inspired by the success of the Human Genome Project, SG groups have a different mission than university and private sector labs have. 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). Although the 15-year effort did not solve the structure of every known protein, SG groups have achieved a much 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 are 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." We view structures deposited by SG groups as a set of structures that were published by scientists who were not subject to the usual level of competition for priority.

We can identify SG deposits in our data by looking at the structure authors in the PDB. If the structure was solved by an SG group, the group name will be listed as the last structure author (e.g., the last author might be "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 III provides summary statistics for our analysis sample separately for non-SG structures and SG structures. SG structures make up about 20% 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

TABLE III

SUMMARY STATISTICS: NON–STRUCTURAL GENOMICS SAMPLE VERSUS STRUCTURAL-GENOMICS SAMPLE

	Non–structural genomics sample						Structural genomics 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	2.2	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	0.11	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	1.8	0.0	30.9	0.0	0.4	0.0	1.0	0.0	13.7	0.0
Maturation measures												
Years between collection and deposition	1.7	1.2	1.8	0.0	22.8	9.7	0.6	0.2	1.1	0.0	12.6	1.8
Competition measures												
Priority-race indicator	0.08	0.00	0.27	0.00	1.00	0.0	0.03	0.00	0.17	0.00	1.00	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	3.9	1.0	88.0	16.7	11.6	8.0	12.1	2.0	72.0	80.7
Complexity measures												
Number of entities	1.4	1.0	0.9	1.0	14.0	0.0	1.0	1.0	0.3	1.0	7.0	0.0
Molecular weight (1,000s of Daltons)	101.5	56.6	471.0	1.9	47,370.7	0.0	73.2	50.6	80.1	5.6	1,641.1	0.0
Residue count (1,000s of amino acids)	0.8	0.5	1.0	0.0	33.1	0.0	0.7	0.4	0.7	0.0	15.5	0.0
Atom site count (1,000s of atoms)	5.7	3.7	7.0	0.1	261.5	0.0	4.8	3.3	5.3	0.3	113.3	0.0
UniProt papers	6.7	3.0	11.5	0.0	196.0	0.0	2.2	0.0	5.5	0.0	103.0	0.0
Deposition year	2008.6	2009.0	5.9	1993.0	2018.0	0.0	2008.6	2008.0	3.9	1997.0	2018.0	0.0
Total number of structures	16,215						4,219					

TABLE III
CONTINUED

	Non–structural genomics sample						Structural genomics sample					
	Mean	Median	Std. dev.	Min	Max	% Missing	Mean	Median	Std. dev.	Min	Max	% Missing
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	0.37	0.00	1.00	0.0	0.19	0.00	0.40	0.00	1.00	0.0
Three-year citations	17.3	9.0	29.2	0.0	811.0	28.9	11.8	5.0	28.2	0.0	324.0	81.7
Total number of papers/projects	16,215						4,219					

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, in the analysis sample. Structures are grouped into a single paper based on PubMed ID. For structures with no associated PubMed ID, we impute which structures were part of the same project (see the text and Online Appendix B for details). Complexity variables (molecular weight, residue count, atom site count) are divided by 1,000 for ease of interpretation.

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 a lower probability of a priority race. Only 19% of SG deposits have an associated publication, compared with 83% 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 the mission of the SG groups, which seek to provide coverage for less studied proteins. Despite the difference in means, the potential distribution for SG and non-SG structures has substantial overlap, as shown in Online Appendix Figure E7. This suggests that we can draw reasonable comparisons between how SG and non-SG structures are affected by competition and potential.

2. *Analysis of SG Consortia.* Figure VII compares the relationship between potential and maturation for 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 series have differing support.

The level shift between the 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 in part by their very low publication rates (only 19% 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.

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

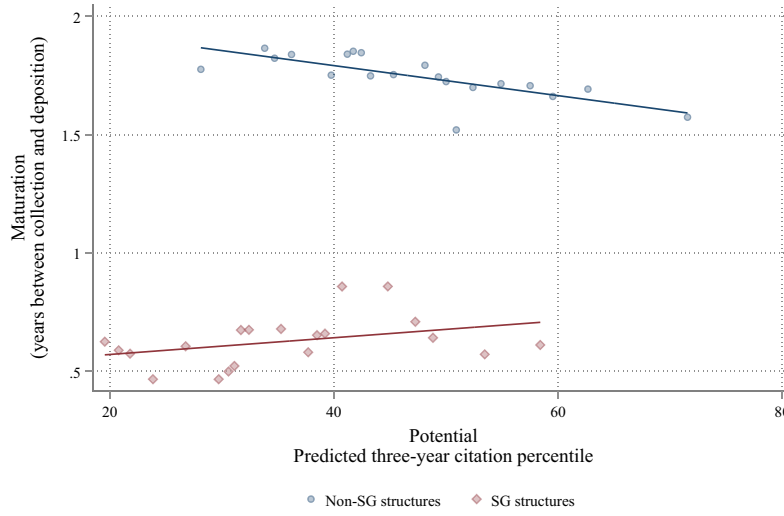

FIGURE VII

The Effect of Potential on Maturation by Structural-Genomics Status

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 plot is presented as two separate binned scatterplots, overlaid on the same axes. To construct these binned scatterplots, we residualize potential and maturation with respect to a set of deposition-year indicators (separately by SG status). We divide each 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 where the maturation variable is nonmissing.

Although our maturation measure does not capture the full maturation period, these results are suggestive.

Figure VIII is similar but presents the effects on quality. Here 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 less competitive) structures behave like their SG counterparts. It is only at high levels of potential (and high levels of competition) that the gap becomes meaningful. This pattern is consistent across the individual quality measures as well (see Online Appendix Figure E8).

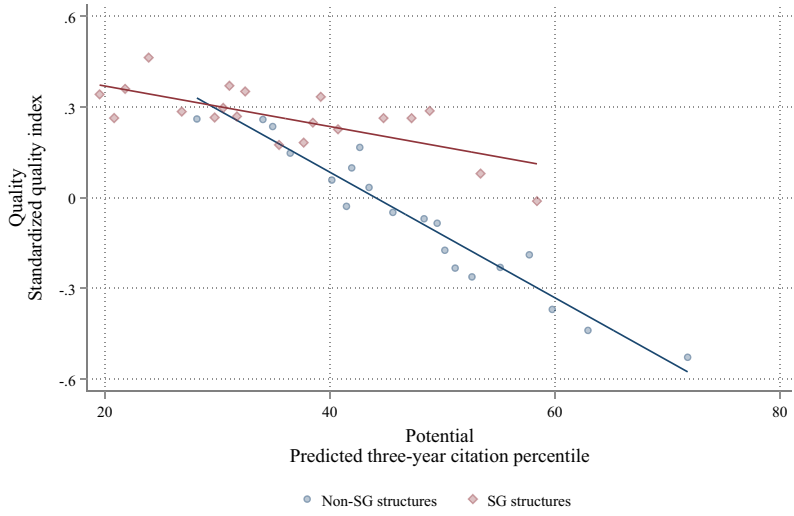

FIGURE VIII

The Effect of Potential on Quality by Structural Genomics Status

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 standardized quality index described in detail in Section III.B.1. The plot is presented as two separate binned scatterplots, overlaid on the same axes, constructed as described in Figure VII. The sample is the full analysis sample.

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

$$Y_{it} = \alpha + \beta P_i + \lambda NonSG_i + \delta(P_i \times NonSG_i) + \tau_t + X_i'\gamma + \varepsilon_{it}, \quad (6)$$

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 the outcome and potential for the SG group. λ measures the average difference in outcomes for non-SG structures relative to SG structures. δ, the coefficient of interest, measures the difference in the slope for non-SG structures relative to SG structures.

