Race to the Bottom: Competition and Quality in Science*

Ryan Hill† Carolyn Stein‡

January 29, 2024

Abstract

This paper investigates how competition to publish first and thereby establish priority impacts the quality of scientific research. We begin by developing a model where scientists decide whether and how long to work on a given project. When deciding how long to let their projects mature, scientists trade off the marginal benefit of higher quality research against the marginal risk of being pre-empted. The most important (highest potential) projects are the most competitive because they induce the most entry. Therefore, the model predicts these projects are also the most rushed and lowest quality. We test the predictions of this model in the field of structural biology using data from the Protein Data Bank (PDB), a repository for structures of large macromolecules. An important feature of the PDB is that it assigns objective measures of scientific quality to each structure. As suggested by the model, we find that structures with higher ex-ante potential generate more competition, are completed faster, and are lower quality. Consistent with the model, and with a causal interpretation of our empirical results, these relationships are mitigated when we focus on structures deposited by scientists who — by nature of their employment position — are less focused on publication and priority. We estimate that the costs associated with improving these low-quality structures are on the order of two to six billion dollars since the PDB's founding in 1971.

JEL Codes: D82, O31, O34

*We are deeply grateful to our advisors Heidi Williams, Amy Finkelstein, and Pierre Azoulay for their enthusiasm and guidance. Thomas Barden provided excellent research assistance. 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 their thoughtful comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 (Hill and Stein) and the National Institute of Aging under Grant No. T32-AG000186 (Stein). All remaining errors are our own.

†Northwestern University, ryan_hill@kellogg.northwestern.edu

‡Corresponding author. UC Berkeley, carolyn_stein@berkeley.edu. Both authors contributed equally.

1 Introduction

Credit for new ideas is the primary currency of scientific careers. Credit allows scientists to build reputations, which translate to grant funding, promotion, and prizes (Tuckman and Leahey, 1975; Diamond, 1986; 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. It is not surprising, then, 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 pre-emption 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 a third of scientists reported being moderately to very concerned about being scooped in their current work. Newer survey evidence focusing on experimental biologists (Hong and Walsh, 2009) and structural biologists more specifically (Hill and Stein, 2023) suggests that pre-emption remains common, and that 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),[2,3] to even more extreme behaviors, such as the intentional sabotage of competitors (Anderson et al., 2007). In this paper, we focus on a particular behavior: the pressure to publish quickly and pre-empt competitors may lead to "quick and dirty experiments" rather than "careful, methodical work" (Yong, 2018; Anderson et al., 2007).[4] In other words, the faster pace of research may lead to lower quality science. The goal of this paper 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 which supports a causal relationship between competition and lower-quality research rather than a spurious relationship driven by omitted factors.

[1] To name but a few examples: Isaac Newton and Gottfried Leibniz famously sparred over who should get credit as the inventor of calculus. Charles Darwin was distraught upon receiving a manuscript from Alfred Wallace, which bore an uncanny resemblance to Darwin's (yet unpublished) *On the Origin of Species* (Darwin, 1887). More recently, Robert Gallo and Luc Montagnier fought bitterly and publicly over who first discovered the HIV virus. The dispute was so acrimonious (and the research topic so important) that two national governments had to step in to broker a peace (Altman, 1987). For more examples, see Chapter 10 of Lamb and Easton (1984).

[2] Indeed, Dasgupta and David (1994) even highlight structural biology (the field we study) as an example of when researchers would intentionally delay sharing their experimental data. Since 1999, most scientific journals now require structural biologists to deposit their data at the time of publication.

[3] An 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.

[4] 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).

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 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 R&D spending does not affect the probability of success.[5] No team can visibly pull ahead, and therefore researchers compete vigorously. In our model, researchers work to develop ideas which arise exogenously. However, ideas alone cannot be published. Similar to Hopenhayn and Squintani (2016) and in particular Bobtcheff et al. (2017), ideas must be developed. The longer ideas are allowed to mature, the better quality the resulting research will be. 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 pre-empted). As a result, the threat of competition leads to lower quality projects than if the scientist knew she was working in isolation.[6]

However, in a departure from Bobtcheff et al. (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 in turn leads scientists to prematurely publish their findings. Thus, the key prediction of the model is that high-potential projects — those tackling questions that the scientific community has deemed the most important — are the projects that will also be executed with the lowest quality.

While the model provides a helpful framework, the primary contribution of this paper is to provide empirical evidence for the theoretical forces that it describes. Compared to 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). In order to test the predictions of our specific model, we require a setting which 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 to testing the core predictions of our model. Lastly, we require measures

[5] 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.

[6] Tiokhin et al. (2020) develop a model of a similar spirit, where researchers choose a specific dimension of quality — the sample size. Studies with larger sample sizes take longer to complete, and so more competition leads to smaller sample sizes and less reliable science. Tiokhin and Derex (2019) test this hypothesis in a lab experiment.

of the quality of scientific work. By quality, we mean quality of execution — not a measure of the paper's interest or importance.[7]

We make progress on all four of these challenges in the field of 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-on work (Berman et al., 2000; Strasser, 2019). This rich setting satisfies the four criteria outlined above. First, in structural biology, research projects center around using 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 (Kim, 2023), and progress is timestamped. Together, this allows us to identify competitive proteins: structures that are identical and being worked on contemporaneously.[8] 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 that the researcher would observe before starting her project, and would inform her view of its potential. Therefore, we construct a measure of potential by using these characteristics to predict the number of citations that the structure will ultimately receive. Lastly, every macromolecular structure is scored on a variety of quality metrics. At a high level, structural biologists are concerned with fitting three-dimensional structure models to experimental data, and so these quality metrics are measures of goodness-of-fit. They allow us to compare quality across different projects in an objective, science-based manner. To give an example of one of our quality metrics, consider refinement resolution, which measures the distance between crystal lattice planes. Nothing about this measure is subjective, nor can it be manipulated by the researcher. Figure 1 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 90^{th} versus 10^{th} percentile of the potential distribution, we find that high-potential projects induce meaningfully more competition. High-potential structures are 4 percentage points (60 percent) 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 then look at how project potential impacts maturation and quality. We find that high-potential structures are completed over two months faster, and have quality measures that are about 0.7 standard deviations lower than low-potential structures. These results echo recent findings by a pair of structural biologists (Brown and Ramaswamy, 2007), who show that structures published in top general interest journals tend to be of lower quality than structures published in less prominent field journals.[9]

[7] Some studies (Hengel, 2022) have used text analysis to measure a paper's readability as a proxy for paper quality, but such writing-based metrics fail to measure the underlying scientific content. Another strategy might be to use citations, but this fails to disentangle the quality of the project from the importance of the topic or the prominence of the author (Azoulay et al., 2013) — a distinction which is critical for our research question.

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

[9] Brown and Ramaswamy propose multiple reasons why this might be the case. One is increased competition

However, 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 use an alternative measure of potential: whether the protein originates from a human as opposed to another organism. Human proteins are significantly more competitive, and this simpler measure allows us to probe the omitted variables bias issue more carefully. In fact, we find that human structures are slightly less complex than non-human structures on average. Yet, we find that they are also over 0.2 standard deviations lower in quality than their non-human counterparts.

Lastly, we leverage another source of variation — namely, whether the protein was deposited by a structural genomics group. The majority of PDB structures are deposited by university- or industry-based scientists, both of 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 make future structure discovery easier (The Structural Genomics Consortium, 2020; Zhou, 2023). Qualitative evidence suggests these groups are less focused on publication and priority, which is consistent with the fact that only about 20 percent of SG structures ever appear in journal publications, compared to over 80 percent of non-SG structures. Because the SG groups are less motivated by competition, we can contrast the relationships between potential and quality for SG structures versus non-SG structures. If complexity is correlated with potential, then this should be the case for both the SG and non-SG structures. Intuitively, by comparing the slopes across both groups, we can "net out" the potential omitted variables bias. Consistent with competition acting as the causal channel, we find more negative relationships between potential and quality among non-SG (i.e., more competitive) structures.

Finally, we turn 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 individual, lab-based researchers to their SG counterparts, we can estimate the welfare costs that arise from racing.[10] 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

among structures published in top journals. Another 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.

[10] We thank one of the referees for this excellent suggestion.

an entirely new experiment. Thus, 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 projections of $100,000 per protein structure, we estimate that researchers have spent between $1.9 and $5.5 billion in an effort to improve low-quality structures generated by racing behavior since the PDB's founding in 1971.

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

2 A Model of Competition and Quality in Scientific Research

The idea that competition for priority drives researchers to rush and cut corners in their work is perhaps intuitive. Our goal in this section is to develop a model that both 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.[11] We allow projects to differ in terms of their 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, but direct interested readers to a more formal treatment in Appendix A.

2.1 Preliminaries

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

Timing, Investment, and Maturation. Time is continuous and indexed by t. From the perspective of each scientist, the model consists of two stages. In the first stage, scientist i has an idea. We denote the moment the idea arrives as the start time, or t_i^S. However, the scientist must pay an upfront cost in order to pursue the idea. At t_i^S, scientist i must decide how much to invest in starting the project. If she invests I_i, she has probability $g(I_i) \in [0,1]$ 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 generate preliminary results. In our setting, a natural interpretation is that I represents the time and resources spent trying to grow a protein crystal.

[11] This is consistent with views put forth by Merton (1957) and Stephan (2012), though it stands in contrast with the idea that scientists are purely motivated by the intrinsic satisfaction derived from "puzzle-solving" (Hagstrom, 1965).

The second stage occurs if the scientist successfully starts the project.[12] Then, she must decide how long to work on the project before publicly disclosing her findings. Let m_i denote the time she spends on the project, or the "maturation period." The project is then complete at $t_i^F = t_i^S + m_i$.

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

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

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

We make no restrictions on these weights, other than to specify that they are both positive and $\overline{\theta} \geq \underline{\theta}$. Importantly, we do not assume that the race is winner-take-all (i.e., $\underline{\theta} = 0$), as is common in the theoretical patent and priority race literature (for example, Loury (1979); Fudenberg et al. (1983); Bobtcheff et al. (2017)). Rather, consistent with empirical work on priority races (Hill and Stein, 2023) and anecdotal evidence (Ramakrishnan, 2018), we allow for the second-place team to share some of the credit.

Information Structure. The competing scientists have limited information about their competitor's progress in the race. Scientist i does not observe I_j, and so she doesn't know the probability her opponent enters, although she will have correct beliefs about this probability in equilibrium. In addition, she does not know her competitor's start time t_j^S. We assume that she believes that it is uniformly distributed around her own start time. In other words, she believes

[12] Note that prior to 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.

that $t_j^S \sim \text{Unif}\left[t_i^S - \Delta, t_i^S + \Delta\right]$ for some $\Delta > 0$.[13] Appendix Figure A1 summarizes the model setup.

2.2 The Maturation Decision

We work backwards, first 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 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 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 (a) the difference between $\overline{\theta}$ and $\underline{\theta}$ is large (priority rewards are more lopsided), (b) 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).

2.3 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 Appendix A that there is a unique and symmetric pure strategy Nash equilibrium for investment.

2.4 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.

[13] 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).

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 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 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 will take to the data in Section 4.

3 Structural Biology and the Protein Data Bank

This section provides some scientific background on structural biology and describes our data. We take particular care to explain how we map key variables from our model into measurable objects in our data. Our empirical work focuses on structural biology precisely because there is such a clean link between our theoretical model and our empirical setting. Section 3.1 provides an overview of the field of structural biology, while sections 3.2 and 3.3 describe our datasets. Section 3.4 describes how we construct our primary analysis sample and provides summary statistics. Appendix B provides additional detail on our data sources and construction.

3.1 Structural Biology

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

Proteins are composed of building blocks called amino acids. These amino acids are arranged into a single chain, which folds up onto itself, creating a three-dimensional structure. While the shape of these proteins is of great interest to researchers, the proteins themselves are too small to

observe directly under a microscope.[14] Therefore, structural biologists use experimental data to propose three-dimensional models of the protein shape to better understand biological function.

Structural biology has several unique features that make it amenable for our purposes, but it is also an important field of science. Proteins contribute to nearly every process inside the body, and understanding the shape and structure of proteins is critical to understanding how they function. Moreover, many heritable diseases — such as sickle-cell anemia, Alzheimer's disease, and Huntington's disease — are the direct result of protein mis-folding. Protein structures also play a critical role in drug development and vaccine design (Westbrook and Burley, 2018).[15] Over a dozen Nobel prizes have been awarded for advances in the field (Martz et al., 2019).

3.1.1 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 the three-dimensional 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. Moreover, projects include timestamps indicating when particular milestones were reached in the process. The combination of these two features allows us to identify proteins that are (a) identical or near-identical and (b) 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 4.1.

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 will discuss these measures in the subsequent sections, but we want to highlight the importance of this feature: it is difficult to imagine how one might objectively rank the quality (distinct from the importance or relevance) of papers in other fields, such as economics or mathematics. Our empirical work hinges on the fact that structural biologists have developed unbiased, science-based measures of structure quality.

3.1.2 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

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

[15] 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).

called x-ray crystallography. The vast majority (89 percent) of structures are solved using this method.

X-ray crystallography broadly consists of three steps (see Figure 2). Individual proteins are too small to analyze or observe directly. Therefore, as a first step, the scientist must distill a concentrated solution of the protein into orderly crystals. Growing these crystals is a slow and difficult process, often described as "more art than science" (Rhodes, 2006) or at times simply "dumb luck" (Cudney, 1999). Success typically comes from trial and error.[16]

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

Finally, the scientist will use these diffraction patterns to first build an electron density map, and then an initial atomic model. Building the atomic model is an iterative process: the scientist will compare simulated diffraction data from her model to her actual experimental data and adjust the model until she is satisfied with the goodness of fit. This process is known as "refinement," and depending on the complexity of the structure can take an experienced crystallographer anywhere from hours to weeks to complete. Refinement can be a "tedious" process (Strasser, 2019), and involves "scrupulous commitment to the iterative improvement and interpretation of the electron density maps" (Minor et al., 2016). 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.

3.2 The Protein Data Bank

Our primary data source is the Protein Data Bank (PDB). The PDB is a worldwide repository of biological macromolecules, 95 percent of which are proteins.[17] It was established in 1971 with just seven entries, and today contains upwards of 150,000 structures. Its goal is to promote the dissemination and further use of protein structures, both by structural biologists and by scientists in other fields.[18] Since the late 1990s, the vast majority of journals and funding agencies have

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

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

[18] 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 (2021) in the context of plasmid repositories.

required that scientists deposit their findings in the PDB (Barinaga, 1989; Berman et al., 2000, 2016; Strasser, 2019). Therefore, the PDB represents a near-universe of macromolecule structure discoveries. Below, we describe the data collected by the PDB. The primary unit of observation in the PDB is a structure, representing a single protein. Most variables in our data are indexed at the structure level.[19]

3.2.1 Measuring Quality

The PDB provides several measures intended to assess quality. These quality 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, thereby increasing overall user confidence, and it makes the PDB more useful and accessible for scientists who do not possess the specialized knowledge to critically evaluate structure quality. Below, we describe the three measures that we use in our empirical analysis. We selected these three because they are scientifically distinct and have good coverage in our data. We also combine these three measures into a single quality index, described below. Together, these measures map 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 all either calculated or independently validated by the PDB, leaving no scope for misreporting or manipulation by authors. Since 2013, the PDB has required that x-ray structures undergo automatic validation reports prior to deposition. These reports take the researcher's proposed model and experimental data as inputs, and use a suite of software programs to produce and validate various quality measures. In 2014, the PDB ran the same validation reports retrospectively on all structures that were already in the PDB (Worldwide Protein Data Bank, 2013), so we have full historical coverage for these quality measures. Appendix Figure E1 provides a snapshot from one of these reports.

Refinement resolution. Refinement resolution measures the smallest distance between crystal lattice planes that can be detected in the diffraction pattern. It is somewhat analogous to resolution in a photograph. Resolution is measured in angstroms (Å), which is a unit of length equal to 10^{-10} meters. Smaller resolution values are better, because they imply that the diffraction data is more detailed. This in turn allows for better electron density maps, as shown in Figure 1. At resolutions less than 1.5Å, individual atoms can be resolved and structures have almost no errors. At resolutions greater than 4Å, individual atomic coordinates are meaningless and only secondary structures can be determined. Scientists can improve resolution by spending time improving the quality of the protein crystals and by fine-tuning the experimental conditions during x-ray exposure. In our main analysis, we will standardize refinement resolution so that the units are in standard deviations and higher values represent better quality.

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

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

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

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

Quality index. Finally, we combine the three measures above into a single quality index. All three measures are correlated, with correlation coefficients in the 0.4 to 0.6 range (see Appendix Table E1). We create the index by adding all three standardized quality measures and then 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 Appendix.

3.2.2 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, i.e., shortening the maturation period. In most scientific fields, it would be impossible to measure the time researchers spend on each project, but the PDB metadata provides unique insight about project timelines.

As shown in Figure 3, the PDB collects two key dates which allow us to infer the maturation

period: the collection date and the deposition date. The collection date is self-reported and it corresponds to the date that the scientist subjected her crystal to x-rays and collected her experimental data. The deposition date corresponds to the date that the scientist deposited (i.e., uploaded) her structure to the PDB. Because journals require evidence of deposition before publishing articles, the deposition date corresponds roughly to when the scientist submitted her paper for peer review.[20] The timespan between these two dates represents the time it takes the scientist to go from the raw diffraction data to a completed draft (the "diffraction pattern" stage to the "completed structure" stage in Figure 2). In other words, it is the time spent determining the protein's structure, refining the structure, and writing the paper.

However, note that this maturation period only includes time spent working on the structure once the protein was successfully crystallized and taken to a synchrotron. Anecdotally, crystallizing the protein (the first step in Figure 2) 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. However, 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 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.[21]

3.2.3 Measuring Investment

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

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

[21] 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.

the text or performing complementary analyses.

3.2.4 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. Since 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 are able to 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 identical or near-identical proteins will both belong to the same similarity cluster.[22] Second, the timeline measures shown in Figure 3 allow us to focus on proteins that are not only near-identical, but are also being worked on concurrently. Following the procedure described in Hill and Stein (2023), 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 researcher 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 will run into issues of attenuation bias due to measurement error. We discuss how we handle this in Section 4.6.

3.2.5 Complexity Covariates

Proteins can be difficult to solve because (a) they are hard to crystallize, and (b) once crystallized, they are hard to model. In general, predicting whether a protein will be easy or hard to crystallize is a difficult task. Researchers have failed to discover obvious correlations between crystallization conditions and protein structure or family (Chayen and Saridakis, 2008). Often, a single amino acid can be the difference between a structure that forms nice, orderly crystals and one that evades all crystallization efforts. The fact that crystallization is not easily predictable bodes well for us, because it suggests that it is not correlated with easily observable protein characteristics, which in turn makes it less likely to be correlated with a protein's potential.