TABLE IV
THE EFFECT OF POTENTIAL ON MATURATION AND QUALITY, BY STRUCTURAL-GENOMICS STATUS

Dependent variable	Maturation Years (1)	Quality Std. index (2)
Panel A: Without complexity controls		
Potential	0.0046***	−0.0082***
	(0.0014)	(0.0011)
Non–structural genomics	1.4897***	0.2665***
	(0.0793)	(0.0524)
Potential × non–structural genomics	−0.0112***	−0.0121***
	(0.0019)	(0.0013)
R-squared	0.091	0.082
Panel B: With complexity controls		
Potential	0.0052***	−0.0069***
	(0.0014)	(0.0010)
Non–structural genomics	1.4839***	0.2665***
	(0.0796)	(0.0493)
Potential × non–structural genomics	−0.0114***	−0.0115***
	(0.0019)	(0.0012)
R-squared	0.093	0.217
Mean of dependent variable	1.499	0.000
Observations	18,779	20,434

Notes. This table shows the relationship between competition/maturation/quality and potential, estimating equation (5). The level of observation is a structure-paper. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include log molecular weight, log residue count, and log 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 (2) is lower because maturation is missing for a subset of observations. The number of observations in columns (3) and (5) are lower because we drop singleton-PI observations when adding PI fixed effects. 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.
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table IV presents the results. Focusing first on Panel A, column (1), we see that our estimate of β (the coefficient on potential) is positive and significant at the 1% level, reflecting the fact that SG groups spend longer on 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 at the 1% level. The negative estimate of the δ coefficient suggests

that the 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 the competition channel. This comparison assumes that the groups suffer from the same omitted-variables bias, 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 90th percentile of the potential distribution) to be completed over four months faster than low-potential structures (structures that fall in the 10th percentile of the potential distribution). The typical low-potential, non-SG structure has a maturation period of about 1.8 years, so this represents a meaningful (20%) reduction.

Column (2) focuses on quality. 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 the 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 (consistent with what we see in Figure VIII).

The primary coefficient of interest, δ, is negative and statistically significant at the 1% level. The estimated δ coefficient implies 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. Online Appendix Table E7 shows that the results are consistent for each of the three separate quality measures.

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, we would expect there to be no relationship between potential and quality (see Online Appendix A for more detail on the no-competition case). 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 do. Recall that they publish about 20% of their structures. This could lead to a negative but less steep slope. If this lesser (but nonzero) competition is the reason for the negative slope, 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 completely 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 work. 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 emphasize that it is likely a conservative lower bound.

IV.E. The Relationship Between Competition and Quality

Competition is the channel through which high-potential projects are ultimately executed with lower quality. This is clarified by Proposition 1, 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. As previously discussed, we cannot directly measure this risk. Instead, we measure ex post realized competition. This noisy proxy may lead to attenuated estimates of the effect of competition on quality. In addition, we might worry about omitted-variables bias in an OLS regression of quality on competition.

A natural solution to these issues is to instrument for competition. We need an instrument that is correlated with competition (first stage) but uncorrelated with quality except through its effect on competition (exclusion restriction). We use the organism that a protein originates from as our instrument. More specifically, we construct a dummy variable that equals one if the protein comes from a hu-

man (as opposed to another organism). Different species often have genetically similar proteins, but the demand for structures in human organisms might be higher because of potential medical applications of those discoveries.

We selected this instrument in a data-driven way, by running five first-stage regressions using indicators for the five most common species in our sample. Online Appendix Table E8 shows that the human indicator was the only instrument with a strong first stage. Human proteins are 3.3 percentage points more likely to be in a priority race, with an F-statistic of nearly 40. We want to test the exclusion restriction: proteins that originate from humans have different quality only because of their differing level of competition. While this is hard to show definitively, we can assess one major threat to this claim: that human proteins are more complex. In Online Appendix Table E9 we check for balance on our complexity measures for human and nonhuman proteins. Human proteins are different on average than their nonhuman counterparts, but if anything, they are less complex and the differences are small. To the extent that this biases our results, it should push us away from finding that competition causes lower quality.

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

$$Y_{it} = \alpha_1 + \beta_1 C_i + X_i' \gamma_1 + \tau_{1t} + \varepsilon_{1it}, \tag{7}$$

where Y is our outcome of interest (maturation or quality) and C is our proxy for competition, X is our vector of complexity controls, τ_1 is the deposition-year fixed effect, and ε_1 is the error term.

We also estimate an alternative specification, using 2SLS and instrumenting for competition using the human indicator. The second-stage regression for structure i deposited in year t is given by:

$$Y_{it} = \alpha_2 + \beta_2 \hat{C}_i + X_i' \gamma_2 + \tau_{2t} + \varepsilon_{2it}, \tag{8}$$

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, τ_2 is the deposition-year fixed effect, and ε_2 is the error term. β_2 is the coefficient of interest, as it measures the effect of competition on quality.

Table V shows the results from both of these specifications. Comparing the coefficients of β_1 (in Panel A) and β_2 (in Panel

TABLE V

THE EFFECT OF COMPETITION ON MATURATION AND QUALITY

Dependent variable	Maturation Years (1)	Quality Std. index (2)
Panel A: OLS		
Competition	−0.427***	−0.018
	(0.045)	(0.028)
Complexity controls?	Y	Y
Panel B: 2SLS (instrument = human)		
Competition	−5.152***	−7.061***
	(1.474)	(1.266)
Complexity controls?	Y	Y
First-stage *F* statistic	26.0	37.9
Panel C: 2SLS (instrument = potential)		
Competition	−5.657***	−15.821***
	(1.691)	(2.689)
Complexity controls?	Y	Y
First-stage *F* statistic	25.2	36.0
Mean of dependent variable	1.75	−0.07
Observations	14,638	16,215

Notes. This table shows the relationship between maturation/quality and competition. Panel A presents the results from an OLS regression, following equation (7). Panels B and C present the results from a 2SLS regression, where competition is instrumented with a human indicator and potential respectively, following equation (8). The *F*-statistic is the Montiel Olea and Pflueger (2013) robust *F*-statistic. The level of observation is a structure-paper. Competition is measured as an indicator for whether the structure was involved in a priority race. Complexity controls include log molecular weight, log residue count, and log 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.

* $p < .1$, ** $p < .05$, *** $p < .01$.

B), we see that competition is correlated with shorter maturation periods and lower quality in both specifications. However, as expected, we see that the estimates in Panel A are attenuated compared with the estimates in Panel B. The estimates in Panel B are large and represent the change in maturation or quality that arises when a structure goes from a 0% chance of a priority race to a 100% chance. This is an extreme comparison. In our data, we do not observe that level of variation. The first-stage regression suggests that our instrument increases $\hat{C}$ by 3.3 percentage points. A shift of this magnitude would translate to a maturation period that is about two months shorter and quality that is 0.2 standard deviations lower.

Our model also suggests an alternative instrument for competition: our measure of potential. The intuition for this instrument is very similar to the human instrument: the combination of protein characteristics that are used to predict potential should also predict competition. However, they should be uncorrelated with quality except through their effect on competition. It is harder to assess the validity of the exclusion restriction for this more complex instrument. We show the results of the 2SLS regression using potential as an instrument for competition in Panel C. Again, we see results that are much larger than the OLS estimates. We also find estimates in column (2) that are about double the magnitude of those in Panel B. It is worth noting that when using potential as an instrument, we generate substantially more variation in $\hat{C}$ than when using the binary human instrument. Thus, we hesitate to overinterpret these differences because they could easily arise due to any nonlinearity in the relationship between competition and quality (Angrist, Graddy, and Imbens 2000). Ultimately, we view the results in Panel C as an additional robustness check, which strengthens the argument that competition is the key causal channel for the maturation and quality effects that we observe.

IV.F. 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 low-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 to be useful for structure-based drug design, the structures must have a resolution of 2.5 Å or lower, and an R-free of 0.25 or lower.[36] Though 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 have worse quality than these thresh-

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

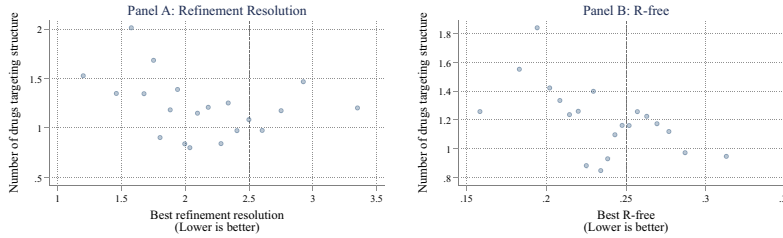

FIGURE IX

The Relationship Between Structure Quality and Drug Development

This figure plots the relationship between structure quality and the 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 binned scatterplots, constructed as described in Figure IV. The dashed lines indicate the quality thresholds for drug development proposed by Anderson (2003). The sample is the full analysis sample.

olds. About 35% of the non-SG structures in our analysis sample do not meet the resolution cutoff. About 45% of these same structures do not meet 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 hypothesized quality thresholds, 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. 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.

Figure IX, Panel A shows the relationship between drug development and resolution in a binned scatterplot.[37] Here we plot unstandardized resolution, so recall that lower values correspond to higher quality. We also plot the 2.5 Å 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 kink near the 2.5 Å cutoff. Panel B repeats this

37. If a structure has been deposited multiple times, we use resolution from the best (i.e., highest-quality) structure. The idea behind this choice is that a pharmaceutical firm would always use the best structure available. We discuss this in more detail in Section VI.A.

procedure with R-free (again, lower values of unstandardized R-free correspond to higher quality). We see a drop-off in drug development at lower quality. Again, the kink occurs near the 0.25 threshold proposed by Anderson (2003). 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 affect downstream drug development.

V. Survey Experimental Evidence

The results in the previous section document a negative correlation between project potential and project quality. We argue that the primary channel is increased competition among high-potential projects. However, these analyses have two shortcomings. First, they lack randomized variation that could best identify the causal effect of competition on quality. Second, the results are limited to a single field of science: structural biology. To address both of these concerns, we designed a large-scale survey experiment that allows us to generate our own variation (Stantcheva 2023) and explicitly test our model. We sent the survey to structural biologists and to researchers in other fields of science.

V.A. Survey Experiment Setup

The survey experiment consisted of two questions that were designed to explicitly test Propositions 1 and 2 of the model. All responses were anonymous. The first question asked about perceived competition. Participants randomly saw one of two versions of the following question (boldface highlights the randomized text):

V1: "Consider the following scenario: You are working on a project and you have generated some preliminary results. Based on the research question and your results, **you expect that it will publish in a high impact journal (such as Science, Nature, or the top journal in your field)**. How likely is it that another research team is working on a very similar project?"

V2: "Consider the following scenario: You are working on a project and you have generated some preliminary results. Based on the research question and your results, **you expect that it will publish in a medium impact field journal.** How likely is it that another research team is working on a very similar project?"

Respondents were given a slider between 0% and 100% to answer. By randomly varying potential, this question was intended to directly test the effect of potential on perceived competition (Proposition 2).

The second question asked about maturation and quality. In this case, participants randomly saw one of two versions of the following question (boldface highlights the randomized text):

V1: "Consider a different scenario: Suppose you have generated some preliminary results for a project. **You are fairly confident that nobody else is working on a very similar project (less than a 10% chance).** Answer the following questions with this scenario in mind."

- How long would it take you to complete the project and submit the paper to a journal?
- Which of the following would you do prior to submitting the paper?

V2: "Consider a different scenario: Suppose you have generated some preliminary results for a project. **You are fairly confident that another team is working on a very similar project (greater than a 90% chance).** Answer the following questions with this scenario in mind."

- How long would it take you to complete the project and submit the paper to a journal?
- Which of the following would you do prior to submitting the paper?

For the question about time until completion, respondents were given a slider between 0 and 24 months. For the question about which of the following respondents would do prior to submission, we provided six categories that were meant to capture the care and quality of the work. The options were replicate key experiments, run additional supporting experiments, perform a thorough code review, perform a thorough review of analytical

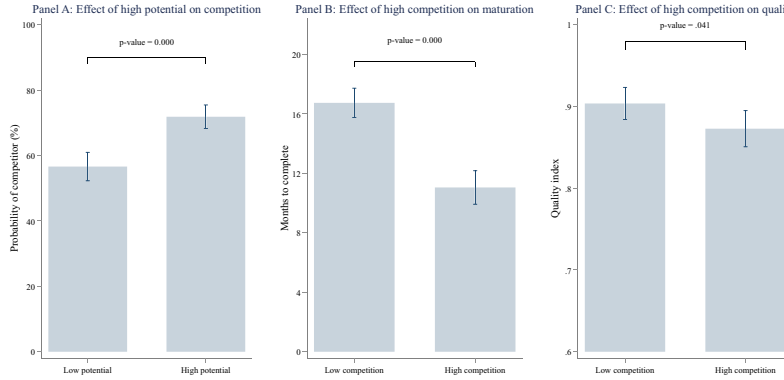

FIGURE X

Survey Experiment Results: PDB Respondents

This figure shows the results from our survey experiment of 309 active structural biologists. Details of the survey experiment can be found in Section V and Online Appendix C. In Panel A, respondents were randomly assigned to receive a prompt about a low-potential or high-potential project and asked to assess the probability of a competitor. In Panel B, respondents were randomly assigned to receive a prompt about a low-competition or high-competition project and asked how long they would spend completing the project. In Panel C, respondents were randomly assigned to receive a prompt about a low-competition or high-competition project and asked which quality control items they would complete before submitting the project.

work, perform a thorough proofread of the manuscript, and perform a thorough literature review. For each category, we asked respondents to choose between yes, maybe, no, or not applicable. In this case, by randomly varying competition, we are able to directly test the effect of competition on maturation and quality (Proposition 1). The survey instruments as they appeared to respondents can be found in Online Appendix C.

V.B. Results Within Structural Biology

We distributed the survey to just over 3,000 structural biologists who were active in the PDB. More details on the sampling process can be found in Online Appendix C. We received 309 complete responses. The results of the survey experiment are shown in Figure X. Panel A shows how potential affects perceived competition, with respondents who are told that their project is low potential believing there is a 57% chance of a priority race, whereas those who are told that their project is high potential believing

there is a 72% chance. This represents a 27% increase and is statistically significant at the 1% level.

Panel B shows how competition affects maturation. Respondents who are told their project is unlikely to be competitive report they will spend 17 months completing the project, whereas those who are told their project is likely to be competitive report they will spend 11 months (a 34% decrease, statistically significant at the 1% level). Finally, Panel C shows how competition affects quality. For each of the quality measures we ask about, we convert a yes to 1 point, a maybe to 0.5 points, and a no to 0 points. Panel C reports the average across all six measures (the "quality index"). Separate results for each individual quality measure can be found in Online Appendix Figure E9. We see that high competition is associated with a 3.4% decrease in the quality index. This decline is statistically significant at the 5% level. For the more substantive quality measures (replicating the main experiment and performing additional experiments) we see that the effects are slightly larger (between 5% and 10% decrease).

These results are consistent with our model and provide additional credibility to our observational results. Although surveys may suffer from their own issues (social-desirability bias and response bias are potential concerns in our context), the combination of revealed-preference observational data and survey experiment results paint a consistent and compelling picture.