However, as a general rule, larger and "floppier" proteins are more difficult to crystallize than

[22] More specifically, there are different "levels" of sequence similarity clusters. Two proteins belonging to the same 100 percent similarity cluster share 100 percent of their amino acids in an identical order. Two proteins belonging to the same 90 percent similarity cluster share 90 percent of their amino acids in an identical order. We use all clusters at the 50 percent level and above, consistent with the scientific literature. For more detail, see Hill and Stein (2023).

their smaller and more rigid counterparts (Rhodes, 2006). Moreover, since these larger proteins are more complex, with more folds, they are harder to model once the experimental data are in hand. Therefore, despite the general uncertainty of protein crystallization, size is a predictor of difficulty. The PDB contains several measures of structure size, which we use as covariates to control for complexity. These include molecular weight (the structure's weight), atom site count (the number of atoms in the structure), and residue count (the number of amino acids the structure contains). Because these variables are heavily right-skewed, we take their logs. We then include these three variables and their squares as complexity controls. 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. [23]

3.2.6 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.), the taxonomy (organism the structure comes from), and the gene that expresses the protein. We use these detailed classification variables to estimate a protein's scientific relevance, a topic discussed in more detail in Section 4.1.

3.3 Other Data Sources

3.3.1 Web of Science

The Web of Science links over 70 million scientific publications to their respective citations.[24] Our version of these data start in 1990 and end in 2018. Broadly, we are able to link the Web of Science citations data to the PDB using PubMed identifiers, which are unique IDs assigned to research papers in the medical and life sciences by the United States National Library of Medicine. The PDB manually links all structures to the published paper that "debuts" the structure, and includes the PubMed ID in this linkage. The Web of Science includes a paper-PubMed ID crosswalk. This allows us to link the Web of Science to the PDB.

We then use these linked data to compute citation counts for PDB linked papers. We compute citations by counting citations in the three years following publication[25] and exclude any self-citations. 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

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

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

[25] 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.

citations. Note that these citation variables are unique at the *paper* level, rather than at the structure level. Structures are linked to papers in a many-to-one fashion. In other words, while some papers only have one affiliated structure, other papers may have multiple affiliated structures. We discuss how we handle multiple matching of structures to a single paper in Section 3.4.

3.3.2 UniPROT Knowledgebase

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

Scientists often study and publish papers about proteins long before their structures are solved. Therefore, we can count the number of papers that were published about a protein *prior* to the protein's structure publication. We view this as a measure of ex-ante demand for the protein's structure.[26] In other words, if a protein is heavily studied before anyone has solved and released its structure, there is probably more interest in the structure. We use this to help proxy for a protein's importance, a topic discussed in more detail in Section 4.1.

3.3.3 DrugBank

DrugBank is a comprehensive database containing information on both drugs, their mechanisms, their interactions, and their protein targets. It is widely used by researchers, physicians, and the pharmaceutical industry (Wishart et al., 2018). The current release contains over 11,000 drugs, including about 2,600 approved drugs (approved by the FDA, Health Canada, EMA, etc.), 6,000 experimental (i.e., pre-clinical) drugs, and about 4,000 investigational drugs (in Phase I/II/III human trials).[27] 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.

3.4 Sample Construction

We begin with the full sample of 128,876 PDB structures that were deposited and solved using x-ray crystallography between 1971 and 2018. These structures are linked to 63,809 unique publications. From here, we make a series of sample restrictions to construct our final analysis sample. Following (Hill and Stein, 2023) we drop a few hundred exceptionally large proteins (structures with 15 or more sub-structures, known as entities).[28] This leaves us with 128,270 structures. Key variables in our

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

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

[28] 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 Appendix B. These aggregation choices in some cases

data are indexed at two distinct levels: the structure level and the paper level. Therefore, we start by restricting to publications with just one structure. This leaves us with 35,541 structures linked to 35,541 papers (or "projects" in the case of structures without an associated publication).[29] The resulting data have a one-to-one mapping between a given paper and structure. This restriction allows us to assign paper-level characteristics, such as expected citations, directly to individual structure deposits in the PDB.

Because we are interested in the behavior of scientists who are potentially racing, we further restrict our analysis sample to new structure discoveries. In other words, we drop PDB deposits if a structure of the protein had previously been deposited. In practice, we use the similarity clusters and only keep the first protein to be released in each cluster. This leaves us with 22,128 structures. Finally, we drop structures that are missing any of our three quality measures. We also drop membrane proteins.[30] This leaves us with a final sample of 20,435 structures.

Table 1 provides summary statistics for both the full sample and our analysis sample. Panel A presents structure-level statistics and Panel B presents paper-level statistics. Although our analysis sample comprises a small subset of the total structures, it appears fairly representative of the full sample 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, and 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 prior to 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.[31] For more detail on the full distributions of our key outcome variables, see the histograms in Appendix Figure E3.

4 Testing the Model: Empirical Strategy and Results

In this section, we test the predictions laid out by the model in Section 2. 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 therefore be more rushed and lower quality. We provide a variety of evidence which points to increased competition and rushing — rather than other omitted factors — as the primary channel.

Finally, we return to Proposition 1, which states that more competitive projects (projects at

become more difficult when the entity count is very high, so we drop the less than one percent of structures with over 15 entities.

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

[30] 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 impact our results.

[31] 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 under-represented in the analysis sample.

higher risk of having multiple teams competing simultaneously) are more likely to be rushed and lower quality. We do not have a clean measure of ex-ante competition — as discussed in Section 3.2.4, we only measure ex-post realized competition. This noise will lead to attenuation bias in our estimates. However, the model sets up a natural instrumental variables specification: we can instrument for competition with project potential. Proposition 2 functions as the first stage, while Proposition 3 is the reduced form.

4.1 Defining Project Potential

Before we can begin testing the model, we must define an empirical analog to the project potential variable in our model. Project potential captures the notion that ex-ante, some proteins are likely to be 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 which began in 2019 spurred a sudden and large interest in a particular virus and its associated proteins (Corum and Zimmer, 2020). The scientists who successfully determined the structures of these key proteins were ex-ante likely to publish in the top science journals and receive high levels of citations, acclaim, and publicity — indeed, the first structure-paper pair to describe the structure of the SARS-CoV-2 viral spike protein has received over 9,000 citations in the roughly four years since publication (Wrapp et al., 2020; also see PDB ID 6VSB). While not all important proteins are related to a specific disease, many other features of proteins are predictive of the ex-ante demand for their structure.

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

This kind of prediction is possible due to extremely detailed data describing and categorizing every structure in the PDB. Each structure is given a detailed classification (over 500 different classifications, such as "transcription protein" or "signaling protein"), a taxonomy (over 1,000 different organisms, such "homo sapiens" (human) or "mus musculus" (mouse)), and a link to the gene which codes for the protein (over 2,500 different genes). We also take advantage of the UniPROT prior paper measure (described in Section 3.3.2) 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 then 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. Therefore, to avoid overfitting, we implement Least Absolute Shrinkage and Selection Operator (LASSO) to select predictors in a data-driven manner. LASSO regularization helps avoid overfitting, but it also shrinks the fitted coefficients towards zero. To remove this bias, we re-estimate an ordinary least squares regression using the LASSO-selected covariates (Belloni and

Chernozhukov, 2011). We then use the post-LASSO coefficients to generate predicted citations.[32]

In our analysis sample of 20,435 structures, 8,129 (about 40 percent) do not have a three-year citation count. This happens because either the associated paper was published after 2015 (since our citation data only runs through 2018), or because the structure has no associated paper. Rather than drop these observations, we use the LASSO coefficients to impute the predicted citation percentiles, just as we do for the observations with non-missing citation counts.

Figure E4 compares actual versus predicted citation percentiles, to help assess the prediction quality. Panel A shows a histogram of actual versus predicted percentiles. While the predicted values are more clustered toward the middle percentiles, we are able to generate fairly good dispersion. Panel B shows the binned scatterplot of actual percentiles on the y-axis versus predicted percentiles on the x-axis. The fit along the $y = x$ line appears quite good throughout the distribution. Taken together, these figures suggest our prediction exercise is reasonably successful. Appendix Table E2 shows the LASSO-selected covariates and the post-LASSO ordinary least squares coefficients. While many of the coefficients are difficult to interpret, it is reassuring to see some common-sense coefficients — for example, 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 ordinary least squares regression suggests that we are able to capture about 18 percent of the variation in actual citation percentile with our predictions.

4.2 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 discussed in Section 3.2.4. We measure investment using the number of structure authors and paper authors, as discussed in Section 3.2.3.

Figure 4 shows the relationship between competition and potential. We illustrate the relationship using a binned scatterplot. As Figure 4 demonstrates, high-potential projects are more likely to be involved in a priority race. The highest-potential structures are in priority races over 10 percent of the time on average, while the lowest-potential structures are in priority races less than 6 percent of the time.

Column (1) of Table 2 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

[32] In Section 4.4 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 then 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.

effect, 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.[33]

Panel A presents the estimates of β with deposition year fixed effects, which corresponds to the plot shown in Figure 4. Throughout the remainder of this paper, we will find it convenient to benchmark effect sizes by comparing structures in the 90^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 63^{rd} percentile of the citation distribution, as shown in Panel A of Appendix Figure E4) to structures in the 10^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 31^{st} percentile of the citation distribution). We will term 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.[34] Since the typical low-potential structure has a mean of 6%, this represents over a 60 percent increase. This effect is significant at the one percent level.

We also see evidence that researchers invest more in high-potential structures. Appendix Figure E5 is similar to Figure 4, 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, while 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 pursue with citations in mind. In other words, it does *not* appear that researchers simply choose topics they are interested in, with no regard for the citations or acclaim their work will garner. This provides credibility for the setup of our model, where we assume that researchers are behaving as strategic citation-maximizers.

4.3 The Relationship between Potential and Quality

In this section, we turn to the core predictions from our model. The first part of Proposition 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 maturation will lead to lower quality among the high-potential projects. Figure 5 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, while the lowest-potential projects have maturation periods of nearly 1.9 years — a difference of about three months. While our maturation measure is imperfect, as discussed in Section 3.2.2, we view

[33] We report heteroskedacity-robust standard errors. However, as argued by Pagan (1984) and Murphy and Topel (1985), because our measure of potential is a generated (i.e., estimated) regressor, OLS standard errors will be too small. In Appendix Table E3, we re-compute the standard errors using a two-step bootstrap procedure. First, we randomly draw from our sample with replacement, creating a new sample with the same number of observations as the original sample. We use this new sample to re-generate our potential variable, allowing LASSO to re-select the model. Second, we use these generated potential measures and the same sample to estimate the OLS relationship between potential and our dependent variable. We repeat this procedure 200 times. The standard deviation in the sample of 200 coefficient estimates is our bootstrapped standard error. In practice, the boostrapped standard errors do not differ meaningfully from those reported in the main text.

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

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 lifespan.

Figure 6 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. Moreover, in Appendix Figure E6 we show that these trends are consistent across each of the individual quality measures, with very similar magnitudes.

Columns (2) to (5) of Table 2 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 -0.0064 implies that high-potential structures are completed about 0.20 years (or 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 percent. This effect is statistically significant at the one percent 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 Appendix Table E4), and all the coefficients are statistically significant at the one percent 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.[35] Columns (3) and (5) report the results from these regressions for maturation and quality respectively. The signs and magnitudes are broadly unchanged, though the coefficient on maturation becomes statistically insignificant. Researcher sorting does not appear to explain our results. Rather, 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 (Appendix Table E5), suggesting that our results are not driven by strategic submission to journals with different quality standards.

Taken 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 following two sections.

[35] 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.

4.4 Competition or Complexity?

Our model suggests that the negative relationship we document between potential and quality is caused by scientists rushing. However, an alternative explanation is that high-potential proteins might be more complex and therefore 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 following section, we provide two distinct pieces of evidence which together suggest 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. Therefore, our next strategy is to include controls for structure complexity, in an effort to achieve conditional independence. These controls, which are outlined in Section 3.2.6, proxy for the size of the protein structure. While it is generally difficult for researchers to anticipate which structures will be difficult to solve, larger structures tend to be more challenging. These two facts lead us to believe that there are few (if any) protein characteristics which are correlated with complexity and observed by researchers, but not by the econometrician.

Panel B of Table 2 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 2, we observe that the coefficients remain stable. In particular, looking at our quality index outcome in column (4), we see that complexity controls reduce the magnitude of our estimate by just ten percent. Across all four quality outcomes, the coefficients remain negative and statistically significant at the one percent level (see Appendix Table E4).

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

4.5 Investigating Structural Genomics Groups

In this section, we contrast structures deposited by structural genomics (SG) groups and those deposited by other researchers, in order to separate the effect of researcher rushing from other omitted factors (such as project complexity). As we discuss below, researchers in SG groups are less focused on competing for priority. Therefore, these researchers will choose longer maturation periods and higher quality when working on competitive structures as compared to their non-SG

counterparts. This in turn implies that the relationship between potential and quality should be flatter for SG researchers.[36] Comparing the SG and non-SG structures is helpful, because it allows us to "net out" potential omitted variables bias. Intuitively, if we are concerned that the negative relationship between potential and quality is driven by structure complexity, that concern likely applies to both the SG and non-SG samples. Therefore, the *difference* in slopes between the two samples is not driven by complexity, but rather by differing levels of concern over competition.

4.5.1 Background on Structural Genomics Consortia

We focus on structural genomics (SG) groups because we argue that researchers in these groups face different competitive incentives than the typical academic lab. Since the early 2000s, SG consortia around the world have focused their efforts on solving and depositing protein structures in the PDB. In the US, these efforts were coordinated through the Protein Structure Initiative funded by the National Institutes of Health (NIH). Inspired by the success of the Human Genome Project, SG groups have a different mission than university and private-sector labs. These groups focus on achieving comprehensive coverage of the protein folding space, and eventually full coverage of the human "proteome," the catalog of all human proteins (Grabowski et al., 2016). 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 their university counterparts. Indeed, the vast majority of structures solved by structural genomics groups are never published, suggesting that researchers in these groups are focused on data dissemination rather than priority. For example, The Structural Genomics Consortium (an SG center based in Canada and the United Kingdom) describes its primary aim as "to advance science and [be] less influenced by personal, institutional or commercial gain." Therefore, we view structures deposited by SG groups as a set of structures which were published by scientists who were not subject to the usual level of competition for priority.

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

Table 3 provides summary statistics for our analysis sample separately for non-SG structures and SG structures. SG structures comprise about 20 percent of the analysis sample. The two groups differ in several ways. The SG deposits appear to be higher quality (lower refinement resolution, R-free, and Ramachandran outliers, all of which correspond to higher quality). However, these deposits also appear to be less complex. They have fewer entities, and lower molecular weight,

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

residue count, and atom site count — all of which point to these structures being smaller and simpler to solve than their non-SG counterparts. SG structures are completed more quickly, and have more authors. In line with their stated mission, the SG structures appear to be less studied, with fewer UniPROT papers and a lower probability of a priority race. Only 19 percent of SG deposits have an associated publication, compared with 83 percent of non-SG deposits. When they do publish, they receive fewer citations.

Given these facts, it is not surprising that SG structures are lower potential on average. This is in line with mission of the SG groups, which seek to provide coverage for less-studied proteins. However, despite the difference in means, the potential distribution for SG and non-SG structures has substantial overlap as shown in Appendix Figure E7. This suggests that we can draw reasonable comparisons between how SG and non-SG structures are impacted by competition and potential.

4.5.2 Analysis of Structural Genomics Consortia

Figure 7 compares the relationship between potential and maturation for both SG and non-SG structures. The two binned scatterplots are constructed separately and overlaid on the same set of axes. Because we bin each series separately, there are the same number of observations in each marker within the same series (but not across series). The fact that the markers do not line up vertically over the x-axis reflects the fact that the two series have different supports.

The level shift between the two groups is immediately apparent: at all levels of potential, SG structures have shorter maturation periods. The difference is over a full year on average. This gap is consistent with the mission of the SG groups, and is likely driven in part by their very low publication rates (only 19 percent of SG structures have an associated publication). These groups endeavor to get their results into the scientific domain as quickly as possible, and often do not write or release a paper to accompany the structure. Non-SG scientists, on the other hand, typically do not deposit their structures until they have a draft manuscript ready to submit.

However, the key takeaway from Figure 7 is that there is also a visible difference in slopes. As previously illustrated, the higher-potential non-SG structures are have shorter maturation periods (are completed more quickly). By contrast, the higher-potential SG structures appear to have have slightly *longer* maturation periods. While our maturation measure does not capture the full maturation period, these results are suggestive.

Figure 8 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 therefore less competitive) structures behave like their SG counterparts. It is only at high levels of potential (and therefore high levels of competition) that the gap becomes meaningful. This pattern is consistent across the individual quality measures as well (see Appendix Figure E8).

We formalize the trends shown in Figures 7 and 8 using a difference-in-differences framework.

For structure i deposited in year t, we estimate the following regression:

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

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 4 presents the results. Focusing first on column (1) of Panel A, we see that our estimate of β (the coefficient on potential) is positive and significant at the one percent 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 one percent level. The negative estimate of the δ coefficient suggests that relationship between potential and maturation is more negative for non-SG structures relative to SG structures. In fact, it is large enough to more than offset β, implying that non-SG researchers spend less time on high-potential structures, in contrast with their SG counterparts.

If we believe that our estimates of β are contaminated by omitted variables bias, then the difference in the slopes between the SG structures ($\beta + \delta$) and the non-SG structures (β) yields the causal effect of potential via the competition channel. This comparison assumes that both groups suffer from the same omitted variables bias, and so it is "netted out" when we take the difference. Interpreting δ in this way implies that competition causes high-potential structures (structures that fall in the 90^{th} percentile of the potential distribution) to be completed over four months faster than low-potential structures (structures that fall in the 10^{th} percentile of the potential distribution). Recall that the average non-SG structure has a maturation period of about 1.8 years, so this represents a meaningful (20 percent) 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 8).

The primary coefficient of interest, δ, is negative and statistically significant at the one percent 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. Appendix Table E6 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, then we would expect there to be no relationship between potential and quality (see 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. Recall that they do publish about 20 percent of their structures. This could lead to negative but less steep slope. If this lesser (but non-zero) competition is the reason for the negative slope, then the effect of potential on quality due to competition in the non-SG group would be $\beta + \delta$ — in other words, we would not want to net out β.