V.C. Generalizability to Other Fields of Science

A lingering question is whether our results translate to other fields of science, or if they are specific to structural biology. This is both a conceptual question about the nature of scientific inquiry across fields and an empirical question about the consequences of racing elsewhere in science. On the conceptual side, our model can help us think about which characteristics a field would need to show similar results. For example, consider the prediction that high-potential projects are more competitive. This will only happen if there is agreement across researchers as to which projects have high potential and if researchers can accurately forecast this potential. This might be very relevant in a field with a well-defined map of unsolved problems (the protein-folding space), but less relevant for areas with a less structured search space or perhaps the search to open new lines of inquiry. For the prediction that competition leads to shorter maturation periods, we

need for there to be a meaningful penalty for losing the priority race. Otherwise, there is no cost to finishing second and no inducement to rush. Last, for the prediction that competition leads to lower quality, we require that shorter maturation periods translate directly to lower quality. If research quality comes from stochastic breakthroughs, this assumption may not hold. However, if research quality comes from the steady application of effort, this assumption seems more reasonable.

In an effort to address the question of external validity empirically, we extended our survey experiment to nine additional fields of science: cell biology, ecology, immunology, biochemistry, inorganic chemistry, condensed-matter physics, optics, and social psychology. Online Appendix C describes how we selected these fields. We obtained researcher contact information from the Web of Science using the corresponding-author information on academic papers and classified researchers into subfields using the field assignments from Microsoft Academic Graph (MAG). Unfortunately, MAG does not include structural biology as a subfield. We constructed a comparison group of structural biologists in two different ways: first, we took authors who had deposited in the PDB (closely analogous to researchers in our main analysis). Second, we took scientists who published in the MAG subfields who were most likely to link to a PDB deposit. Ultimately, we contacted nearly 100,000 researchers and received 8,237 complete survey responses.

Figure XI shows the survey experimental estimates across all fields. Panel A estimates the effect of potential on perceived competition. Although this effect is positive in all fields and statistically significant in all but one, structural biology is an outlier in terms of the magnitude. Panel B shows the effect of competition on maturation. Again, the results are consistent—negative and statistically significant across all fields. Structural biology is a small outlier, but the results are in line with other life science fields (cell biology and immunology, for example). Finally, Panel C shows the effect of competition on quality. Ten of the 11 point estimates are negative, and 9 are statistically significant. Moreover, structural biology is quite similar in terms of effect sizes to other fields.

Together, this suggests that structural biology is not unique in how its researchers respond to increased competition. However, it is an outlier in terms of how much potential effects perceived competition. This likely means that the potential-quality gradi-

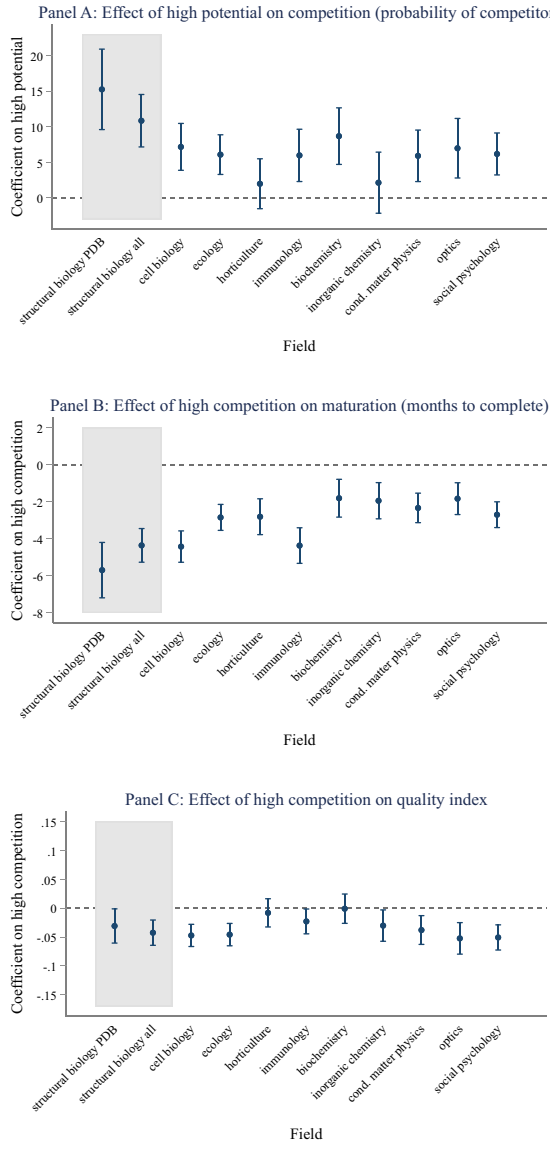

FIGURE XI

External Validity: Cross-Field Results

This figure shows the causal effects for our three survey-experiment questions. We surveyed 8,237 researchers across 10 fields of science. Each dot corresponds to an estimated causal effect, with bars representing 95% confidence intervals. Additional details of the survey experiment can be found in Section V and Online Appendix C.

ent is steeper in structural biology than in other fields. Overall, however, it appears that the forces in our model apply (at least to some extent) to many fields of science.

VI. Welfare Implications

Thus far, we have focused entirely on the positive predictions of the model. Normative conclusions are more difficult to draw. Nevertheless, we make the case that follow-on researchers cannot easily "fix" low-quality structures, and the quality effects we measure capture a real inefficiency in the generation of new scientific knowledge. We argue that this implies that there are at least two potential costs associated with racing. First, it may lead to lower-quality work, even after accounting for work that builds and improves on the original rushed work. Second, because improving low-quality work requires resinking many of the same costs, the improvement itself is costly. We try to estimate both of these costs and benchmark them relative to related scientific endeavors.

VI.A. 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 on over time. 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 samples from as many birds as possible to better understand differences among species. In 1960, he published a survey of over 5,000 proteins from over 700 different species (Sibley 1960; Strasser 2019). Suppose Sibley had been concerned that a competitor was working on a similar project, and instead released his survey a year earlier, 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. Thus, we would not consider this type of rushing inefficient.

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 2.5 Å structure. Alternatively, she could rush—spending just six months, which would yield a 3.0 Å structure. If she rushes, consider the incentives for another researcher to improve the structure from 3.0 Å to 2.5 Å. This researcher would have to start almost from scratch, especially if the first researcher had cut corners early in the process during the crystal-growing phase. The improvement would require a new crystal, and thus new experimental data and a new structural model. The second researcher would have to sink a whole year—not to mention the financial cost—to achieve the marginal 0.5 Å quality improvement. Even if the new researcher decides the improvement is worth the cost, it is inefficient. The first researcher could have achieved the 2.5 Å structure with one year of work. Instead, the combined researchers spend a year and a half to get the same quality. The key point is that—in contrast to quality-ladder models (and the naturalist example), which assume that researchers can frictionlessly build on most current work—the new researcher has to resink the same costs to generate a marginal improvement. This duplication-of-costs distortion likely applies in many experimental fields of science, where rushing early in the process may cause downstream problems that are difficult to correct.[38]

VI.B. Quantifying the Costs of Competition

Our work so far suggests at least two possible inefficiencies associated with racing. First, there is the loss of structure quality, which as Figure IX illustrates, has the potential to translate to lost downstream innovation. Second, there are costs associated

38. For example, mistakes such as a failure to correctly randomize or contamination of samples make the ultimate conclusions of a study less reliable. However, the study can only be improved by starting (nearly) from scratch. An interesting example of this phenomenon is AstraZeneca's COVID-19 vaccine clinical trial. The company accidentally gave some subjects half doses instead of full doses. The mistake likely arose from the extreme time pressure, and scientists said that the error "eroded their confidence in the reliability of the results" (Robbins and Mueller 2020). Correcting this study would require enrolling new subjects and starting from scratch.

with the duplicative effort involved in improved redeposits. We estimate both of these costs in the following sections.

1. *Computing Missing Quality.* How much quality is lost due to scientists competing to publish first? We can try to answer this question by using the SG researchers as a set of scientists who behave in a socially optimal (i.e., noncompetitive) way. In other words, we ask: "what would happen if university researchers behaved like structural-genomics researchers?" We attribute the difference in their actual behavior and their counterfactual behavior to competition. Note that this is inherently conservative: to the extent that SG researchers engage in any competitive behavior at all, we underestimate the amount of missing quality. With this caveat in mind, we use our difference-in-differences results from Table IV to impute counterfactual quality of non-SG structures in our analysis sample. Online Appendix D provides the details of this counterfactual exercise.