Alternatively, SG researchers may be 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 play. The fact that maturation is positively correlated with potential in the SG groups suggests that there may indeed be a correlation between unobserved complexity and potential. We view δ as our preferred estimate, but emphasize that it is likely a conservative lower bound.

4.6 The Relationship between Competition and Quality

Competition is the channel by which high-potential projects are ultimately executed with lower quality. This is clarified by Proposition 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. We cannot measure this risk, as discussed in Section 3.2.4. Instead, we measure ex-post realized competition. This noisy proxy may lead to attenuated estimates of the effect of competition on quality.

However, the model also suggests a solution: we can instrument for competition using project potential. Empirically, we have already demonstrated that there is a first stage (Section 4.2) and a reduced form (Section 4.3). This is enough to tell us that the relationship between competition and quality must be negative. Still, it is informative to recover the magnitudes. For example, if we want to consider policies that reduce the level of competition in science, then it is useful to know the magnitude of the expected quality response.

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

$$Y_{it} = \alpha + \beta C_i + X_i'\gamma + \tau_t + \varepsilon_{it} \tag{7}$$

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

However, we also estimate a separate specification, using two-stage least squares and instrumenting for competition using project potential. The first stage regression is identical to Equation 5, with competition (measured by our priority race indicator) as the dependent variable. The second stage regression for structure i deposited in year t is given by:

$$Y_{it} = \tilde{\alpha} + \tilde{\beta}\hat{C}_i + X_i'\tilde{\gamma} + \tilde{\tau}_t + \nu_{it} \tag{8}$$

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, $\tilde{\tau}$ is the deposition year fixed effect, and ν is the error term. $\tilde{\beta}$ is the coefficient of interest, as it measures the causal effect of competition on quality. The exclusion restriction in this case is that project potential only affects project quality (or maturation) through its impact on competition, conditional on controls. In other words, potential is not correlated with unobserved factors that impact quality directly once we condition on X. Our results in Section 4.4 and 4.5 help bolster this case.

Table 5 shows the results from both of these specifications. Comparing the coefficients of β (in Panel A) and $\tilde{\beta}$ (in Panel B), we see that competition is correlated with shorter maturation periods and lower quality in both specifications. However, as expected, we see that the estimates in Panel A are attenuated compared to 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 zero percent chance of a priority race to a one hundred percent chance. This is an extreme comparison. In our data, we do not observe that level of variation. A more reasonable way to interpret these coefficients (which also requires less out-of-sample extrapolation) is to consider the actual distribution of $\hat{C}$. A protein in the 10th percentile of the competition distribution has a 4.6 percent chance of being in a priority race, whereas a protein in the 90th percentile has an 11.5 percent chance — roughly a 7 percentage point difference. Thus, proteins in the 90th percentile of the predicted competition distribution have maturation periods that are about five months shorter and quality scores that are about 1.1 standard deviations lower than those in the 10th percentile.

Because our potential measure is comprised of many different predictors, readers may still be concerned that some of the inputs into our potential measure correlated with complexity, leading to a possible failure of the exclusion restriction. We have tried to address this concern in the sections above by controlling for complexity and using the SG groups as a comparison group. However, as a further check, we take one input into our potential measure — whether the protein comes from a human — and use it as an instrument on its own. The advantage of this simpler instrument is that we can more easily probe the exclusion restriction.

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. Appendix Table E7 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 then want to demonstrate that proteins which 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. Thus, in Appendix Table E8 we check for balance on our

complexity measures for human and non-human proteins. Human proteins are different on average than their non-human counterparts, but if anything, they are less complex and the differences are small. To the extent that this biases is our results, it should push us *away* from finding that competition causes lower quality.

Appendix Table E9 shows the reduced form results, which suggest that human proteins are completed about 2 months faster and are about 0.2 standard deviations lower in quality. Panel C of Table 5 shows the two-stage least squares results, after scaling for the shift in competition. Compared to Panel B, the maturation results are very similar. The quality results are the same order of magnitude, but about half the size. In this specification, we have even less variation in predicted competition, in part because the instrument is binary. The first stage, while very strong, only gives us a 3.3 percentage point shift in the probability of a priority race. Thus, we hesitate to over-interpret these differences as they could easily arise due to any non-linearity in the relationship between competition and quality Angrist et al. (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.

4.7 Benchmarking the Quality Estimates

Are the negative quality effects we estimate large enough to matter for overall scientific productivity in our setting? Rushing leads to lower quality structures, but are these structures low enough quality to prevent researchers from drawing useful conclusions or using the structure in follow-on work? According to structural biologists, the answer depends on what the researcher wishes to do with the structure. If the researcher simply wants to understand the protein's function, a 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 in order to be useful for structure-based drug design, the structures must have a resolution of 2.5Å or lower, and an R-free of 0.25 or lower.[37] While these cutoffs may not be hard-and-fast, they tell us something about the usefulness of a structure given its quality. It is not uncommon for structures to have worse quality than these thresholds. About 35 percent of the non-SG structures in our analysis sample do not meet the resolution cutoff. About 45 percent 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.

Panel A of Figure 9 shows the relationship between drug development and resolution in a binned

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

scatterplot.[38] Here we plot unstandardized resolution, so recall that lower values correspond to higher quality. We also plot the 2.5Å cutoff for reference. There is a clear positive relationship between higher levels of drug development and lower (i.e., better) resolution. The relationship is nonlinear, with a kink near the 2.5Å cutoff. Panel B repeats this procedure with R-free (again, lower values unstandardized R-free correspond to higher quality). We again see a 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 impact downstream drug development.

4.8 Generalizability to Other Fields of Science

We conclude this section by considering external validity. Do researchers in other fields of science also cut corners on quality in order to publish first? Or is this phenomenon unique to structural biology? This question is difficult for us to answer, because as we discuss in Section 3, structural biology has uniquely rich project-level data in the PDB, which includes measures of project quality. To the best of our knowledge, it is not feasible to run similar analyses in other fields of science.

However, in an effort to address the question of external validity, we ran a large-scale survey across ten fields of science. 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. Thus, we constructed a comparison group of structural biologists in two different ways: first, we took authors who had deposited in the PDB during the 2017-18 time period. Second, we took scientists who published in the MAG subfields that were most likely to link to a PDB deposit. More details of the survey design are available in Appendix C. Ultimately, we contacted nearly 100,000 researchers and received 7,882 complete survey responses.

We asked these researchers two questions. First, "how would you rate the competition to publish first in your field (none / mild / moderate / intense)?" And second, "in general, do you feel that peers in your field ever sacrifice the quality of their research to publish first (never / rarely / some of the time / most of the time)?" We coded responses on a zero to three point scale. Figure 11 shows the results. For the first question about competition, structural biologists score their field between 2.1 and 2.2, depending on which definition of structural biologists we use. This corresponds most closely to "moderate" on our scale. While these values are higher than some other fields, they are roughly in line with the other life science subfields (cell biology, immunology, and biochemistry). Condensed matter physics also reports high levels of competition. Structural biologists also report an average score of 1.9 for the second question about sacrificing quality, corresponding most closely

[38] 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 5.1.

to "some of the time" on our scale. Again, this is higher than average but in line with the other life science subfields we surveyed.

Taken together, these results suggest that neither the overall level of competition nor the incentive that this competition creates to decrease quality is unique to structural biology. Other subfields of the life sciences, such as cell biology, immunology, and biochemistry all report similar answers. Therefore, it seems likely that competition reduces the quality of scientific research in many areas of science, especially those in the life sciences.

5 Welfare Implications

Thus far, we have been focused entirely on the positive predictions of the model. Normative conclusions are more difficult to draw. Nevertheless, in the first part of this section, we make the case that follow-on researchers cannot easily "fix" low-quality structures, and so the quality effects we measure capture a real inefficiency in the generation of new scientific knowledge. We argue that this implies 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 upon the original rushed work. And second, because improving low-quality work requires re-sinking many of the same costs, the improvement itself is costly. In the second part of this section, we try to estimate both of these costs. Finally, we discuss an alternative policy that might mitigate the incentive to cut corners in order to publish first.

5.1 Will Follow-On Work Fix the Problem?

Even if the quality effects we measure are meaningful, is the rush to publish and the subsequent lower-quality work necessarily bad for science? Society values speed of disclosure as well as quality, in part because the quality of a discovery might be improved upon over time. Therefore, in certain circumstances, a rushed low-quality discovery might be preferable to a higher-quality breakthrough that takes longer to develop. The overall costs and benefits of rushing depends in part on the knowledge production model. If science progresses like a quality ladder, where each researcher can build frictionlessly on existing work (Grossman and Helpman, 1991), then quick-and-dirty work is likely not bad for science. To fix ideas, consider the example of ornithologist and molecular biologist Charles Sibley. In 1958, he began collecting egg samples from as many birds as possible in order to better understand the differences between species. In 1960, he published a survey of over 5,000 proteins from over 700 different species (Sibley, 1960; Strasser, 2019). Now, suppose Sibley had been concerned that a competitor was working on a similar project, and instead released his survey a year earlier, 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 2.5Å structure. Alternatively, she could rush —

spending just six months, which would yield a 3.0Å structure. If she rushes, consider the incentives for another researcher to improve the structure from 3.0Å to 2.5Å. This researcher would have to start nearly 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 an entire year — not to mention the financial cost — to achieve the marginal 0.5Å quality improvement. Even if the new researcher decides the improvement is worth the cost, it is inefficient. The first researcher could have achieved the 2.5Å structure with 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 above), which assume that researchers can frictionlessly build on most current work — the new researcher has to re-sink the same costs in order to generate a marginal improvement. 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.[39]

5.2 Quantifying the Costs of Competition

Our work so far suggests that there are at least two possible inefficiencies associated with racing. First, there is the loss of structure quality, which as Figure 9 illustrates, has the potential to translate to lost downstream innovation. Second, as discussed in Section 5.1 above, there are costs associated with the duplicative effort involved in improved re-deposits. We will estimate both of these costs in the following sections.

5.2.1 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 structural genomics researchers as a set of scientists who behave in a socially optimal (i.e., non-competitive) way. In other words, we ask: "what would happen if university researchers behaved like structural genomics researchers?" We then attribute the difference in their actual behavior and their counterfactual behavior to competition. Note that this is inherently conservative: to the extent that structural genomics researchers engage in any competitive behavior at all, we will underestimate the amount of missing quality. With this caveat in mind, we use our difference-in-differences results from Table 4 to impute counterfactual quality of non-SG structures in our analysis sample. Appendix D provides the details of this counterfactual exercise.

Figure 10 visualizes this counterfactual. In blue, we see the initial non-SG structures with the familiar negative relationship between potential and quality. In red, we see the counterfactual quality if these same structures had been deposited by SG researchers. As we expect, the counterfactual

[39]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 trail. 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.

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. However, in green 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 re-deposited. If it has, we check whether the re-deposited 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.[40]

5.2.2 The Costs of Improved Deposits

However, this improvement itself is not free. These improved structures typically require researchers to re-solve the structure, which the PDB estimates costs about $100,000 on average (Sullivan et al., 2017). Other studies have arrived at similar estimates, with older studies citing higher numbers (Stevens, 2003). To estimate the cumulative costs of improving structures, we count the number of proteins that were re-deposited 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 re-deposits is a bit nuanced; we discuss our definition in more detail in Appendix D. Table 6 shows our estimates of the total cost. We present these in a sensitivity analysis format: down the rows of the table, we show an increasingly stringent definition of "re-deposit." Across the columns of the table, we allow the cost of re-deposit to vary (different sources cite different estimates of the costs per structure, and these costs have been falling over time). Ultimately, it appears that between 13 and 40 percent of the x-ray crystallography structures deposited in the PDB are re-deposits, depending on the definition of re-deposit. This translates to a cost of $1.9 to $5.5 billion, using the $100,000 per structure estimate. In the context of science funding, this is a large number. It is similar to the cost of the entire Human Genome Project (estimated to cost $3 billion between 1990 and 2003). It is significantly more than the cost 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 percent of all PDB structures.

5.2.3 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, but 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.[41] These lags have the potential to slow down follow-on research such as drug development, and impose additional costs beyond those that we quantify above. Thus, we view our $1.9 to $5.5 billion estimate of the costs of racing as conservative.

[40] 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.

[41] We focus on single-entity structures when computing these numbers. See Appendix B for details.

Taking a further step back, competition may affect welfare beyond just racing and quality. For example, competition may lead to over-entry in promising areas, as each researcher ignores the externality she imposes on her rivals (Loury, 1979; Mankiw and Whinston, 1986; Hopenhayn and Squintani, 2021). It may also engender a culture of secrecy which 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 paper, they are important considerations for making any type of judgment about the optimal level of competition in science.

5.3 Policy and History

We conclude this section with a brief discussion of policy. In particular, we want to highlight one specific policy that might alleviate the rushing distortion: ending 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. This would lead to teams choosing the socially optimal maturation period because we have removed the distortion that arises from competition (see Appendix A for more detail on this point).

We highlight this one policy for two reasons. First, because it works well due to the somewhat specific nature of our model. And second, because such a policy was informally implemented in structural biology's early days. Taken together, we think this lends credence to our modeling choices. Recall that the uncertainty in our model occurs in the investment stage, while the maturation stage is deterministic. Thus, having two teams competing during the investment stage can be helpful, because it increases the probability that at least one team successfully starts the project. But once one team has entered the project, there is no more uncertainty, and so the second team creates no additional value. Yet, despite the model-specific nature of this policy, we highlight it because it is relevant in structural biology — so relevant in fact, that an informal policy along these lines once existed in the field.

As discussed in Section 3.1, when solving protein structures, the most difficult and risky part of the process is growing the protein crystal. Researchers may try to crystallize a protein under a variety of conditions and fail to generate a usable crystal. Therefore, growing the crystal is analogous in many ways to the investment stage of the model. By contrast, building the atomic model from the diffraction data is a more deterministic process, akin to the maturation phase. Therefore, the analog of ending priority races early in this setting would be to let researchers 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.

In fact, in the early days of structural biology, there was a strong, community-enforced norm that if "someone else is working on [a structure] — hands off" (Strasser, 2019). As Ramakrishnan (2018) explains, scientists would announce (often through publication) that they had successfully crystallized a protein, and "there was a tradition that if someone had produced crystals of something, they were usually left alone to solve the problem." This norm parallels the policy of stopping races once the first research team has successfully entered the project. However, as the field grew and

the number of unsolved structures dwindled, this precedent became too difficult to enforce. Today structural biologists are secretive about what they are working on, knowing that the "hands off" rule no longer applies (Strasser, 2019). Still, it is interesting to note that structural biology organically developed a set of norms which alleviated the problem of rushing and associated lower quality work, even if those norms have not been sustained to the present day.

6 Conclusion

This paper documents that in the field of structural biology, competition to publish first and claim priority causes researchers to release their work prematurely, leading to lower quality science. We explore the implications of this fact in a model where scientists choose which projects to work on, and how long to let them mature. Our model clarifies that because important problems in science are more crowded and competitive, perversely it is exactly these important projects that will be 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. While our results are focused on structural biology, additional survey evidence suggests other fields of life science face similar competition to publish first and feel that their peers also cut corners on quality as a result.

Subsequent work by structural biologists leads to re-solving and re-depositing 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 between $1.9 and $5.5 billion to date. Therefore, the low quality that results from competition has a large cost.

Despite this, we stop short of making broad statements about the optimal level of competition in science. Even if we could perfectly measure the costs generated by the racing distortion 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 paper — 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 potentially 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 also 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

(for example, the Protein Structure Initiative and the Human Genome Project). As emphasized by Bikard et al. (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.

References

Abreu, Dilip and Markus K. Brunnermeier, "Bubbles and Crashes," *Econometrica*, 2003, *71* (1).

Akcigit, Ufuk and Qingmin Liu, "The Role of Information in Innovation and Competition," *Journal of the European Economic Association*, 2016, *14* (4).

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

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

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

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

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," *The Review of Economic Studies*, 2000, *67*.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang, "Matthew: Effect or Fable?," *Management Science*, 2013, *60* (1), 92–109.

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

Barinaga, Marcia, "The Missing Crystallography Data," *Science*, 1989, *245* (4923), 1179.

Belloni, Alexandre and Victor Chernozhukov, "High Dimensional Sparse Econometric Models: An Introduction," in Pierre Alquier, Eric Gautier, and Gilles Stoltz, eds., *Inverse Problems and High-Dimensional Estimation*, Vol. 203 2011.

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

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

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

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

_ , Fiona Murray, and Joshua Gans, "Exploring Trade-offs in the Organization of Scientific Work: Collaboration and Scientific Reward," *Management Science*, 2015, *61* (7).

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

Bobtcheff, Catherine, Jérôme Bolte, and Thomas Mariotti, "Researcher's Dilemma," *The Review of Economic Studies*, 2017, *84* (3), 969–1014.

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

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

Bryan, Kevin A. and Jorge Lemus, "The Direction of Innovation," *Journal of Economic Theory*, 2017, *172*.

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

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

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

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

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

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

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

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

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

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

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

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

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

Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole, "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, 1983, *22* (1), 3–31.

Furman, Jeffrey L. and Scott Stern, "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review*, 2011, *101* (5).

Gans, Joshua and Fiona Murray, "Credit History: The Changing Nature of Scientific Credit," in Adam B. Jaffe and Benjamin F. Jones, eds., *The Changing Frontier: Rethinking Science and Innovation Policy*, University of Chicago Press, 2015.

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

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

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

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

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

Halac, Mariana, Navin Kartik, and Qingmin Liu, "Contests for Experimentation," *Journal of Political Economy*, 2017, *125* (5).

Hengel, Erin, "Publishing While Female," *The Economic Journal*, 2022, *132* (648).

Hill, Ryan and Carolyn Stein, "Scooped! Estimating Rewards for Priority in Science," *Working Paper*, 2023.

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

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

_ **and** _ , "On the Direction of Innovation," *Journal of Political Economy*, 2021, *129* (7).

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

Lamb, David and Susan M. Easton, *Multiple Discovery: The Patterns of Scientific Progress*, Avebury, 1984.

Lazear, Edward P. and Sherwin Rosen, "Rank-order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, *89* (5), 841–864.

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

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

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

Mankiw, N. Gregory and Michael D. Whinston, "Free Entry and Social Inefficiency," *RAND Journal of Economics*, 1986, *17* (1).

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

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

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

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

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

Montiel Olea, Jose Luis and Carolin Pflueger, "A Robust Test for Weak Instruments," *Journal of Business and Economic Statistics*, 2013, *31* (3).

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

Nalebuff, Barry J. and Joseph E. Stiglitz, "Prizes and Incentives: Toward a General Theory of Compensation and Competition," *The Bell Journal of Economics*, 1983, *14* (1).

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

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

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

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

Reinganum, Jennifer F., "Dynamic Games of Innovation," *Journal of Economic Theory*, 1981, *25* (1).

_ , "The Timing of Innovation: Research, Development, and Diffusion," in R. Schmalensee and R.D. Willig, eds., *Handbook of Industrial Organization*, North-Holland, 1989.

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

Robbins, Rebecca and Benjamin Mueller, "After Admitting Mistake, AstraZeneca Faces Difficult Questions About Its Vaccine," 2020.

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

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

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

Stepner, Michael, "Binned Scatterplots: Introducing -binscatter- and Exploring Its Applications," *2014 Stata Conference 4*, 2014.

Stevens, Raymond C., "The Cost and Value of Three-Dimensional Protein Structure," *Drug Discovery World*, 2003.

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

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

The Structural Genomics Consortium, "Mission and Philosophy," 2020.

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

Thompson, Neil and Samantha Zyontz, "Decomposing the "Tacit Knowledge Problem:" Codification of Knowledge and Access in CRISPR Gene-Editing," *Working Paper*, 2021.

Thompson, Neil C. and Jeffrey M. Kuhn, "Does Winning a Patent Race Lead to More Follow-on Innovation?," *Journal of Legal Analysis*, 2020, *12*.

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

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

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

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

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

Williams, Heidi L., "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy*, 2013, *121* (1).

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

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

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

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

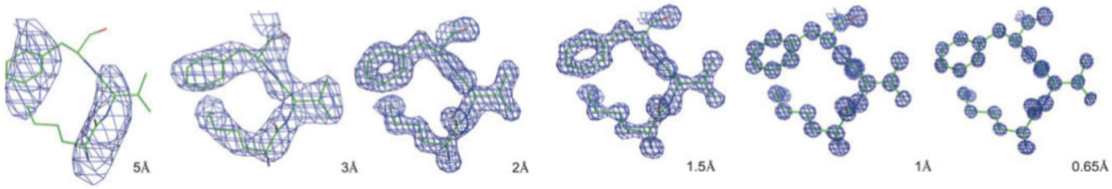
Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan, "Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation," *Science*, 2020, *367* (6483), 1260–1263.

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

Zhou, Ran, "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs," *Working Paper*, 2023.

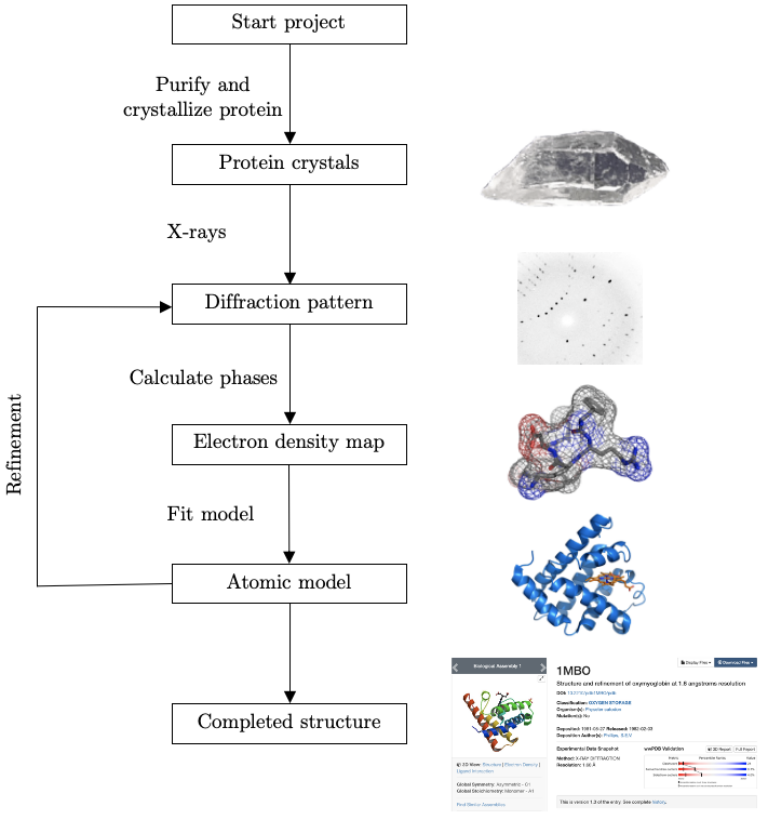
Figures and Tables

Figure 1: Illustration of a Protein Structure at Different Refinement Resolutions



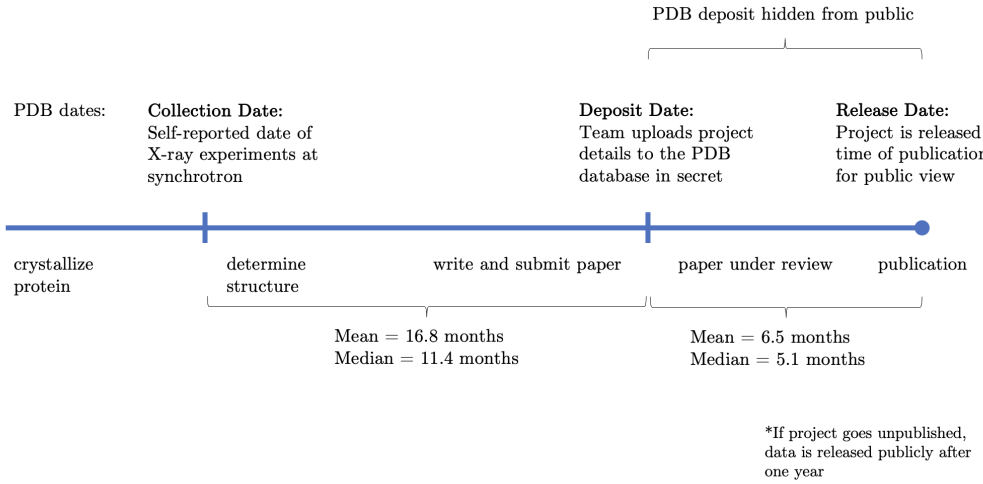
Notes: This figure shows the electron density maps from a fragment of the triclinic lysozyme (PDB ID 2VB1) at different refinement resolutions. The Angstrom (Å) values measure the smallest distance between crystal lattice planes that can be detected in the experimental data. Lower values correspond to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

Figure 2: Summary of the X-Ray Crystallography Process



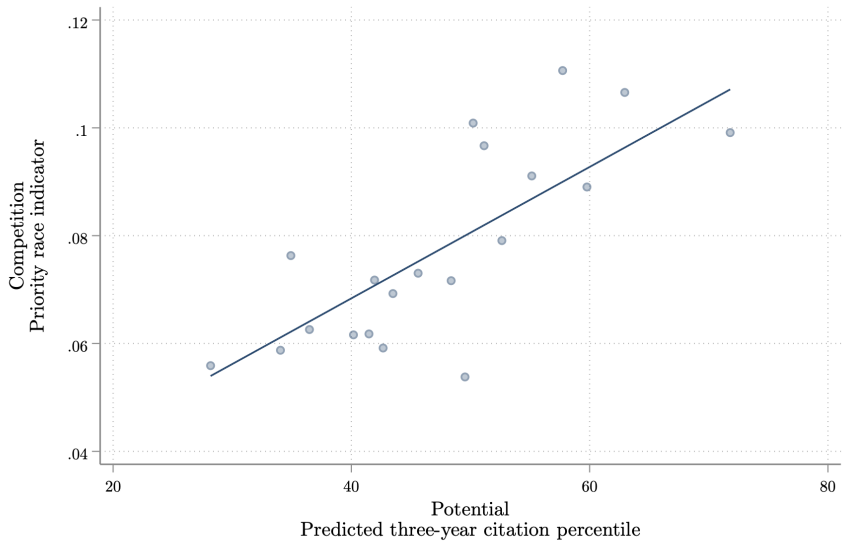
Notes: This figure summarizes the process of solving a protein structure via x-ray crystallography. The images in this figure were taken from Thomas Splettstoesser (www.scistyle.com) and rendered with PyMol based on PDB ID 1MBO.

Figure 3: PDB Timeline



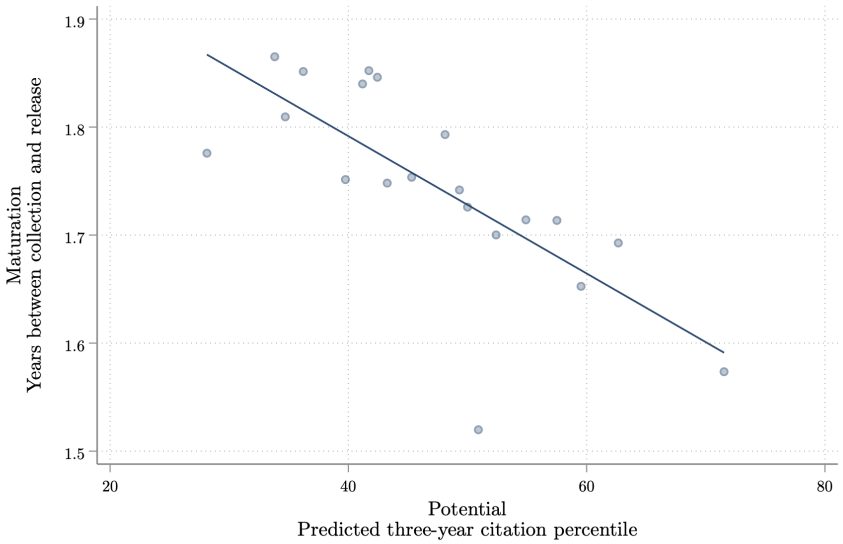
Notes: This figure shows the PDB dates we observe in timeline form. Means and medians are from the full PDB sample. This figure is identical to Figure 1 in Hill and Stein (2023).

Figure 4: The Effect of Potential on Competition



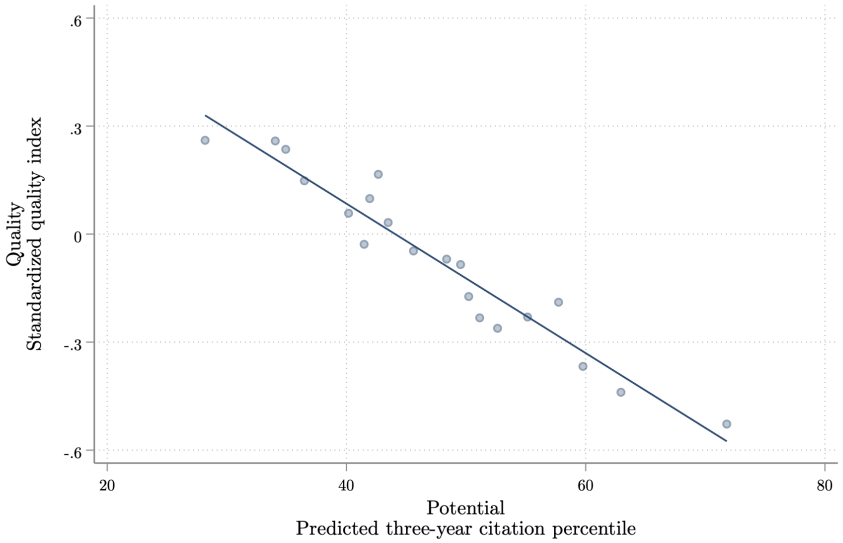
Notes: This figure plots the relationship between potential and competition, testing Proposition 2. Potential is measured as the predicted three-year citation percentile. Competition is measured as an indicator for whether the structure was involved in a priority race. The plot is presented as a binned scatterplot (Stepner, 2014). To construct this binned scatterplot, we first residualize potential and competition with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of competition against the mean of potential in each bin. Finally, we add back the mean competition to make the scale easier to interpret after residualizing. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure 5: The Effect of Potential on Maturation



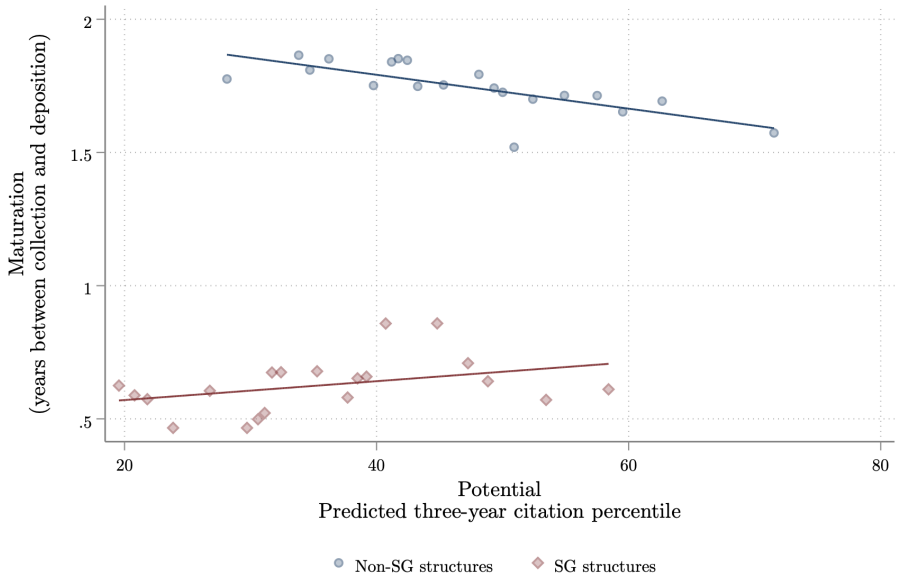
Notes: This figure plots the relationship between potential and maturation, testing Proposition 3. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plot is presented as a binned scatterplot, constructed as described in Figure 4. The sample is the full analysis sample as defined in the text, excluding SG deposits and deposits where the maturation variable is missing.

Figure 6: The Effect of Potential on Quality



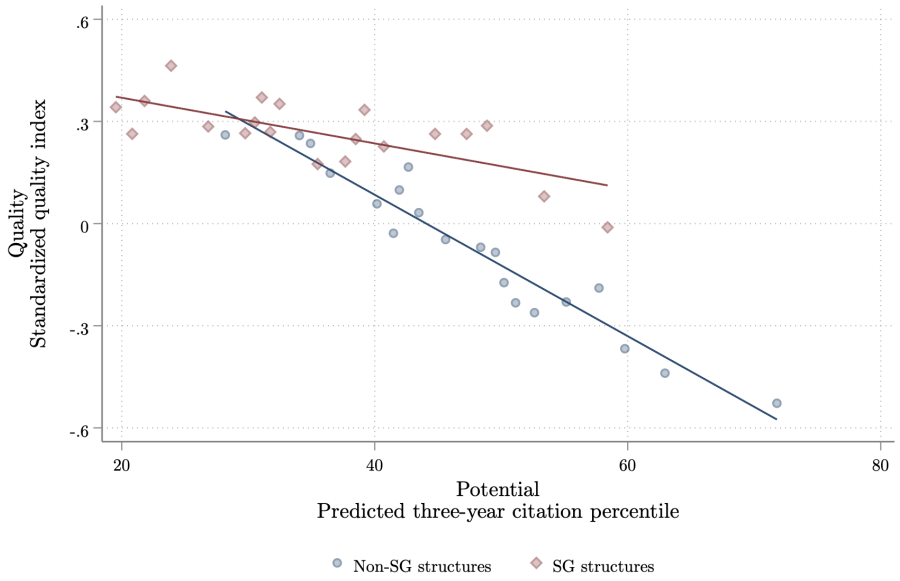
Notes: This figure plots the relationship between potential and quality, testing Proposition 3. Potential is measured as the predicted three-year citation percentile. Quality is measured by our standardized quality index described in detail in Section 3.2.1. The plot is presented as a binned scatterplot, constructed as described in Figure 4. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure 7: The Effect of Potential on Maturation by Structural Genomics Status



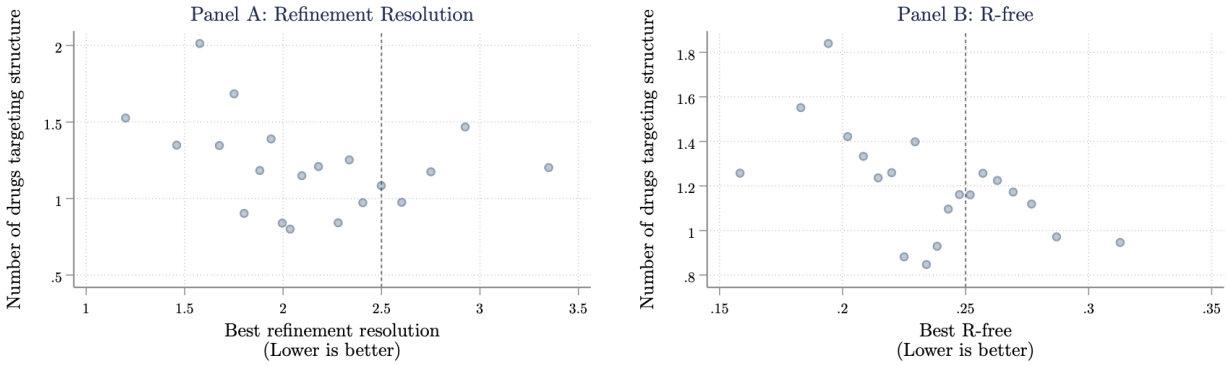
Notes: This figure plots the relationship between potential and maturation, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plot is presented as two separate binned scatterplots, overlaid on the same axes. To construct these binned scatterplots, we first residualize potential and maturation with respect to a set of deposition year indicators (separately by SG status). We then 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 non-missing.