Figure XII visualizes this counterfactual. In blue circles, we see the initial non-SG structures with the familiar negative relationship between potential and quality. In red diamonds, we see the counterfactual quality if these same structures had been deposited by SG researchers. As we expect, the counterfactual SG quality is higher. The gap is large, at over half a standard deviation for the highest-potential structures. Thus, a substantial amount of quality is initially lost due to racing. In green triangles we plot the best version of each protein that is eventually deposited. In other words, we go protein by protein in our sample and see if it has been redeposited. If it has, we check whether the redeposited version is higher quality—if yes, we replace its initial quality with the higher quality. After accounting for these improved repeat deposits, we see that most of the gap between the initial structures and the counterfactual structures is closed.[39]

2. *The Costs of Improved Deposits.* However, this improvement is not free. These improved structures typically require researchers to re-solve the structure from scratch, which is costly. Suppose that the cost of the original (first) structure is c_1^*, which derives from the optimal choice of maturation (m^*) in our priority-

39. Comparing the slopes from the three regressions implies that the gap between initial and best structure accounts for 75% of the gap between initial and counterfactual structure.

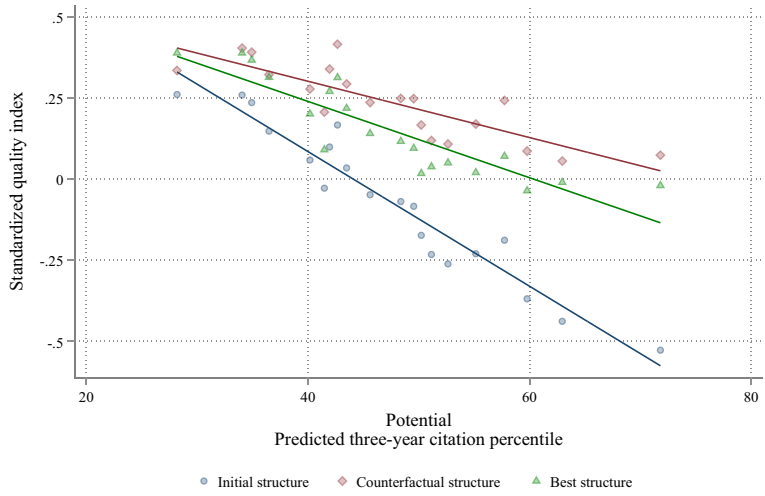

FIGURE XII

Subsequent Structure Deposits and Quality Improvement

This figure plots the relationship between potential and initial quality (in blue circles), counterfactual quality if the researchers behaved like SG researchers (red diamonds), and the best version of the structure's quality (green triangles). The details of how we compute counterfactual quality and the best quality can be found in Online Appendices D and B, respectively. Quality is measured using our quality index described in detail in Section III.B.1. The plots are presented as binned scatterplots, constructed as described in Figure IV. The sample is the full analysis sample.

race model. Further suppose that the cost of the improved redeposit is c_2. The total cost of both deposits is $c_1^* + c_2$. How does this compare to what the costs would be if a social planner dictated maturation and quality? In that case, researchers would behave as if there were no competition. We would have a single structure with quality $Q(m^{NC^*}) > Q(m^*)$ and we would incur costs c^{NC^*} (see Online Appendix A for more discussion of the no-competition case). Thus, the additional costs incurred due to racing can be written as:

$$\underbrace{(c_1^* + c_2)}_{\text{costs with competition, including redeposits}} - \underbrace{c^{NC^*}}_{\text{costs in the no-competition (socially optimal) case}} . \tag{9}$$

Can we arrive at a back-of-the-envelope calculation of these costs? Figure XII suggests that the quality of redeposits is close to the socially optimal quality. Thus, it seems reasonable to assume that these costs are also comparable, that is, that $c_2 \approx c^{NC^*}$. If this

is the case, then we simply need to estimate c_1^* in all instances where we observe a redeposit.[40] If there is no redeposit, we do not estimate any costs. If there are multiple redeposits, we multiply the costs by the number of redeposits.[41]

We surveyed the literature for estimates of c_1^*, and report our findings in Online Appendix Table E10. We primarily rely on a report commissioned by the PDB intended to estimate the economic value of the database (Sullivan, Brennan-Tonetta, and Marxen 2017). In this report, the authors estimate it would cost $100,000 per structure in 2017 to replicate the entire contents of the PDB. This translates to about $120,000 in 2024 dollars. We interpret this as meaning it would cost $120,000 to replicate each structure *at its current quality level*, meaning that $120,000 is a reasonable estimate of c_1^*. Other studies have arrived at similar or larger estimates, with older studies citing higher numbers (see Online Appendix Table E10).

To estimate the cumulative costs of improving structures, we count the number of proteins that were redeposited by scientists to improve quality and multiply by the estimated cost of re-solving a structure. In practice, defining structures that are intentional, quality-improving redeposits is a bit nuanced; we discuss our definition in more detail in Online Appendix D. Table VI shows our estimates of the total cost. We present these in a sensitivity analysis format. Along the rows of the table, we show an increasingly stringent definition of "redeposit." It appears that between 13% and 40% of the X-ray crystallography structures deposited in the PDB are redeposits, depending on the definition of redeposit.[42] Across the columns of the table, we allow the cost of redeposit to vary. Our motivation for this is twofold. First, different sources provide different estimates of the cost to repli-

40. On the other hand, if researchers can leverage insights from the first structure to save time and effort (for example, Kim 2023 highlights the role of molecular replacement—a tool that allows scientists to use similar protein structures as a starting point for model-building—in reducing the costs of solving less novel structures) then it is possible that $c_2 < c^{NC^*}$. In this case, it is possible that we overestimate the costs associated with racing. This motivates our approach of using a wide range of possible estimates of c_1^*.

41. For example, if one structure had two redeposits, the additional costs due to racing would be $(c_1^* + c_2 + c_3) - c^{NC^*}$. Assuming that $c_2 \approx c_3 \approx c^{NC^*}$, this reduces to $c_1^* + c_2$. Assuming that $c_2 > c_1^*$, we conservatively estimate this as $2c_1^*$.

42. One way to interpret this is that roughly 13%–40% of spending in structural biology could have been avoided in the absence of racing.

TABLE VI

THE COSTS OF STRUCTURE IMPROVEMENT

	Duplicate structure definition	# of structures	Cost per structure (2024 dollars) $80,000	$100,000	$120,000	$140,000	$160,000
Least restrictive	All repeated structures	54,816	$4,385,280,000	$5,481,600,000	$6,577,920,000	$7,674,240,000	$8,770,560,000
↓	All repeated, nonracing structures	54,172	$4,333,760,000	$5,417,200,000	$6,500,640,000	$7,584,080,000	$8,667,520,000
↓	All repeated, nonracing structures with some quality improvement	20,420	$1,633,600,000	$2,042,000,000	$2,450,400,000	$2,858,800,000	$3,267,200,000
Most restrictive	All repeated, nonracing structures with full quality improvement	18,963	$1,517,040,000	$1,896,300,000	$2,275,560,000	$2,654,820,000	$3,034,080,000

Notes. This table shows our estimates of the total costs (in 2024 dollars) of duplicative work done to improve the quality of protein structures. We present our estimates in a sensitivity analysis format, with varying costs per structure across columns and varying definitions of a "duplicate structure" across rows. See the Online Appendix for more details of how these different definitions of duplicate structure are constructed.

cate a structure (Online Appendix Table E10 summarizes). Second, in doing this exercise we had to make several assumptions that are challenging for us to validate. Providing a range of costs helps to capture some of this inherent uncertainty. Ultimately, using what we believe to be a conservative estimate of $120,000 per structure, we arrive at cost estimates of $2.5 to $6.6 billion. If we look at different per structure cost estimates, this range widens even further, from as little as $1.5 billion to as much as $8.8 billion.