Figure 8: The Effect of Potential on Quality by Structural Genomics Status



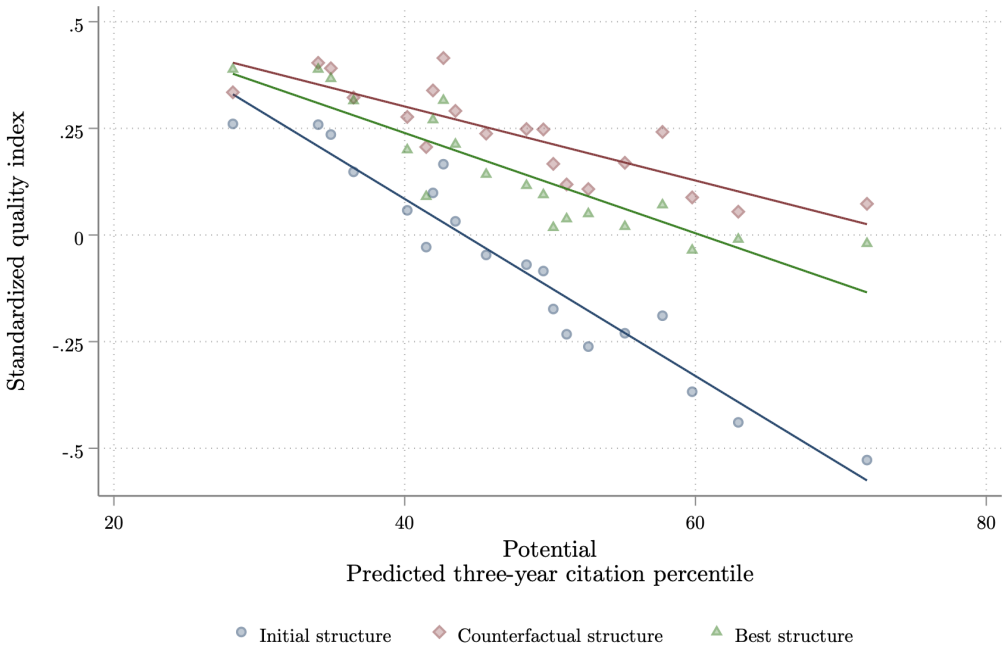
Notes: This figure plots the relationship between potential and quality, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Quality is measured by our standardized quality index described in detail in Section 3.2.1. The plot is presented as two separate binned scatterplots, overlaid on the same axes, constructed as described in Figure 7. The sample is the full analysis sample.

Figure 9: The Relationship between Structure Quality and Drug Development



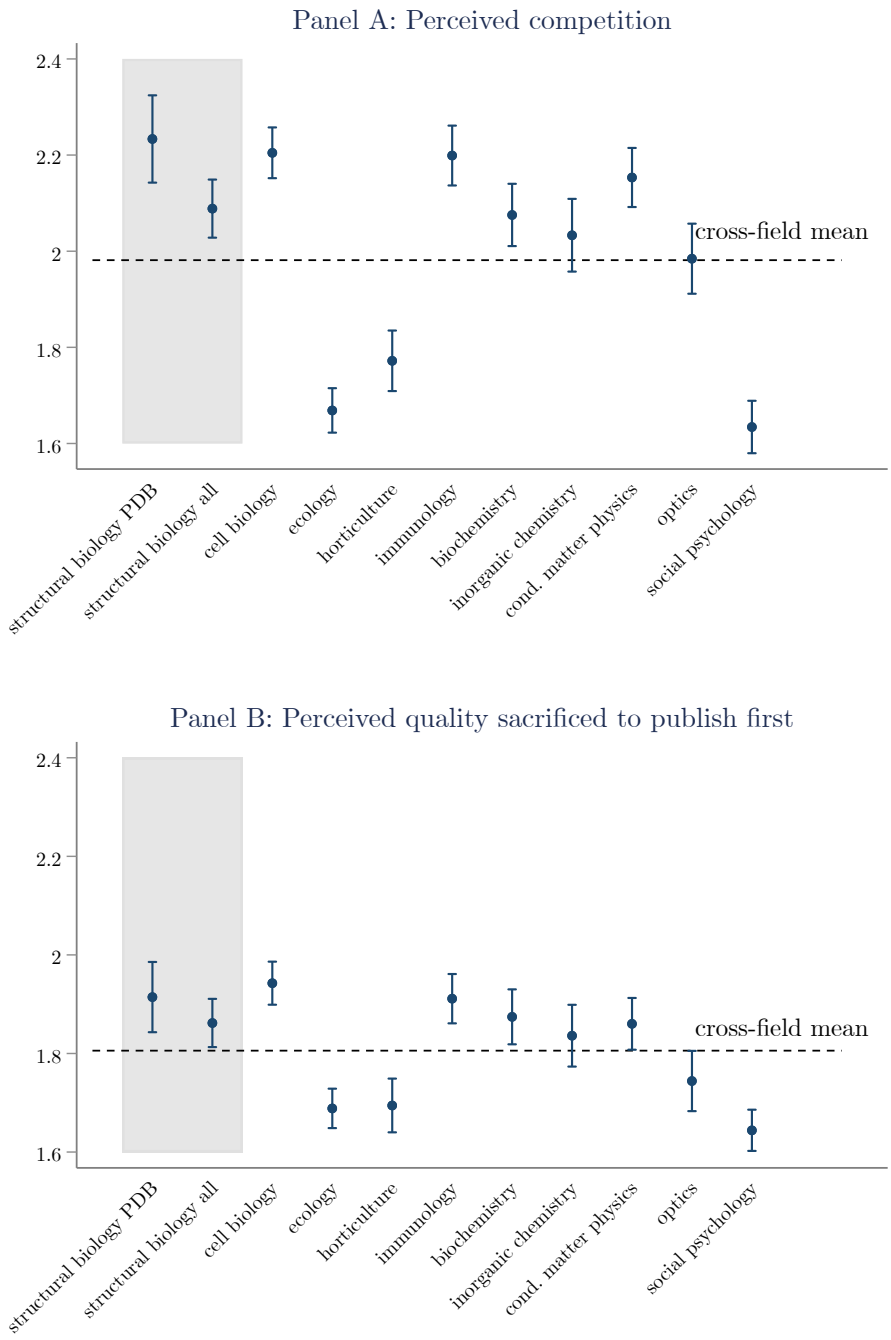
Notes: This figure plots the relationship between structure quality and structure's use in drug design. Quality is measured using unstandardized refinement resolution and R-free, so lower values indicate better quality. In instances where the same structure is deposited in the PDB multiple times, we take the best quality. The results are presented as a binned scatterplots, constructed as described in Figure 4. The dashed lines indicate the quality thresholds for drug development proposed by Anderson (2003). The sample is the full analysis sample.

Figure 10: Subsequent Structure Deposits and Quality Improvement



Notes: This figure plots the relationship between potential and initial quality (in blue), counterfactual quality if the researchers behaved like SG researchers (red), and best version of the structure's quality (green). The details of how we compute counterfactual quality and best quality can be found in Appendix D and B, respectively. Quality is measured using our quality index described in detail in Section 3.2.1. The plots are presented as binned scatterplots, constructed as described in Figure 4. The sample is the full analysis sample.

Figure 11: External Validity: Survey Results



Notes: Panel A shows survey responses to the question "how would you rate the competition to publish first in your field?" Panel B shows survey responses to the question "in general, do you feel that peers in your field ever sacrifice the quality of their research to publish first?" Both questions were answered on a zero to three point scale (with three being the highest). The sample is the 7,882 respondents to our survey. All fields are mutually exclusive, except for "structural biology PDB," which is a subset of "structural biology all." The cross-field means weight all fields equally.

Table 1: Summary Statistics: Full Sample versus Analysis Sample

	All X-Ray Crystallography Sample						Analysis Sample					
	Mean	Median	Std. Dev.	Min	Max	% Missing	Mean	Median	Std. Dev.	Min	Max	% Missing
Panel A. Structure-level statistics												
Quality measures												
Refinement resolution (lower is better)	2.2	2.0	0.6	0.5	15.0	0.2%	2.2	2.2	0.5	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 (1000s of Daltons)	107.1	51.9	600.1	0.3	97730.5	0.0%	95.7	55.3	421.3	1.9	47370.7	0.0%
Residue count (1000s 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 (1000s 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						20,435					
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	63,809						20,435					

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

Table 2: The Effect of Potential on Competition, Maturation, and Quality

	Competition	Maturation		Quality	
	Priority race	Years	Years	Std. index	Std. index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	0.0012***	-0.0064***	-0.0039	-0.0208***	-0.0153***
	(0.0002)	(0.0013)	(0.0025)	(0.0008)	(0.0015)
Principal investigator FEs?			Y		Y
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 FEs?			Y		Y
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	16,216	14,639	12,088	16,216	13,505

Notes: This table shows the relationship between competition / maturation / quality and potential, estimating equation (5) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include 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.

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

Table 3: 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 (1000s of Daltons)	101.5	56.6	471.0	1.9	47370.7	0.0%	73.2	50.6	80.1	5.6	1641.1	0.0%
Residue count (1000s 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 (1000s 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	16,216						4,219					
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	16,216						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, within the analysis sample. Structures are grouped into a single paper based on PubMed ID. For structures with no associated PubMed ID / publication, we impute which structures were part of the same project (see text and Appendix B for details). Complexity variables (molecular weight, residue count, atom site count) are divided by 1000 for ease of interpretation.

Table 4: The Effect of Potential on Maturation and Quality, by Structural Genomics Status

	Maturation	Quality
	Years	Std. index
Dependent variable	(1)	(2)
Panel A. Without complexity controls		
Potential	0.0046***	-0.0082***
	(0.0014)	(0.0011)
Non-structural genomics	1.4900***	0.2663***
	(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	1.4841***	0.2664***
	(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	18,780	20,435

Notes: This table shows the relationship between maturation / quality and potential, interacted with structural genomics status, estimating equation (6) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Structural genomics deposits are defined as 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 structures in the analysis sample. The number of observations in column (1) is lower because maturation is missing for a subset of observations. Heteroskedasticity-robust standard errors are in parentheses.

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

Table 5: The Effect of Competition on Maturation and Quality

	Maturation	Quality
	Years	Std. index
Dependent variable	(1)	(2)
Panel A. Ordinary least squares		
Competition	-0.427***	-0.018
	(0.045)	(0.028)
Complexity controls?	Y	Y
Panel B. Two-stage least squares (instrument = potential)		
Competition	-5.664***	-15.823***
	(1.693)	(2.690)
Complexity controls?	Y	Y
First-stage F statistic	25.2	36.0
Panel C. Two-stage least squares (instrument = human)		
Competition	-5.147***	-7.060***
	(1.473)	(1.266)
Complexity controls?	Y	Y
First-stage F statistic	26.0	37.9
Mean of dependent variable	1.75	-0.07
Observations	14,639	16,216

Notes: This table shows the relationship between maturation / quality and competition. Panel A presents the results from an OLS regression, following equation (7) in the text. Panels B and C present the results from a 2SLS regression, where competition is instrumented with potential and a human indicator respectively, following equations (5) and (8) in the text. The F-statistic is the Montiel Olea and Pflueger (2013) robust F-statistic. The level of observation is a structure-paper pair. 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.

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

Table 6: The Costs of Structure Improvement

		# of Structures	Cost per Structure			
	Duplicate Structure Definition		$ 80,000	$ 100,000	$ 120,000	$ 140,000
least restrictive	All repeated structures	54,816	$ 4,385,280,000	$ 5,481,600,000	$ 6,577,920,000	$ 7,674,240,000
↓	All repeated, non-racing structures	54,172	$ 4,333,760,000	$ 5,417,200,000	$ 6,500,640,000	$ 7,584,080,000
↓	All repeated, non-racing structures with some quality improvement	20,420	$ 1,633,600,000	$ 2,042,000,000	$ 2,450,400,000	$ 2,858,800,000
most restrictive	All repeated, non-racing structures with full quality improvement	18,963	$ 1,517,040,000	$ 1,896,300,000	$ 2,275,560,000	$ 2,654,820,000

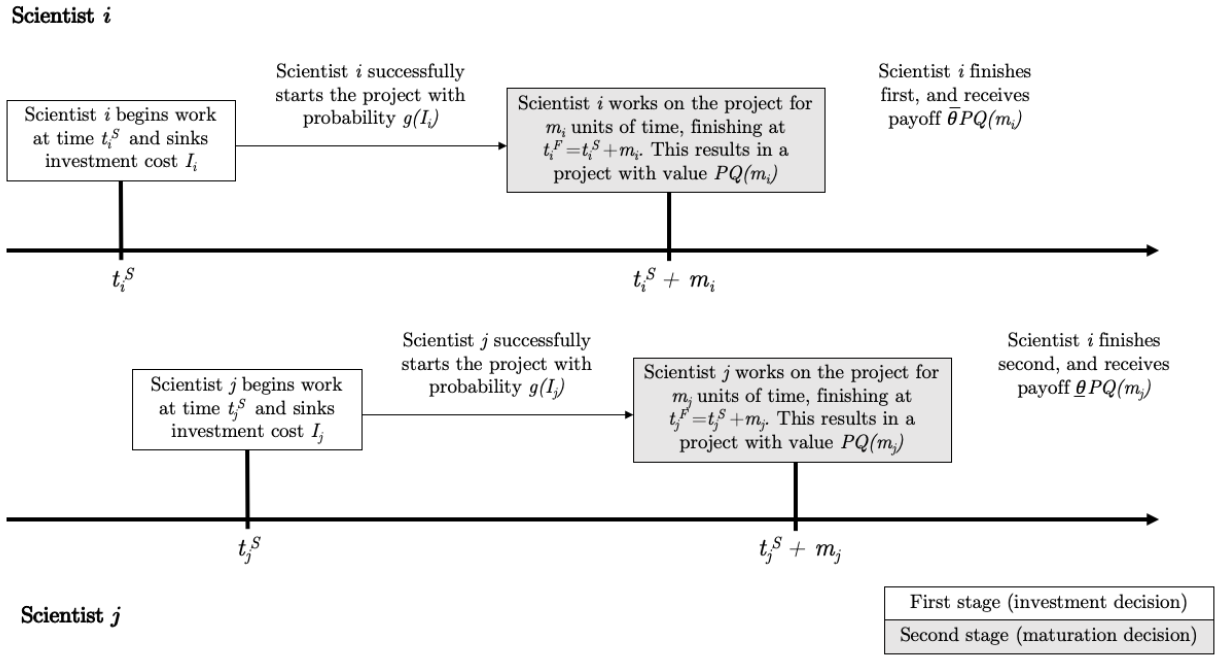
Notes: This paper shows our estimates of the total costs of duplicative work done to improve the quality of protein structures. We present our estimates in a sensitivity analysis format, with varying costs per structure across columns and varying definitions of a "duplicate structure" across rows. See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

Appendices for Online Publication

A Theoretical Appendix

This section formally develops the model outlined in the main text in Section 2, and provides proofs of the propositions. The setup is identical to that of the main text, and is summarized by Figure A1 below.

Figure A1: Model Summary



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

A.1 Maturation

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

A.1.1 The No Competition Benchmark

To build intuition, we start by solving for the optimal maturation period of a scientist who knows that she is not competing for priority. Alternatively, we could consider this the behavior of a naive scientist, who does not recognize the risk of being scooped. This will serve as a useful benchmark

once we re-introduce the possibility of competition.

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

$$m_i^{NC^*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \right\}. \tag{9}$$

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

$$\frac{Q'\left(m^{NC^*}\right)}{Q\left(m^{NC^*}\right)} = r. \tag{10}$$

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

A.1.2 Adding Competition

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

While scientist i knows the probability that j entered the project, she does not know her potential competitor's start time, t_j^S. As described in Section 2, her prior is that t_j^S is uniformly distributed around her own start time. Let $\pi\left(m_i, m_j, I_j\right)$ denote the probability that scientist i wins the race, conditional on successfully entering. Ignoring the choice of I_j for now (simply treating $g(I_j)$ as a parameter g), this can be written as:

$$\pi(m_i, m_j) = (1-g) + gPr(t_i^F < t_j^F) = (1-g) + gPr(t_i^S + m_i < t_j^S + m_j). \tag{11}$$

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

$$m_i^{C^*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \left[\pi(m_i, m_j)\overline{\theta} + (1 - \pi(m_i, m_j))\,\underline{\theta}\right] \right\}. \tag{12}$$

The term outside the square brackets represents the full present discounted value of the project. The terms inside the brackets denote i's expected share of the credit, conditional on i successfully starting the project. The product of these two terms is scientist i's expected payoff conditional on successfully starting the project. Taking the first-order condition of Equation 12 implicitly defines

scientist i's best-response function, which depends on m_j and other parameters:

$$\frac{Q'\left(m_i^*\right)}{Q\left(m_i^*\right)} = r + \frac{1}{\Delta\left(\frac{2\overline{\theta}-g(\overline{\theta}-\underline{\theta})}{g(\overline{\theta}-\underline{\theta})}\right) + m_j - m_i^*}. \tag{13}$$

If we look for a symmetric equilibrium, this yields the proposition below.

Proposition A1. *Assume that first stage equilibrium investment is equal for both researchers, i.e., $I_i^* = I_j^* = I^*$. Further assume that Δ is sufficiently large. Then in the second stage, there is a unique symmetric pure strategy Nash equilibrium where $m_i^* = m_j^* = m^*$ and m^* is implicitly defined by*

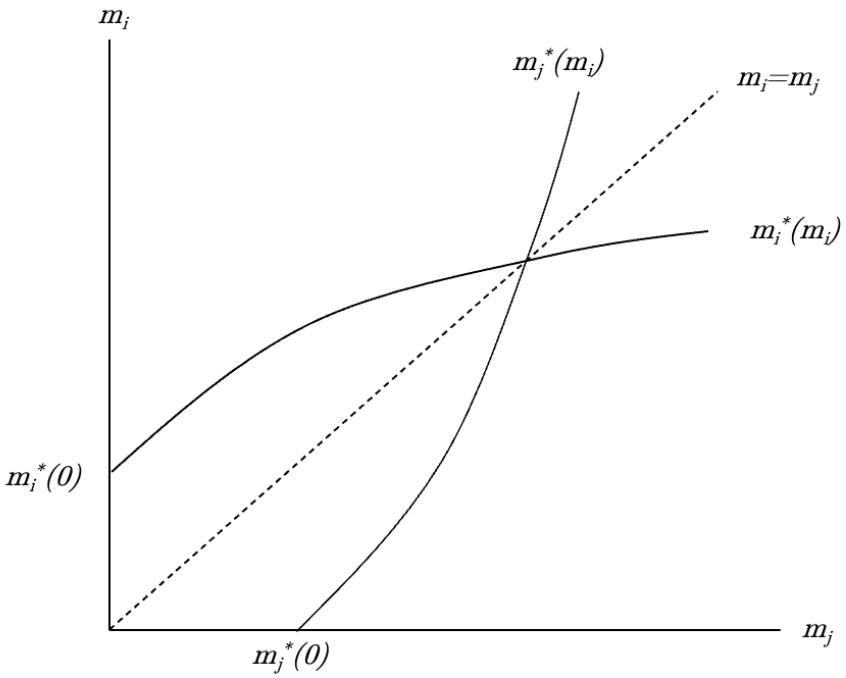
$$\frac{Q'\left(m^*\right)}{Q\left(m^*\right)} = r + \frac{g(I^*)(\overline{\theta}-\underline{\theta})}{\Delta\left(2\overline{\theta} - g(I^*)(\overline{\theta}-\underline{\theta})\right)}. \tag{14}$$

Proof. First, we will expand on how we derive the first-order condition for m_i^* (Equation 13). Taking the derivative of Equation 12 with respect to m_i and setting it equal to zero yields:

$$\frac{Q'(m_i^*)}{Q(m_i^*)} = r - \frac{\frac{\partial\pi}{dm_i}(\overline{\theta}-\underline{\theta})}{\pi(m_i,m_j)\overline{\theta} + (1-\pi(m_i,m_j))\underline{\theta}}. \tag{15}$$

Next, we note that $\pi(m_i,m_j) = (1-g) + g(\frac{1}{2} + \frac{m_j-m_i}{2\Delta})$ and therefore $\frac{\partial\pi}{\partial m_i} = -\frac{g}{2\Delta}$ *if m_i is close enough to m_j*. We will assume this is the case for the moment, and plugging these values into Equation 15 above yields Equation 13 in the text. However, if m_i is much larger than m_j (i.e., if $m_i > m_j + \Delta$), then $\frac{\partial\pi}{\partial m_i} = 0$ and Equation 15 collapses to the no-competition case, i.e., Equation 10. We will return to this caveat, but for now we will assume m_i is close to m_j. Equation 13 implicitly defines $m_i^*(m_j)$ as a function of m_j and parameters. If we can show that (i) $m_i^*(0) > 0$ and (ii) $\frac{dm_i^*}{dm_j} \in (0,1)$, then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $m_i^*(m_j)$ and $m_j^*(m_i)$ will only cross the $m_i = m_j$ line once.

Figure A2: Maturation Best Response Functions



To show (i), plug $m_j = 0$ into Equation 13. This results in an equation that implicitly defines a unique $m_i^*(0) > 0$. To show (ii), we can totally differentiate equation 13 with respect to m_j. For notational ease, define $\zeta \equiv \Delta\left(\frac{2\bar{\theta}-g(\bar{\theta}-\underline{\theta})}{g(\bar{\theta}-\underline{\theta})}\right)$, and note that $\zeta > 0$. Gathering terms and rearranging, we have that

$$\frac{dm_i^*}{dm_j} = \left[\underbrace{\left(\frac{-Q(m_i^*)Q''(m_i^*)+Q'(m_i^*)^2}{Q(m_i^*)^2}\right)\left(\zeta + m_j - m_i^*\right)^2}_{>0} + 1\right]^{-1} \in (0,1). \tag{16}$$

Next, we confirm that the second-order conditions hold. Differentiating the objective function (Equation 12) twice with respect to m_i and evaluating at $m_i = m_j = m^*$ yields

$$Pe^{-rm_i}\left[Q''(m^*) - Q'(m^*)\left(r + \frac{1}{\zeta}\right)\right] < 0. \tag{17}$$

Therefore, $m_i^* = m_j^* = m^*$ is a local optimum. Plugging m^* in for both m_i and m_j (and assuming that $I_i = I_j = I^*$) in Equation 13 yields the expression in Proposition A1. However, as a final check, we need to confirm that this is also a global optimum. Note that Equation 14 tells us that as $\Delta \to 0$, $m_i^* \to 0$. This will yield a payoff of zero for researcher i. This cannot be researcher i's best response, because there is always a $1-g$ probability that her

competitor did not enter. Therefore, she would be better off selecting $m_i = m^{NC^*}$ and hoping that her competitor fails to enter the project. To map this intuition to the math, note that we are now considering a case where $m_i > m_j + \Delta$, and so the relevant first-order condition is now Equation 10. More generally, in order to ensure that $m_i^* = m_j^* = m^*$ is a global optimum we need the payoff from playing $m_i = m^*$ to be larger than the payoff to playing $m_i = m^{NC^*}$:

$$e^{-rm^*} PQ(m^*) \left[(1 - \frac{g}{2})\overline{\theta} + \frac{g}{2}\underline{\theta} \right] > e^{-rm_i^{NC}} PQ(m_i^{NC^*}) \left((1-g)\,\overline{\theta} + g\underline{\theta} \right). \tag{18}$$

Because m^* is increasing in Δ, this defines a lower bound on Δ such that this equation will hold. Therefore, $m_i^* = m_j^* = m^*$ is a symmetric pure strategy Nash equilibrium as long as Δ is sufficiently large. Moreover, this is the only possible pure strategy Nash equilibrium. To see this, note that if $|m_i - m_j| < \Delta$, then the first-order condition in Equation 13 applies and we have the equilibrium defined by $m_i^* = m_j^* = m^*$. Alternatively, if $|m_i - m_j| \geq \Delta$, then the first-order condition defined by Equation 10 applies. But this implies that $m_i^* = m_j^* = m^{NC^*}$, which violates the assumption that $|m_i - m_j| \geq \Delta$. Therefore, if Δ is below some threshold, the Nash equilibrium must be mixed. We will focus on the pure strategy case throughout the remainder of the paper. □

Because $Q(m)$ is increasing and concave, we know Q'/Q is a decreasing function. Therefore, by comparing Equations 10 and 14, we can see that $m^{NC^*} > m^*$. In other words, competition leads to shorter maturation periods. This shortening is exacerbated when the difference between $\overline{\theta}$ and $\underline{\theta}$ is large (priority rewards are more lopsided), Δ is small (competitors start the projects close together, and so the "flow risk" of getting scooped is high), or when g is close to one (the entry of a competitor is likely). On the other hand, if $\overline{\theta} = \underline{\theta}$ (first and second place share the rewards evenly), $\Delta \to \infty$ (competition is very diffuse, so the "flow risk" of getting scooped is low), or $g = 0$ (the competitor doesn't enter), then we recover the no competition benchmark.

A.2 Investment

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

$$I_i^* \in \arg\max_{I_i} \left\{ g(I_i)\, e^{-rm_i^*} PQ\left(m_i^*\right) \left[\overline{\theta} - \frac{1}{2} g(I_j) \left(\overline{\theta} - \underline{\theta} \right) \right] - I_i \right\}. \tag{19}$$

Taking the first-order condition of Equation 19 implicitly defines scientist i's best-response function, which depends on I_j, m_i^*, and other parameters:

$$g'(I_i^*) = \frac{1}{e^{-rm_i^*} PQ\left(m_i^*\right)\left[\overline{\theta} - \frac{1}{2}g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{20}$$

If we look for a symmetric equilibrium, this yields Proposition A2 below.

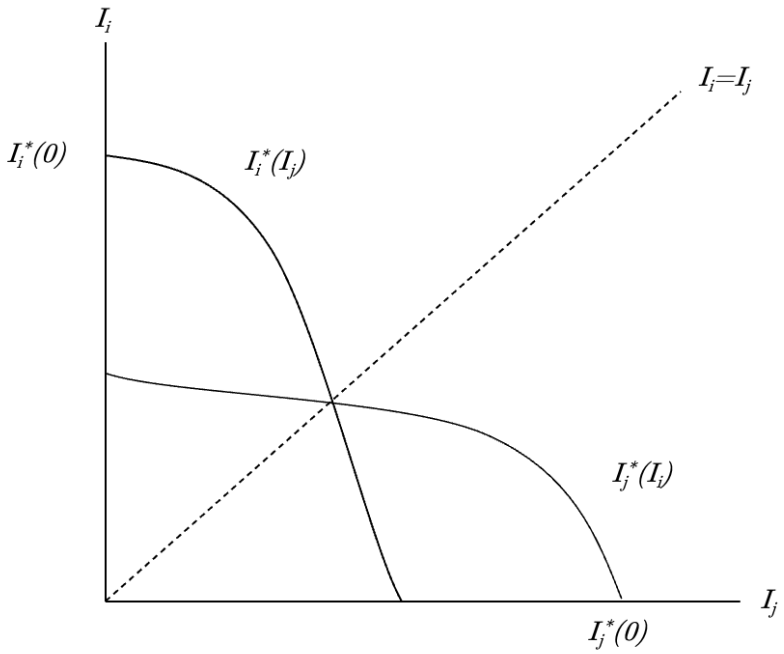
Proposition A2. *Assume that researchers are playing a symmetric pure strategy Nash equilibrium when selecting m in the second stage. Then, in the first stage, there is a unique symmetric pure strategy Nash equilibrium where $I_i^* = I_j^* = I^C$ and I_i^* is implicitly defined by*

$$g'(I^*) = \frac{1}{e^{-rm^*} PQ\left(m^*\right)\left[\overline{\theta} - \frac{1}{2}g\left(I^*\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{21}$$

Together with Proposition A1, this shows that there is a unique symmetric pure strategy Nash equilibrium for both investment and maturation.

Proof. Equation 20 implicitly defines $I_i^*(I_j)$ as a function of I_j, m_i^* (which depends on I_j), and parameters. If we can show that (i) $I_i^*(0) > 0$ and (ii) $\frac{dI_i^*}{dI_j} < 0$ then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $I_i^*(I_j)$ and $I_j^*(I_i)$ will only cross the $I_i = I_j$ line once.

Figure A3: Investment Best Response Functions



To show (i), imagine that j invests zero. Then i should surely invest some positive amount, because the marginal return will be be proportional to $g'(I_i)$. Due to the Inada conditions assumption on $g(\cdot)$, $g'(I_i)$ will be quite large for small values of I_i. To show (ii), we can totally differentiate Equation 20 with respect to I_j. Gathering terms and rearranging, we have that

$$\frac{dI_i^*}{dI_j} = -\frac{e^{-rm_i^*} P \left[\left(\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right) \left(Q'(m_i^*) - rQ(m_i^*)\right) \frac{dm_i^*}{dI_j} - Q(m_i^*) \left(\frac{1}{2} g'(I_j)(\overline{\theta} - \underline{\theta})\right) \right]}{g''(I_j) \left[e^{-rm_i^*} PQ(m_i^*) \left(\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right) \right]^2} < 0 \tag{22}$$

where we can sign this expression by noting that (a) $\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta}) > 0$, (b) $rQ(m_i^*) - Q'(m_i^*) > 0$, and (c) $\frac{dm_i^*}{dI_j} < 0$ and applying assumptions about the function $g(I)$. Therefore, $I_i^* = I_j^* = I^*$ is a unique, pure strategy Nash equilibrium. Plugging in I^* for both I_i and I_j, and plugging in m^* for m_i and m_j yields the expression in Proposition A2. This also confirms our assumption that $I_i = I_j = I^*$ in Proposition A1. □

Equations 21 and 14 together define the optimal investment level and maturation period for scientists when entry into projects is endogenous. This allows us to prove the three key results described in the main text.

Proof of 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. Looking at Equation 14, the left hand side is decreasing in m^*. Looking at the right hand side, we see it is increasing in $g(I^*)$. For the equality to hold as $g(I^*)$ increases, it must be the case that m^* decreases, i.e., that $\frac{dm^*}{dg} < 0$. Because $Q(m)$ is increasing, this also implies that $\frac{dQ(m^*)}{dg} < 0$. □

Proof of 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. Suppose this were not the case. In particular, consider two projects with P_1 and P_2, and further suppose that $P_1 > P_2$. If Proposition 2 is not true, investment for project 1 would be lower than for project 2, i.e., $I_1^* \leq I_2^*$. From Proposition 1, we then know that then $m_1^* \geq m_2^*$ and $Q(m_1^*) \geq Q(m_2^*)$. We also know that $e^{-rm}Q(m)$ is increasing in m for all values of $m < m^{NC*}$. Together, this implies:

$$\underbrace{e^{-rm_1^*} P_1 Q(m_1^*) \left[\overline{\theta} - \frac{1}{2} g(I_1^*)(\overline{\theta} - \underline{\theta})\right]}_{\text{PDV of project 1}} > \underbrace{e^{-rm_2^*} P_2 Q(m_2^*) \left[\overline{\theta} - \frac{1}{2} g(I_2^*)(\overline{\theta} - \underline{\theta})\right]}_{\text{PDV of project 2}}.$$

Therefore, a researcher would want to invest more to enter project 1 than project 2. Thus, we have a contradiction. This implies that $I_1^* > I_2^*$ for any arbitrary pair of projects where $P_1 > P_2$. This implies that $\frac{dI^*}{dP} > 0$ and $\frac{dg(I^*)}{dP} > 0$. □

Proof of 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. □

B Data Appendix

B.1 Description of the Protein Data Bank Data

The first iteration of the Protein Data Bank (PDB) started in 1971. Today, a non-profit organization called the World Wide Protein Data Bank (wwPDB) curates and manages the database. The wwPDB is a collaboration of four existing data banks from around the world: Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB), Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), and Biological Magnetic Resonance Data Bank (BMRB). The data has been standardized and currently represents the universe of discoveries deposited in each of these archives. All new discoveries deposited to any database are transferred to, processed, standardized, and archived by the RCSB (Berman et al. 2006) at Rutgers University. Details about the PDB data can be found on their website.[42]

We access the data directly from the RCSB Custom Report Web Service.[43] The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

- Structure Summary: structure ID, structure title, structure authors, deposit date, release date, experimental technique, molecule classification, macromolecule type, molecular weight, residue count, and atom site count.
- Citation: PubMed ID, publication year, paper authors, and journal name.
- Cluster Entity: entity ID, chain ID, UniPROT accession number, taxonomy, gene name, BLAST sequence 100 percent similarity clusters.
- Data Collection Details: collection date (the self-reported date the scientists generated diffraction data at a major synchrotron or in a home lab).
- Refinement Details: R-free and refinement resolution.

Data about Ramachandran outliers, one of the quality metrics, was not available through RCSB custom reports. Instead, we accessed validation reports data from the PDBe REST API[44] provided by the European Bioinformatics Institute (EMBL-EPI). Data for this study was downloaded on October 25, 2019 and merged using the standard PDB structure identifiers.

[42] http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction
[43] https://www.rcsb.org/pdb/results/reportField.do
[44] https://www.ebi.ac.uk/pdbe/api/doc/validation.html

B.2 Description of the Web of Science Data

Citation data is sourced from the Web of Science produced by Clarivate Analytics and accessed through a license with Stanford University. Our version of the dataset includes digitized academic references through the end of 2018 and is linked to the PDB data using PubMed identifiers. The citation data is restricted to citations between papers linked to PubMed IDs,[45] and self-citations are excluded. Citations are aggregated for each cited paper by publication year of the citing paper. When we report three-year citations, it represents the total number of citations in the publishing year and the subsequent three calendar years.

B.3 Description of the UniPROT Knowledgebase Data

The UniPROT Knowledgebase is a comprehensive, curated database of the biological and functional details of most known proteins. Importantly for our purposes, each protein entry contains a linkage to PDB identifiers of associated structure discoveries. It also contains an annotated bibliography of all associated scientific articles, both structure papers and others, such as articles describing protein function. We count the number of PubMed-linked articles that were published before the first structure discovery as a measure of "potential" or ex-ante demand for a structure model. We only include papers that had been manually reviewed (Swiss-Prot) and exclude those that had only been annotated automatically (TrEMBL). Raw data was accessed on August 26, 2018.[46]

B.4 Description of DrugBank Data

DrugBank is a comprehensive database containing information on FDA-approved drugs and experimental drugs going through the FDA approval process. It includes information on their mechanisms, their interactions, and their targets (Wishart et al., 2018). Academic users may apply for a free license, while all other users require a paid license. We accessed the data on February 20, 2020. Our version of the data includes 11,355 drugs. For every drug, DrugBank provides the protein target(s). We focus on all targets, including both pharmacologically active and inactive targets. There are 5,120 unique protein targets (some protein targets correspond to multiple drugs). If those proteins targets have a PDB ID(s), DrugBank will provide those ID(s). We count the number of times a PDB ID is listed as a drug target as our outcome of drug development use.

B.5 Harmonizing the Data

Variables in our data are reported at three different levels: the entity level, the structure level, and the paper level. Entity is the smallest level, as some protein structures are comprised of multiple entities. Structure is the middle level, as some papers contain multiple structures. Paper is the largest level. The levels fully nest (there is a many-to-one correspondence between entities and

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

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

structures, and a many-to-one correspondence between structures and papers). Below, we report the variables we use at the level they are uniquely indexed:

Variables at the Entity Level:

- Entity ID
- BLAST sequence 100 percent similarity clusters
- Gene linkage
- Taxonomy
- UniProt ID
- UniProt prior articles

Variables at the Structure Level:

- Structure ID
- Determination method
- Classification
- Macromolecule type
- Molecular size (molecular weight, residue count, atom site count)
- Dates (collection date, deposition date, release date)
- Quality measures (refinement resolution, R-free, Ramachandran outliers)
- Structure authors

Variables at the Paper Level:

- PubMed ID
- Paper authors
- Citations

Throughout our analysis, we use a protein structure as our unit of analysis. However, some of the variables we need are indexed at either the entity or the paper level. To create a one-to-one link between papers and structures, we drop all instances where papers are linked to multiple structures (20 percent of PDB-linked papers). Moreover, since about 30% of deposits are never published, we make a similar restriction for groups of structure deposits that appear to have been part of the same unpublished project. We group unpublished structures into the same "project" if the deposits

have the same first and last PDB structure author and share the same release date. Unpublished projects with more than one structure are dropped to mirror the single-structure paper restriction. Appendix Figure E2 assesses this heuristic among the set of published structures.

To similarly create a one-to-one link between structures and entities, we aggregate some of the entity-level measures up to the structure level. While the vast majority of structures have a single entity, about 21 percent have multiple entities. Therefore, we make the following aggregation choices:

- Priority structure. Our sample restricts to the first protein of its kind to be deposited in the PDB (we call this the "priority structure"). However, protein similarity is computed at the entity level. Within each 100 percent similarity, we flag the first deposit (in terms of release date) as the "priority entity." If an entity has not been assigned a 100 percent similarity cluster (this happens if the entity has fewer than 25 amino acids, 12.7 percent of all entities), we do not treat it as a "priority entity." If a structure contains multiple entities and any of those entities is a "priority entity," then we call the structure a priority structure. In other words, if some component of the structure is novel, we treat the entire structure as novel.

- Priority race. As an additional measure of competition, we have an indicator for whether a protein was involved in a priority race. Following Hill and Stein (2023), we code a structure as being involved in a priority race if any of the entities were involved, but drop instances where structures contain more than 15 entities (less than one percent of the sample).

- Gene, taxonomy. Both of these are indexed at the entity level. 9.4 percent of structures are linked to multiple genes, and 5.9 percent of structures are linked to multiple taxonomies. In these cases, we assign the mode as the structure gene / taxonomy (ties broken alphabetically).

- UniProt prior articles. The number of previous articles about a protein is indexed at the entity level. We sum these across all entities to get a number for the structure.