The goal of this exercise is not to provide a precise estimate of the racing distortion—as evidenced by the wide range of estimates, we simply do not have enough information to do so. Instead, our goal is to provide an idea of the order of magnitude. We view these estimate (in the billions) as large in the context of science funding. Estimates at the midpoint of our range (between $3 and $5 billion) are on a par with the Human Genome Project, which was estimated to cost $3 billion between 1990 and 2003. Moreover, these estimates are significantly more than the cost of the Protein Structure Initiative (estimated to cost nearly $1 billion between 2000 and 2015) which gave rise to the structural-genomics consortia and contributed about 20% of all PDB structures. In other words, in the absence of this distortion, enough research funding could have been freed up to pursue significantly more science—in structural biology, or elsewhere.

3. *Additional Costs.* There are additional costs associated with racing that we do not attempt to quantify but which are worth highlighting. First, there is the time lag associated with improved deposits. Improvement is not only expensive, it can also be slow. The average time lag until a structure is improved upon is 3.1 years. The average time lag until the best version of the structure appears is 4.1 years after the initial structure is released.[43] These lags have the potential to slow down follow-on research such as drug development and to impose additional costs beyond those that we quantify. Thus, we view our $1.5 to $8.8 billion estimate of the costs of racing as conservative.

Taking a further step back, competition may affect welfare beyond just racing and quality. For example, competition may lead to overentry in promising areas, as each researcher ig-

43. We focus on single-entity structures when computing these numbers. See Online Appendix B for details.

nores the externality that she imposes on her rivals (Loury 1979; Mankiw and Whinston 1986; Hopenhayn and Squintani 2021). It may also engender a culture of secrecy that prevents the exchange of ideas and possible collaborations (Walsh and Hong 2003; Anderson et al. 2007). While these forces are beyond the scope of this article, they are important considerations for making any type of judgment about the optimal level of competition in science.

VII. Conclusion

This article 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 the most poorly executed. We find strong evidence of this negative relationship between project potential and project quality in our data, and complementary analyses suggest that competition—rather than other omitted factors—is what drives this negative relationship. Though our results are focused on structural biology, additional survey evidence suggests other fields of life science face similar competition to publish first and cut corners on quality as a result.

Subsequent work by structural biologists leads to re-solving and redepositing of low-quality but high-potential structures. Accounting for this subsequent work mostly eliminates the negative relationship between potential and quality. However, this follow-on work requires researchers to re-solve the protein structures from scratch and is therefore expensive: we estimate that it has cost the field billions of dollars to date. These costs are reminiscent of those discussed in Gans and Murray (2015), who point out that strategic effects might distort efficiency in cumulative innovation systems.[44]

44. They highlight a "salami slicing" phenomenon, where scientists try to publish in smaller and smaller units of output to claim credit for opening a new research line. These "pioneers" capture citations from subsequent work by others, but they don't need to bear the costs of replicating structures or making marginal

What are some potential policies that could help mitigate this distortion? One policy is to end priority races early. More specifically, this policy would end priority races when the first team successfully starts the project and let that team carry out the maturation phase without threat of competition by barring other teams from entering. In our model, this would lead to teams choosing the socially optimal maturation period because we have removed the distortion that arises from competition. Recall that the uncertainty in our model occurs in the entry stage, while the maturation stage is deterministic. Thus, having multiple teams competing during the entry stage can be helpful because it increases the probability that at least one team successfully starts the project. Once one team has entered the project, there is no more uncertainty, and the second team provides no additional value. 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.

In structural biology, the entry stage corresponds most closely to the crystallization phase. By contrast, building the model 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 claim exclusivity on a protein structure once they successfully crystallize it. Then they can build the structure from their experimental data, without fear of being scooped. Interestingly, in the early days of structural biology, an informal policy along these lines existed. At the time, there was a strong, community-enforced norm that if "someone else is working on [a structure], hands off" (Strasser 2019, 167). As Ramakrishnan (2018, 104) 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 parallels the policy of stopping races once the first research team has successfully entered the project. 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

quality improvements. This behavior has parallels with publishing low-quality work quickly to claim priority.

2019). Still, it is striking to note that structural biology organically developed a set of norms that alleviated the distortions associated with racing, even if those norms have not been sustained to the present. A policy along these lines might work well in fields of science where the majority of the uncertainty occurs early in the process, whereas quality is determined more deterministically later in the process.

More broadly, policies that reduce the level of competition could also help to mitigate the quality distortion that arises from racing. One concrete way to reduce competition is to reduce the priority premium, making the rewards for finishing first and second more equal. Some scientific journals have been pursuing policies along these lines, publicly committing to treating scooped papers the same as novel papers (Marder 2017; Justman 2018; PLOS Biology Staff Editors 2018). These journals often cite concerns about researchers rushing their papers to publication when announcing these policies.[45]

We stop short of making broad statements about the optimal level of competition in science or advocating for reduced competition. Even if we could perfectly measure the costs generated by the racing distortion that we study, such an analysis would almost surely be incomplete. Competition shapes the field of science in numerous ways, and other margins—while beyond the scope of this article—are likely important as well. Heightened competition likely encourages costly effort, which, given the public goods nature of science, benefits society. It may also induce positive selection of researchers, if only the top scientists enjoy the rewards. On the other hand, heightened competition may reduce poten-

45. For example, the journal *eLife* released the following statement in 2017: "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, 1)

tially productive collaborations across different labs, promoting secrecy and ultimately slowing the pace of innovation. It may influence or distort the direction of research, as argued by Bryan and Lemus (2017), or lead to excessive clustering in certain areas (Dasgupta and Maskin 1987; Hopenhayn and Squintani 2021). 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 important inputs to consider when determining how competitive science ought to be or how scientific competitions ought to be designed (Halac et al. 2017). There is growing interest in alternative and more collaborative ways of organizing science (e.g., the Protein Structure Initiative and the Human Genome Project). As emphasized by Bikard, Murray, and Gans (2015) and Gans and Murray (2015), an understanding of how credit and competition shape incentives will be critical in determining whether these cooperative organizations are successful.

Northwestern University, United States
University of California, Berkeley, United States

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

Data Availability

The data underlying this article are available in the Harvard Dataverse, https://doi.org/10.7910/DVN/KD7A8B (Hill and Stein 2025).

References

Abreu, Dilip, and Markus K. Brunnermeier, "Bubbles and Crashes," *Econometrica*, 71 (2003), 173–204. https://doi.org/10.1111/1468-0262.00393.
Akcigit, Ufuk, and Qingmin Liu, "The Role of Information in Innovation and Competition," *Journal of the European Economic Association*, 14 (2016), 828–870. https://doi.org/10.1111/jeea.12153.
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, 111 (2014), 5773–5777. https://doi.org/10.1073/pnas.1404402111.
Altman, Lawrence K., "U.S. and France End Rift on AIDS," *New York Times*, April 1, 1987.
Anderson, Amy C., "The Process of Structure-Based Drug Design," *Chemistry and Biology*, 10 (2003), 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002.
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*, 13 (2007), 437–461. https://doi.org/10.1007/s11948-007-9042-5.
Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens, "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67 (2000), 499–527. https://doi.org/10.1111/1467-937X.00141.
Azoulay, Pierre, Toby Stuart, and Yanbo Wang, "Matthew: Effect or Fable?," *Management Science*, 60 (2014), 92–109. https://doi.org/10.1287/mnsc.2013.1755.
Bai, Xiao-Chen, Greg McMullan, and Sjors H. W. Scheres, "How Cryo-EM Is Revolutionizing Structural Biology," *Trends in Biochemical Sciences*, 40 (2015), 49–57. https://doi.org/10.1016/j.tibs.2014.10.005.
Barinaga, Marcia, "The Missing Crystallography Data," *Science*, 245 (1989), 1179–1181. https://doi.org/10.1126/science.2781276.
Belloni, Alexandre, and Victor Chernozhukov, "High Dimensional Sparse Econometric Models: An Introduction," in *Inverse Problems and High-Dimensional Estimation*, Pierre Alquier, Eric Gautier, and Gilles Stoltz, eds. (Cham, Switzerland: Springer, 2011), 121–156.
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*, 28 (2000), 235–242. https://doi.org/10.1093/nar/28.1.235.
Berman, Helen M., Stephen K. Burley, Gerald J. Kleywegt, John L. Markley, Haruki Nakamura, and Sameer Velankar, "The Archiving and Dissemination of Biological Structure Data," *Current Opinion in Structural Biology*, 40 (2016), 17–22. https://doi.org/10.1016/j.sbi.2016.06.018
Bikard, Michaël, "Idea Twins: Simultaneous Discoveries as a Research Tool," *Strategic Management Journal*, 41 (2020), 1528–1543. https://doi.org/10.1002/smj.3162.
Bikard, Michaël, Fiona Murray, and Joshua Gans, "Exploring Trade-Offs in the Organization of Scientific Work: Collaboration and Scientific Reward," *Management Science*, 61 (2015), 1473–1495. https://doi.org/10.1287/mnsc.2014.2052.
Bloom, Floyd E., "Policy Change," *Science*, 281 (1998), 175. https://doi.org/10.1126/science.281.5374.175c.
Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti, "Researcher's Dilemma," *Review of Economic Studies*, 84 (2017), 969–1014. https://doi.org/10.1093/restud/rdw038.
Brown, Eric N., and S. Ramaswamy, "Quality of Protein Crystal Structures," *Acta Crystallographica Section D*, 63 (2007), 941–950. https://doi.org/10.1107/S0907444907033847.
Brünger, Axel T., "Free *R* Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures," *Nature*, 355 (1992), 472–475. https://doi.org/10.1038/355472a0.
Bryan, Kevin A., and Jorge Lemus, "The Direction of Innovation," *Journal of Economic Theory*, 172 (2017), 247–272. https://doi.org/10.1016/j.jet.2017.09.005.
Budish, Eric, Benjamin N. Roin, and Heidi Williams, "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials," *American Economic Review*, 105 (2015), 2044–2085. https://doi.org/10.1257/aer.20131176.

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*, 16 (2008), 5–11. https://doi.org/10.1016/j.str.2007.12.002.
Campbell, Philip, "New Policy for Structural Data," *Nature*, 394 (1998), 105. https://doi.org/10.1038/27971.
Carpenter, Elisabeth P., Konstantinos Beis, Alexander D. Cameron, and So Iwata, "Overcoming the Challenges of Membrane Protein Crystallography," *Current Opinion in Structural Biology*, 18 (2008), 581–586. https://doi.org/10.1016/j.sbi.2008.07.001.
Chayen, Naomi E., and Emmanuel Saridakis, "Protein Crystallization: From Purified Protein to Diffraction-Quality Crystal," *Nature Methods*, 5 (2008), 147–153. https://doi.org/10.1038/nmeth.f.203.
Cockburn, Iain, and Rebecca Henderson, "Racing to Invest? The Dynamics of Competition in Ethical Drug Discovery," *Journal of Economics and Management Strategy*, 3 (1994), 481–519. https://doi.org/10.1111/j.1430-9134.1994.00481.x.
Corum, Jonathan, and Carl Zimmer, "Bad News Wrapped in Protein: Inside the Coronavirus Genome," *New York Times*, April 3, 2020. https://www.nytimes.com/interactive/2020/04/03/science/coronavirus-genome-bad-news-wrapped-in-protein.html.
Cudney, Bob, "Protein Crystallization and Dumb Luck," *Rigaku Journal*, 16 (1999), 1.
Darwin, Charles, *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*, vol. 1, (London: John Murray, 1887).
Dasgupta, Partha, and Paul A. David, "Toward a New Economics of Science," *Research Policy*, 23 (1994), 487–521. https://doi.org/10.1016/0048-7333(94)01002-1.
Dasgupta, Partha, and Eric Maskin, "The Simple Economics of Research Portfolios," *Economic Journal*, 97 (1987), 581–595. https://doi.org/10.2307/2232925.
Dasgupta, Partha, and Joseph Stiglitz, "Uncertainty, Industrial Structure, and the Speed of R&D," *Bell Journal of Economics*, 11 (1980), 1–28. https://doi.org/10.2307/3003398.
Diamond, Arthur M., "What Is a Citation Worth?," *Journal of Human Resources*, 21 (1986), 200–215. https://doi.org/10.2307/145797.
Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole, "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, 22 (1983), 3–31. https://doi.org/10.1016/0014-2921(83)90087-9.
Furman, Jeffrey L., and Scott Stern, "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review*, 101 (2011),1933–1963. https://doi.org/10.1257/aer.101.5.1933.
Gans, Joshua, and Fiona Murray, "Credit History: The Changing Nature of Scientific Credit," in *The Changing Frontier: Rethinking Science and Innovation Policy*, Adam B. Jaffe and Benjamin F. Jones, eds. (Chicago: University of Chicago Press, 2015), 107–132.
Goodsell, David S., "Guide to Understanding PDB Data," Technical Report PDB 101, Protein Data Bank, 2019. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction.
Grabowski, Marek, Ewa Niedzialkowska, Matthew D. Zimmerman, and Wladek Minor, "The Impact of Structural Genomics: The First Quindecennial," *Journal of Structural Functional Genomics*, 17 (2016), 1–16. https://doi.org/10.1007/s10969-016-9201-5.
Grossman, Gene M., and Elhanan Helpman, "Quality Ladders in the Theory of Growth," *Review of Economic Studies*, 58 (1991), 43–61. https://doi.org/10.2307/2298044.
Hagstrom, Warren O., *The Scientific Community*, (New York: Basic Books, 1965).
———, "Competition in Science," *American Sociological Review*, 39 (1974), 1–18. https://doi.org/10.2307/2094272.