- Best quality by structure. Quality measures are at the structure level, but again protein similarity is computed at the entity level. To compute the best quality level for each structure, we first assign the same quality score to every entity in the structure (using the quality index as our measure). Then, for clusters with multiple entities, we compute the maximum quality and call this the best quality entity within the cluster. Then, we merge these best quality entities back to their respective structures, and collapse back to the structure level by averaging. Thus, the "best quality" may come from a combination of structures. We think this is an accurate way to think about best quality, because scientists have the option of looking at multiple structures.

- Quality improvement. As part of our effort to measure the cost of improving structures, we develop an indicator which codes for whether a protein structure represents an improvement over a prior structure. Quality measures are at the structure level, but again protein similarity is computed at the entity level. For every entity in the protein structure, we measure

whether it represents an improvement over the "priority entity." We then aggregate up to the structure level. A structure with any improved entity is coded as having "any improvement" and a structure with all entities being better than the priority entity is coded as having "full improvement."

Our results are qualitatively similar if we restrict to structures with just one entity (results available upon request). Therefore, we do not believe our aggregation choices are driving the results.

C Survey Details

C.1 Selection of Comparison Fields

We collect email addresses from corresponding authors listed on publications in the Web of Science. We focus on papers published in 2017 and 2018, which is the most recent sample for which we have Web of Science data access. We want to sample across different fields of science, but the Web of Science does not have field tags for papers or authors. We therefore merge the data to the Microsoft Academic Graph (MAG) using DOI paper identifiers and use the paper-level field tags in the MAG dataset. MAG has a hierarchy of field codes, and sometimes assigns multiple codes at each level. We simplify each paper field tag to the combination of the level-0 and level-1 codes that have the highest classification score according to the MAG field clustering algorithm (e.g. physics-astronomy). We then assign each author to a field based on their modal paper-level field.

In an effort to write survey questions that would be sensible to scientists in different fields, we decided to focus on experimental fields of science. This allowed us to tailor our questions. Therefore, our first step was to classify MAG fields based on the share of papers that have the word stub "experiment" in their abstract (we sample 1000 abstracts from each field for computational convenience to do this step). From there, we sort all level-0/level-1 field combinations by experimental share and pick fields by hand that have a mix of high experimental share and a high number of email addresses. We also looked for breadth of scientific methodology and topics, choosing some subfields of the life sciences, physical sciences, and social sciences. Our final list of nine comparison fields includes: biology-cell biology, biology-ecology, biology-horticulture and biology-agronomy (combined), biology-immunology, chemistry-biochemistry, chemistry-inorganic chemistry, physics-condensed matter physics, physics-optics, and psychology-social psychology.

C.2 Selection of Structural Biologists

Structural biology is not listed as a specific field in the MAG taxonomy. Therefore, we use two approaches to constructing the structural biology group. First, we find all email addresses that are listed on papers directly linked to the PDB, giving us 3,038 addresses. We call this the "structural biology - PDB" group. This is the sample that most directly matches the authors in our main analysis, but we were concerned that the sample size might be too small. Therefore, we also used a second approach to supplement this sample. In this approach, we calculate the share of all level-0/level-1 fields that contain a link to a PDB publication, and select fields that have the largest share

of PDB-linked papers. The final combinations we chose for this broader category are: biology-stereochemistry, biology-crystallography, biology-biophysics, and chemistry-stereochemistry. Not including the email addresses directly linked to the PDB, this broader category consists of 7,195 email addresses. We combine our PDB group with this group to create a larger sample that we call "structural biology - all."

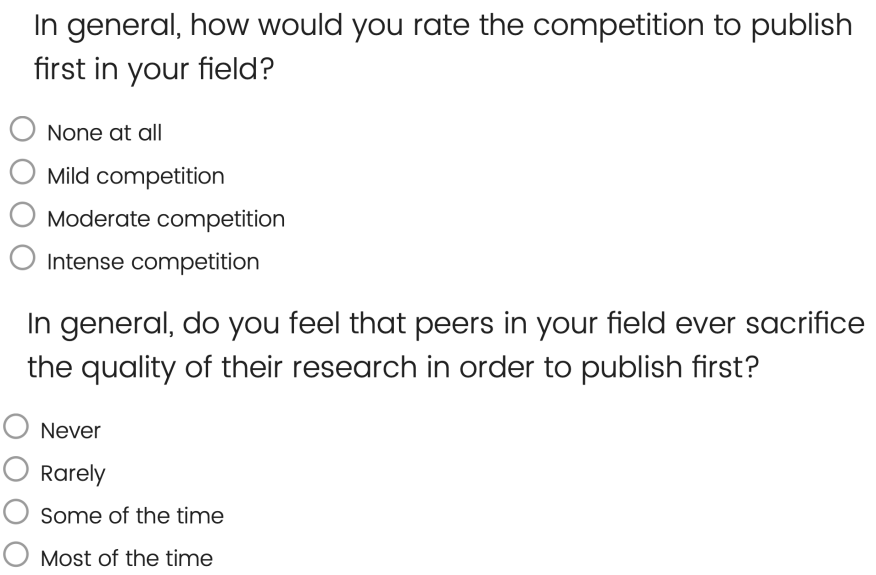
C.3 Survey Implementation and Text

Power calculations based on piloting and detailed in our survey pre-registration (available on the AEA RCT pre-registry, ID #AEARCTR-0011356) suggested that we would need around 1,000 responses per field in order to draw meaningful comparisons across fields. We expected a 8-10% response rate based on piloting, and therefore randomly selected 10,000 email addresses per field (or used all addresses if the total number was less than 10,000). No personally identifiable information was collected from respondents, and the survey was deemed exempt by the UC Berkeley IRB (protocol #2023-05-16350).

We ran the survey using Qualtrics, and sent the initial email on May 15, 2023 to 99,282 email addresses. We sent a reminder to anyone who had not filled out the survey on May 18 and May 24. We closed the survey on June 5. In total, we received 10,557 complete responses (10.6% response rate). We dropped all responses that Qualtrics coded as likely spam, leaving us with 9,211 responses. Unfortunately, due to an autocomplete error on Qualtrics, one of the questions had an incorrect response option. We discovered this error 29 minutes after we launched the survey. We immediately corrected the error and dropped all responses that we received prior to fixing it. This left us with a sample of 7,882 responses. 88 percent of respondents completed the survey, leaving us with a final sample of 6,955 responses.

The survey consisted of two questions. The exact text of the survey as it appeared to respondents is below.

Figure C1: Qualtrics Survey Questions



D Welfare Calculations

D.1 Imputing Missing Quality

Recall the key difference-in-differences estimating equation for the SG and non-SG structures:

$$Y_{it} = \alpha + \beta P_i + \lambda NonSG_i + \delta(P_i \times NonSG_i) + \tau_t + X_i'\gamma + \varepsilon_{it} \tag{23}$$

The regression includes potential (P), an indicator for non-SG structures ($NonSG$), the interaction between the two ($P \times NonSG$), year fixed effects (τ_t), structure covariates (X_{it}), and an unobserved individual shock (ε_{it}). To compute the counterfactual quality of a non-SG structure if they behaved like an SG researcher, we simply plug in $NonSG = 0$ for these non-SG structures. This yields:

$$Y_{it}^{CF} = \alpha + \beta P_i + \tau_t + X_i'\gamma + \varepsilon_{it} = Y_{it} - \lambda - \delta P_i. \tag{24}$$

D.2 Costs of Improved Deposits

In principle, we simply want to count the number of deposits that were strict improvements to past deposits and multiply this count by an estimated cost per deposit. In practice, defining an improved deposit is nuanced. We lay out the details here. In an effort to be complete and conservative, we have four different definitions, each increasingly restrictive.

The first challenge arises because, as discussed in Appendix B, different variables are defined at different levels. In particular, quality is defined at the structure level. However, similarity is defined at the entity level, and some structures have several entities. We use both of these variables to define structure improvements, as detailed below.

Definition 1 (least restrictive). We start with all of the protein structures in our sample (144,173 structures). We drop all non x-ray structures, since we don't have quality scores for these. This leaves us with an initial sample of 128,876 structures. Our broadest definition counts the number of structures with zero novel entities (where novel is defined as being the only entity in a 100% similarity cluster). This results in 54,816 structures (44% of the x-ray sample).

On the one hand, it is possible that even this definition is conservative. Using the 100% similarity clusters to define repeated entities means that highly similar entities will not count as repeated. And because we require all entities to be repeats, a multi-entity structure that is mostly (but not exclusively) repeated entities will not count. On the other hand, some of these structures may be unintentional repeat deposits, in the sense that they were engaged in a priority race and were novel structures at the time they were being worked on. We are interested in computing the number of structures that were *intentionally* re-deposited as a direct replication to the original project. This motivates our next definition.

Definition 2. We take our subsample of 54,816 structures from definition 1 but we drop any structures that were involved in a priority race (see Hill and Stein (2023) for more details on how priority races are defined). The goal is to exclude unintentional re-deposits, i.e., projects that would have been produced anyway because the racing teams were working contemporaneously. This leaves us with 54,172 structures (42% of the x-ray sample).

Definition 3. So far, we have not imposed any restrictions that these re-deposits represent improvements over the initial deposits. In our model, the sole purpose of re-deposits is to improve the quality. Thus, we might argue that only these should count in our calculation of the costs of improved deposits. However, to the extent that there is ex-ante uncertainty about a structure's completed quality, then perhaps some of these re-deposits that do not represent quality improvements were still solved with the intention of improving quality and should be counted. Rather than taking a strong stand, our next two definitions will further restrict the sample to improved deposits whereas definitions 1 and 2 do not make this restriction.

Here we run into the issue of quality being defined at the structure level whereas similarity is defined at the entity level, as discussed in Appendix B. Aggregating up to the structure level, we call any re-deposit an "improved re-deposit" if at least one entity is an improvement over the priority entity. This leaves us with 23,318 structures (18% of the x-ray sample).

Definition 4 (most restrictive). Definition 4 is the same as definition 3, except that we require all entities (rather than at least one) to be an improvement over the priority entity. This leaves us with 21,793 structures (17% of the x-ray sample).

E Appendix Figures and Tables

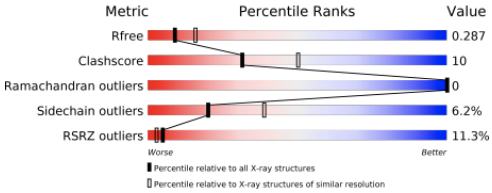
Figure E1: Validation Report for PDB ID 4CMP — Crystal Structure of S. pyogenes Cas9

1 Overall quality at a glance

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

The reported resolution of this entry is 2.62 Å.

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



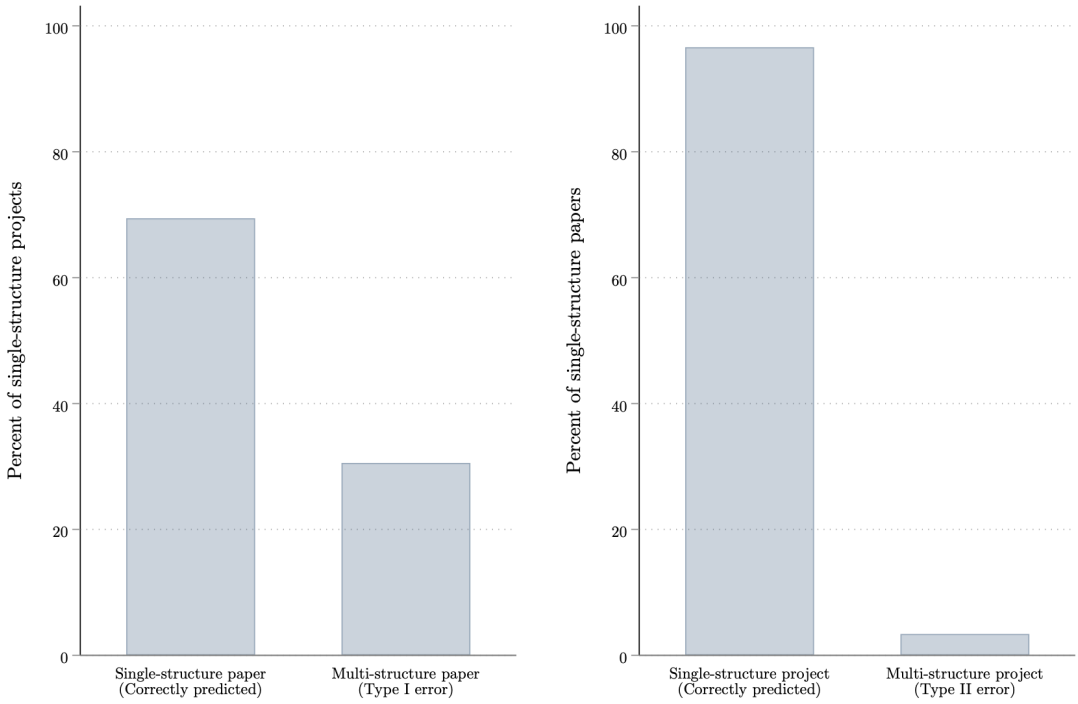
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	3285 (2.64-2.60)
Clashscore	122126	3641 (2.64-2.60)
Ramachandran outliers	120053	3586 (2.64-2.60)
Sidechain outliers	120020	3586 (2.64-2.60)
RSRZ outliers	108989	3218 (2.64-2.60)

4 Data and refinement statistics

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

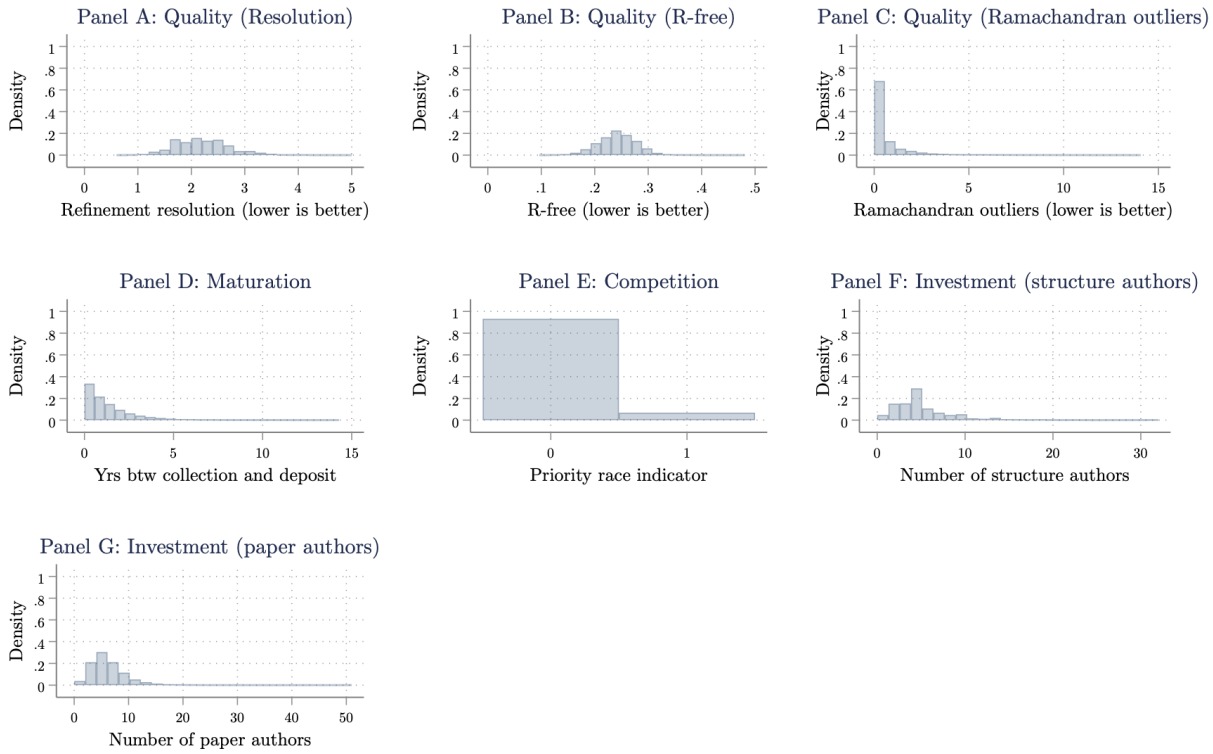
Notes: This figure presents some snapshots from the PDB x-ray structure validation report for PDB ID 4CMP. The "Source" column describes the software package (if applicable) that calculated the quality measure / property.

Figure E2: Predicting Single-Structure Projects



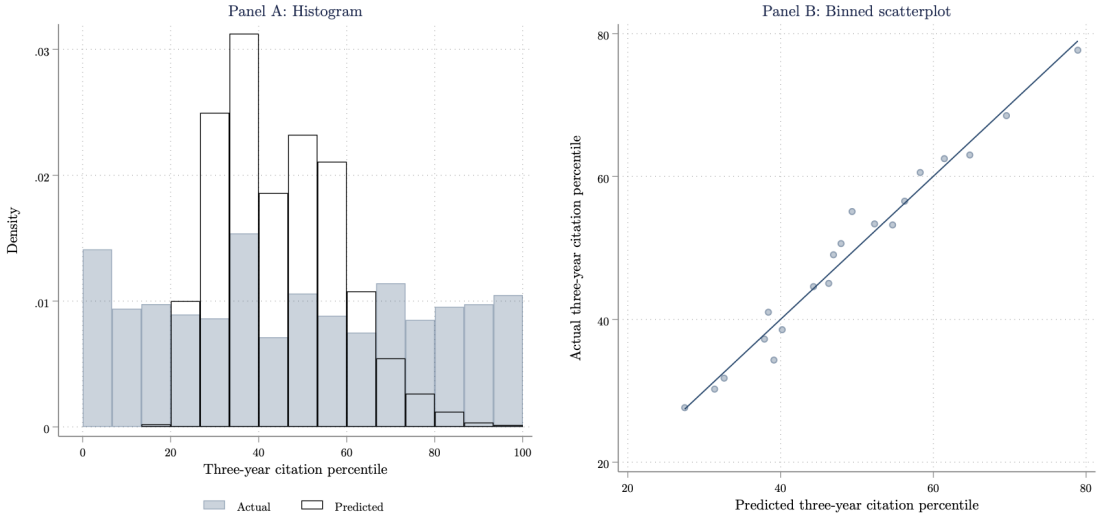
Notes: This figure assesses how well we predict whether a structure will be the only structure in a paper. Panel A looks at the set of structures we predict will fall in single-structure papers ("single structure projects"). About 70 percent of these are indeed single-structure papers, implying a 30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

Figure E3: Distributions of Key Outcome Variables



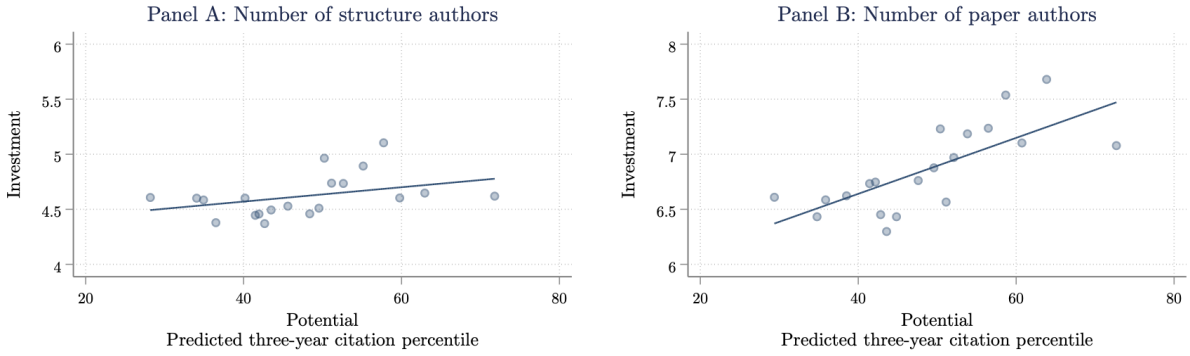
Notes: This figure provides histograms of the distributions of our key outcome variables. All variables have been winsorized at the 99.9^{th} percentile to make the figures easier to read. The sample is the full analysis sample.