Halac, Mariana, Navin Kartik, and Qingmin Liu, "Contests for Experimentation," *Journal of Political Economy*, 125 (2017), 1523–1569. https://doi.org/10.1086/693040.
Hengel, Erin, "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review," *Economic Journal*, 132 (2022), 2951–2991. https://doi.org/10.1093/ej/ueac032.
Hill, Ryan, and Carolyn Stein, "Replication Data for: 'Race to the Bottom: Competition and Quality in Science'," (2025), Harvard Dataverse. https://doi.org/10.7910/DVN/KD7A8B.
———, "Scooped! Estimating Rewards for Priority in Science," *Journal of Political Economy*, forthcoming. https://doi.org/10.1086/733398.
Hong, Wei, and John P. Walsh, "For Money or Glory? Commercialization, Competition, and Secrecy in the Entrepreneurial University," *Sociological Quarterly*, 50 (2009), 145–171. https://doi.org/10.1111/j.1533-8525.2008.01136.x.
Hopenhayn, Hugo A., and Francesco Squintani, "Patent Rights and Innovation Disclosure," *Review of Economic Studies*, 83 (2016), 199–230. https://doi.org/10.1093/restud/rdv030.
———, "On the Direction of Innovation," *Journal of Political Economy*, 129 (2021), 1991–2022. https://doi.org/10.1086/714093.
Justman, Quincey, "'Scooping' Hurts Science and Scientists," *Cell Systems*, 7 (2018), 469–470. https://doi.org/10.1016/j.cels.2018.11.001.
Kim, Soomi, "Shortcuts to Innovation: The Use of Analogies in Knowledge Production," Columbia Business School Working Paper, 2023.
Lamb, David, and Susan M. Easton, *Multiple Discovery: The Patterns of Scientific Progress*, (Buckinghamshire: Avebury, 1984).
Lazear, Edward P., and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89 (1981), 841–864. https://doi.org/10.1086/261010.
Lee, Tom, and Louis L. Wilde, "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, 94 (1980), 429–436. https://doi.org/10.2307/1884551.
Lerner, Josh, " An Empirical Exploration of a Technology Race," *RAND Journal of Economics*, 28 (1997), 228–247. https://www.jstor.org/stable/2555803.
Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, 93 (1979), 395–410. https://doi.org/10.2307/1883165.
Mankiw, N. Gregory, and Michael D. Whinston, "Free Entry and Social Inefficiency," *RAND Journal of Economics*, 17 (1986), 48–58. https://www.jstor.org/stable/2555627.
Marder, Eve, "Scientific Publishing: Beyond Scoops to Best Practices," *eLife*, 6 (2017). https://doi.org/10.7554/eLife.30076.
Martz, Eric, Wayne Decatur, Joel L. Sussman, Michal Harel, and Eran Hodis, "Nobel Prizes for 3D Molecular Structure," Proteopedia, 2019. https://proteopedia.org/wiki/index.php/Nobel_Prizes_for_3D_Molecular_Structure.
Martz, Eric, and Eran Hodis, "Free R," Proteopedia, 2013. https://proteopedia.org/wiki/index.php/Free_R.
Merton, Robert K., "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *American Sociological Review*, 22 (1957), 635–659. https://doi.org/10.2307/2089193.
———, "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science," *Proceedings of the American Philosophical Society*, 105 (1961), 470–486.
Minor, Wladek, Zbigniew Dauter, and Mariusz Jaskolski, "A Young Person's Guide to the PDB," *Postepy Biochemii*, 62 (2016), 242–249. https://doi.org/10.18388/pb.2016_1.
Montiel Olea, José Luis, and Carolin Pflueger, "A Robust Test for Weak Instruments," *Journal of Business and Economic Statistics*, 31 (2013), 358–369. https://doi.org/10.1080/00401706.2013.806694.

Murphy, Kevin M., and Robert H. Topel, "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, 3 (1985), 370–379. https://doi.org/10.1080/07350015.1985.10509471.
Nalebuff, Barry J., and Joseph E. Stiglitz, "Prizes and Incentives: Toward a General Theory of Compensation and Competition," *Bell Journal of Economics*, 14 (1983), 21–43.
Pagan, Adrian, "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25 (1984), 221–247. https://doi.org/10.2307/2648877.
PLoS Biology Staff Editors, "The Importance of Being Second," *PLoS Biology*, 16 (2018), e2005203. https://doi.org/10.1371/journal.pbio.2005203.
Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of Polypeptide Chain Configurations," *Journal of Molecular Biology*, 7 (1963), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6.
Ramakrishnan, Venki, *Gene Machine: The Race to Decipher the Secrets of the Ribosome*, (New York: Basic Books, 2018).
Read, Randy J., Paul D. Adams, and W. Bryan Arendall III, et al., "A New Generation of Crystallographic Validation Tools for the Protein Data Bank," *Structure*, 19 (2011), 1395–1412. https://doi.org/10.1016/j.str.2011.08.006.
Reinganum, Jennifer F., "Dynamic Games of Innovation," *Journal of Economic Theory*, 25 (1981), 21–41. https://doi.org/10.1016/0022-0531(81)90015-6.
———, "The Timing of Innovation: Research, Development, and Diffusion," in *Handbook of Industrial Organization*, R. Schmalensee and R. D. Willig, eds. (Amsterdam: North-Holland, 1989), 849–908. https://doi.org/10.1016/S1573-448X(89)01017-4.
Rhodes, Gail, *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, (Cham, Switzerland: Elsevier Science and Technology, 2006).
Robbins, Rebecca, and Benjamin Mueller, "After Admitting Mistake, AstraZeneca Faces Difficult Questions About Its Vaccine, *New York Times*, November 25, 2020.https://www.nytimes.com/2020/11/25/business/coronavirus-vaccine-astrazeneca-oxford.html
Sibley, Charles G., "The Electrophoretic Patterns of Avian Egg-White Proteins as Taxonomic Characters," *Ibis*, 102 (1960), 215–284. https://doi.org/10.1111/j.1474-919X.1960.tb07114.x.
Stantcheva, Stefanie, "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible," *Annual Review of Economics*, 15 (2023), 205–234. https://doi.org/10.1146/annurev-economics-091622-010157.
Stephan, Paula E., "The Economics of Science," *Journal of Economic Literature*, 34 (1996), 1199–1235.
———, *How Economics Shapes Science*, (Cambridge, MA: Harvard University Press, 2012).
Stepner, Michael, "Binned Scatterplots: Introducing -binscatter- and Exploring Its Applications," *2014 Stata Conference*, 4 (2014). https://ideas.repec.org/c/boc/bocode/s457709.html.
Strasser, Bruno J., *Collecting Experiments*, (Chicago: 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 University, New Brunswick, NJ, 2017.
The Structural Genomics Consortium, "Mission and Philosophy," 2020. https://www.thesgc.org/about/what_is_the_sgc.
The UniProt Consortium, "UniProt: A Worldwide Hub of Protein Knowledge," *Nucleic Acids Research*, 47 (2019), D506–D515. https://doi.org/10.1093/nar/gky1049.

Thompson, Neil C., and Jeffrey M. Kuhn, "Does Winning a Patent Race Lead to More Follow-on Innovation?," *Journal of Legal Analysis*, 12 (2020), 183–220. https://doi.org/10.1093/jla/laaa001.
Thompson, Neil, and Samantha Zyontz, "Decomposing the 'Tacit Knowledge Problem': Codification of Knowledge and Access in CRISPR Gene-Editing," SSRN Working Paper, 2017. https://dx.doi.org/10.2139/ssrn.3073227.
Tiokhin, Leonid, and Maxime Derex, "Competition for Novelty Reduces Information Sampling in a Research Game—A Registered Report," *Royal Society Open Science*, 6 (2019). https://doi.org/10.1098/rsos.180934.
Tiokhin, Leonid, Minhua Yan, and Thomas Morgan, "Competition for Priority Harms the Reliability of Science, but Reforms Can Help," *Nature Human Behavior*, 5 (2021), 857–867. https://doi.org/10.1038/s41562-020-01040-1.
Tuckman, Howard P., and Jack Leahey, "What Is an Article Worth?," *Journal of Political Economy*, 83 (1975), 951–967. https://doi.org/10.1086/260371.
Walsh, John P., and Wei Hong, "Secrecy Is Increasing in Step with Competition," *Nature*, 422 (2003), 801–802. https://doi.org/10.1038/422801c.
Westbrook, John D., and Stephen K. Burley, "How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals," *Structure*, 27 (2019), 211–217. https://doi.org/10.1016/j.str.2018.11.007.
Williams, Heidi L., "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy*, 121 (2013), 1–27. https://doi.org/10.1086/669706.
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*, 46 (2018), 1074–1082. https://doi.org/10.1093/nar/gkx1037.
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*, 27 (1998), 249–284. https://doi.org/10.1146/annurev.biophys.27.1.249.
Wlodawer, Alexander, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski, "Protein Crystallography for Non-Crystallographers, or How to Get the Best (But Not More) From Published Macromolecular Structures," *FEBS Journal*, 275 (2008), 1–21. https://doi.org/10.1111/j.1742-4658.2007.06178.x.
Worldwide Protein Data Bank, "wwPDB 2013 News," 2013, https://www.wwpdb.org/news/news?year=2013#5764490799cccf749a90cdc9.
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*, 367 (2020), 1260–1263. https://doi.org/10.1126/science.abb2507.
Yong, Ed, "In Science, There Should Be a Prize for Second Place," *The Atlantic*, February 1, 2018.
Zhou, Ran, "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs," *University of Michigan Ross School of Business Working Paper*, 2023. https://ranzhuo17.github.io/files/RanZhuo_JMP_main_current.pdf.