Figure E4: LASSO Validation



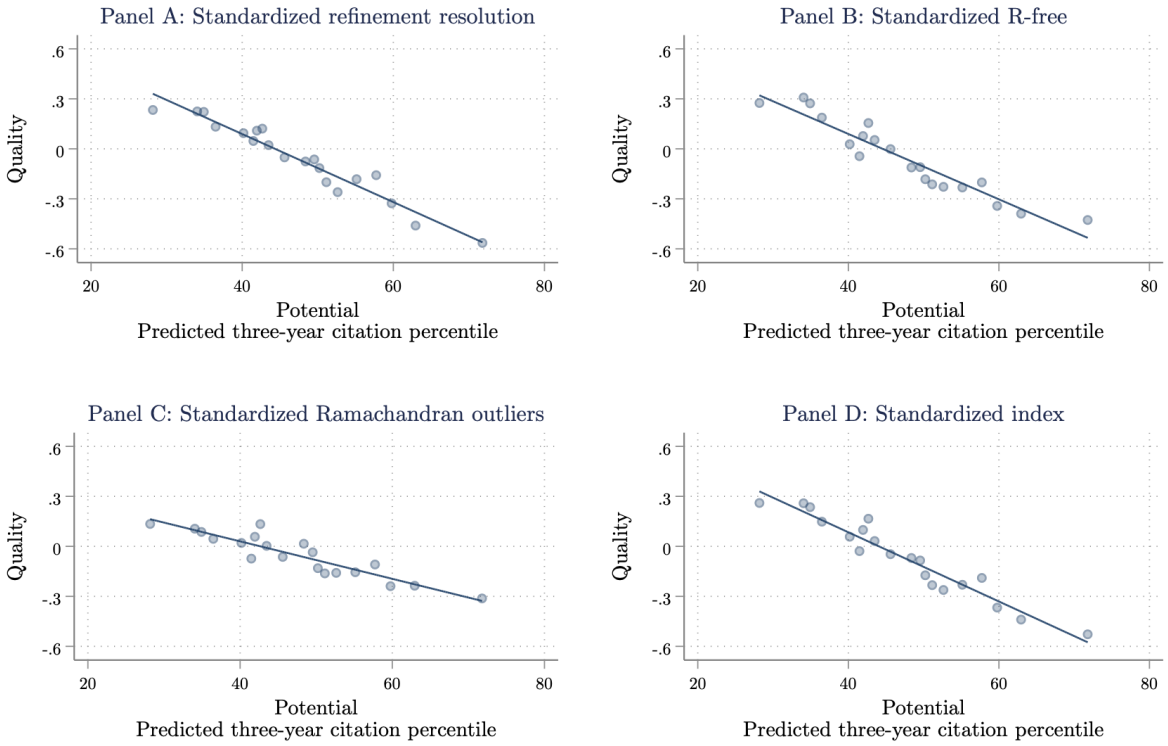
Notes: Panel A of this figure plots the distribution of actual and predicted potential. Panel B presents a graph of actual versus predicted potential as a binned scatterplot. In both panels, potential is measured by the percentile of the structure's three-year citation count. The sample is all structures in the analysis sample that have a three-year citation count.

Figure E5: The Effect of Potential on Investment



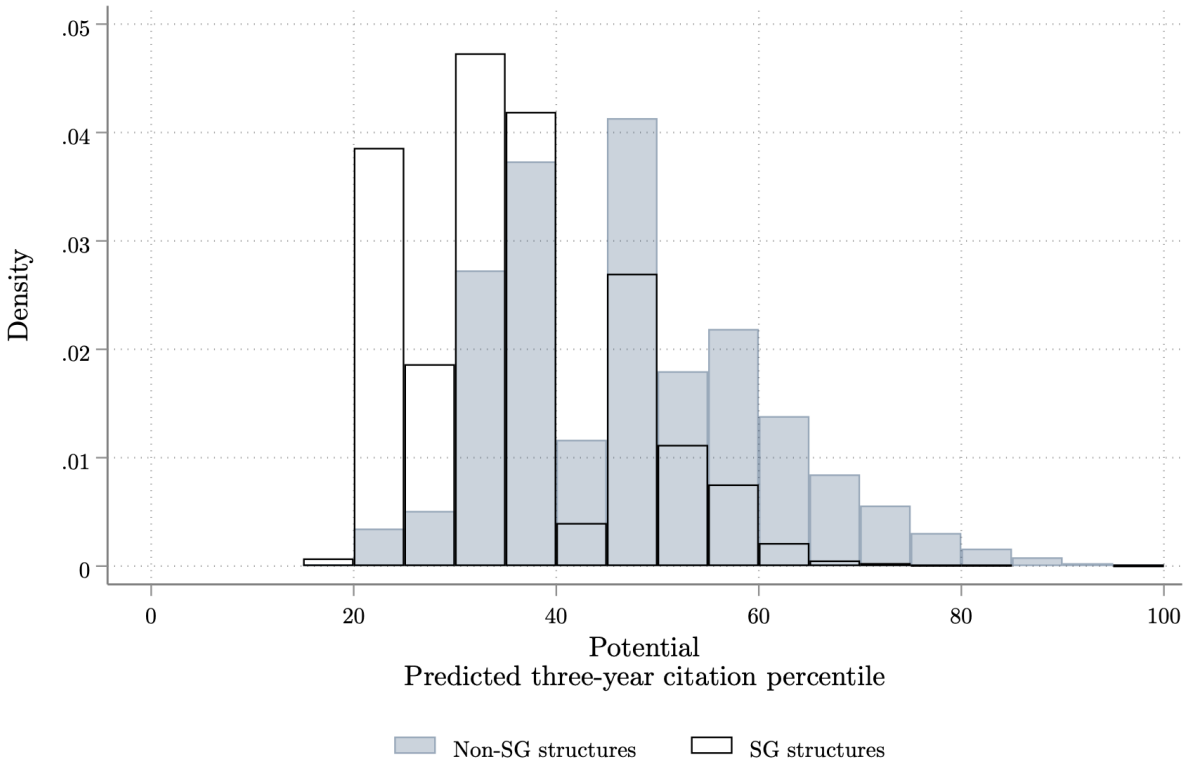
Notes: This figure plots the relationship between potential and investment, testing Proposition 2. Potential is measured as the predicted three-year citation percentile. Investment is measured as the number of structure or paper authors. The plot is presented as a binned scatterplot, constructed as described in Figure 4. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure E6: The Effect of Potential on Quality (Additional Measures)



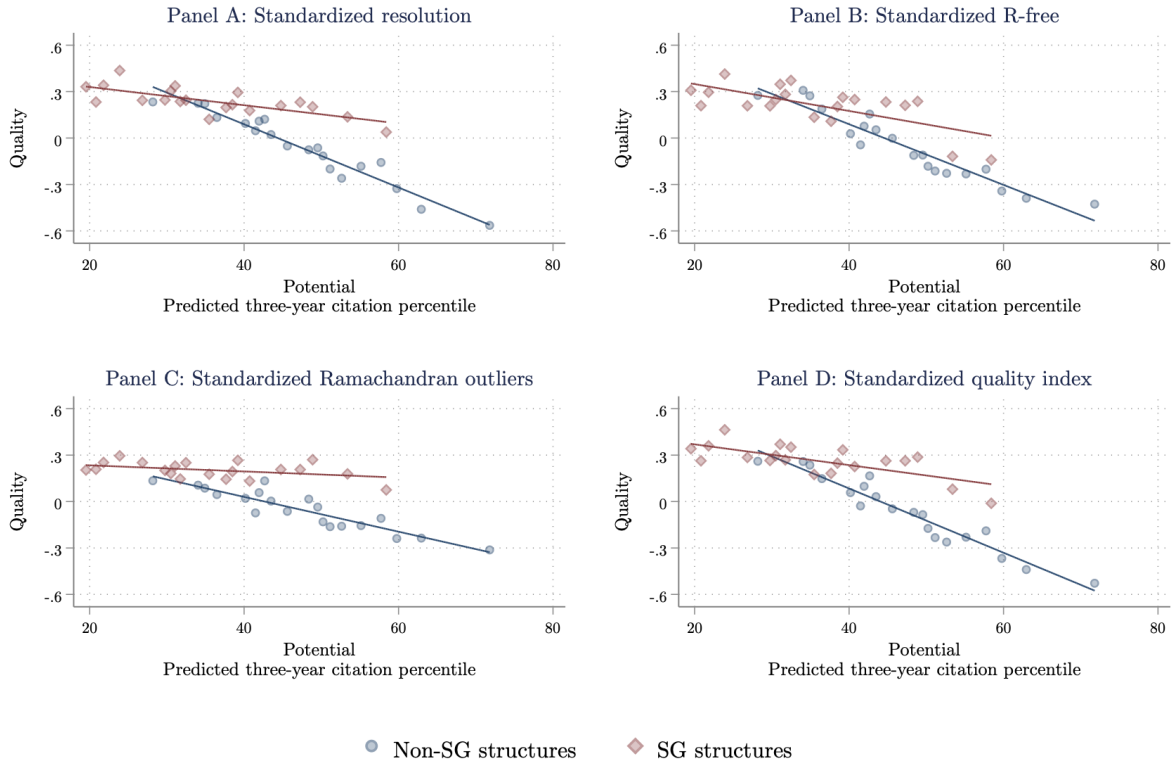
Notes: This figure plots the relationship between potential and additional quality measures, testing Proposition 3. Potential is measured as the predicted three-year citation percentile. Quality measures are described in the text. The plots are presented as a binned scatterplot, constructed as described in Figure 4. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure E7: Potential Distributions by Structural Genomics Status



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

Figure E8: The Effect of Potential on Quality by Structural Genomics Status (Additional Quality Measures)



Notes: This figure plots the relationship between potential and additional quality measures, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Quality measures are described in the text. The plots are presented as two separate binned scatterplots, overlaid on the same axes, constructed as described in Figure 7. The sample is the full analysis sample.

Table E1: Correlation Between Quality Outcomes

	Resolution	R-free	Rama. Outliers
Resolution	1.00		
R-free	0.66	1.00	
Rama. Outliers	0.43	0.46	1.00

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

Table E2: LASSO-Selected Covariates

LASSO-selected variables	Post-LASSO OLS coefficients	LASSO-selected variables	Post-LASSO OLS coefficients
Molecule classification		ISIB	-12.49
Isomerase	-12.72	LINA	14.80
Lyase	-12.03	missing	-8.43
Other	6.09	NAGZ	3.53
Oxioreductase	-5.76	NUTF2	2.00
RNA binding protein / RNA	19.27	PEPT	-6.52
Serine esterase	-8.24	PTPN13	0.29
Transferase	-5.90	SOXA	8.60
Transport Protein	10.87	TC3A	9.24
Unknown function	-16.71	THYX	-6.95
		VP40	1.08
Macromolecule Type		YWLE	3.34
Protein-RNA complex	14.86		
		Other	
Taxonomy		UniProt publications (prior to PDB)	0.190
Homo sapiens	7.04		
Mycobacterium avium	1.25	*Publication Year*	
Sapporo virus	3.53	1996	26.30
Streptomyces himastatinicus	-2.09	1997	22.10
		1998	19.71
Gene		1999	17.60
AGO1	2.78	2000	15.64
ALR1	1.25	2001	13.95
BETVIA	1.61	2002	9.67
BSHA	9.43	2003	9.14
CBFB	11.89	2015	-4.11
DESI1	3.34		
FKBP14	-0.55	Constant	45.29
HPGDS	-0.38		
IGBP1	2.46	R-squared	0.183
INAD	1.25	Observations	12,306

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

Table E3: The Effect of Potential on Competition, Maturation, and Quality (Bootstrapped Standard Errors)

	Competition	Maturation		Quality	
	Priority Race	Years	Years	Std. index	Std. index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	0.0012***	-0.0064***	-0.0039	-0.0208***	-0.0153***
OLS SE	(0.0002)	(0.0013)	(0.0025)	(0.0008)	(0.0015)
Bootstrapped SE	(0.0002)	(0.0015)	(0.0025)	(0.0009)	(0.0014)
Principal investigator FEs?			Y		Y
Panel B. With complexity controls					
Potential	0.0012***	-0.0060***	-0.0034	-0.0190***	-0.0141***
OLS SE	(0.0002)	(0.0014)	(0.0026)	(0.0008)	(0.0014)
Bootstrapped SE	(0.0002)	(0.0015)	(0.0025)	(0.0009)	(0.0012)
Principal investigator FEs?			Y		Y

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

Table E4: The Effect of Potential on Quality (Additional Outcomes)

	Std. refinement resolution	Std. R-free	Std. Rama outliers
Dependent variable	(1)	(2)	(3)
Panel A. Without complexity controls			
Potential	-0.020***	-0.020***	-0.011***
	(0.001)	(0.001)	(0.001)
R-squared	0.049	0.082	0.064
Panel B. With complexity controls			
Potential	-0.019***	-0.018***	-0.010***
	(0.001)	(0.001)	(0.001)
R-squared	0.273	0.160	0.101
Mean of dependent variable	-0.062	-0.056	-0.053
Observations	16,216	16,216	16,216

Notes: This table shows the relationship between additional quality measures and potential, estimating equation (5) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include 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 mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses. *$p<0.1$, **$p<0.05$, ***$p<0.01$.

Table E5: The Effect of Potential on Quality, Controlling for Journal

Dependent variable	Std. refinement resolution (1)	Std. R-free (2)	Std. Rama outliers (3)	Std. index (4)
Panel A. Without complexity controls				
Potential	-0.014***	-0.015***	-0.009***	-0.015***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.128	0.133	0.098	0.124
Panel B. With complexity controls				
Potential	-0.015***	-0.015***	-0.009***	-0.016***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.321	0.198	0.130	0.244
Mean of dependent variable	-0.062	-0.056	-0.053	-0.069
Observations	16,216	16,216	16,216	16,216

Notes: This table shows the relationship between additional quality measures and potential, controlling for the journal of publication. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include 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 that have been published. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table E6: The Effect of Potential on Quality by Structural Genomics Status (Additional Outcomes)

Dependent variable	Std. refinement resolution (1)	Std. R-free (2)	Std. Rama outliers (3)
Panel A. Without complexity controls			
Potential	-0.007***	-0.010***	-0.003***
	(0.001)	(0.001)	(0.001)
Non-structural genomics	0.357***	0.215***	0.086*
	(0.054)	(0.056)	(0.048)
Potential * Non-structural genomics	-0.013***	-0.009***	-0.008***
	(0.001)	(0.001)	(0.001)
R-squared	0.056	0.090	0.073
Panel B. With complexity controls			
Potential	-0.006***	-0.008***	-0.003***
	(0.001)	(0.001)	(0.001)
Non-structural genomics	0.361***	0.217***	0.080*
	(0.049)	(0.054)	(0.048)
Potential * Non-structural genomics	-0.012***	-0.009***	-0.007***
	(0.001)	(0.001)	(0.001)
R-squared	0.265	0.170	0.106
Mean of dependent variable	0.000	0.000	0.000
Observations	20,435	20,435	20,435

Notes: This table shows the relationship between additional quality measures and potential, interacted with structural genomics status, estimating equation (6) in the text. The level of observation is a structure-paper pair. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Structural genomics deposits are defined as 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 structures in the analysis sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table E7: First Stage Results from Most Common Species

Dependent Variable: Competition	(1)	(2)	(3)	(4)	(5)
Taxonomy	Human	E. coli	Mouse	Yeast	Hay bacillus
Taxonomy indicator	0.033***	0.005	-0.003	0.015	-0.020
	(0.005)	(0.009)	(0.009)	(0.012)	(0.016)
Complexity controls?	Y	Y	Y	Y	Y
First-stage F statistic	38.0	0.3	0.1	1.6	1.5
Count of taxonomy observations	4,005	992	826	616	221
Total observations	16,216	16,216	16,216	16,216	16,216

Notes: This table shows the results from a first-stage regression of a taxonomy indicator on competition. The level of observation is a structure-paper pair. Competition is measured as an indicator for whether the structure is involved in a priority race. All regressions control for deposition year and complexity. The F-statistic is the Montiel Olea and Pflueger (2013) robust F-statistic. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample. Heteroskedacity-robust standard errors in parentheses.

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

Table E8: Assessing Balance Between Non-Human and Human Structures

	Non-human structures	Human structures	Difference
Molecular weight	11.01	10.94	-0.065***
Residue count	6.26	6.20	-0.062***
Atom site count	8.27	8.20	-0.069***
Observations	12,211	4,005	

Notes: This table computes the difference in our complexity measures between human and non-human proteins. The level of observation is a structure. The total number of observations corresponds to the number of non-structural genomics structures in the analysis sample.

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

Table E9: Reduced Form Effect of Human on Maturation and Quality

	Maturation	Quality
	Years	Std. quality index
Dependent variable	(1)	(2)
Human	-0.144***	-0.230***
	(0.032)	(0.018)
Complexity controls?	Y	Y
Mean of dependent variable	1.75	-0.07
Observations	14,639	16,216

Notes: This table shows the relationship between maturation / quality and an indicator for a human protein. The level of observation is a structure-paper pair. 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 (1) is lower because maturation is missing for a subset of observations. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses.

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