Race to the Bottom: Competition and Quality in Science*

Ryan Hill† Carolyn Stein‡

November 5, 2020

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 preempted. The most important (highest potential) projects are the most competitive because they induce the most entry. Therefore, the model predicts these projects are also the most rushed and lowest quality. We test the predictions of this model in the field of structural biology using data from the Protein Data Bank (PDB), a repository for structures of large macromolecules. An important feature of the PDB is that it assigns objective measures of scientific quality to each structure. As suggested by the model, we find that structures with higher ex-ante potential generate more competition, are completed faster, and are lower quality. Consistent with the model, and with a causal interpretation of our empirical results, these relationships are mitigated when we focus on structures deposited by scientists who – by nature of their employment position – are less focused on publication and priority.

This paper is updated frequently. The latest version can be found here.

*We are immensely grateful to our advisors Heidi Williams, Amy Finkelstein, and Pierre Azoulay for their enthusiasm and guidance. Stephen Burley, Scott Strobel, Aled Edwards, and Steven Cohen provided valuable insight into the field of structural biology, the Protein Data Bank, and the Structural Genomics Consortium. We thank David Autor, Jonathan Cohen, Peter Cohen, Glenn Ellison, Chishio Furukawa, Colin Gray, Sam Hanson, Ariella Kahn-Lang, Layne Kirshon, Matt Notowidigdo, Tamar Oostrom, Jonathan Roth, Adrienne Sabety, Michael Stepner, Alison Stein, Jeremy Stein, Sean Wang, Michael Wong, and participants in the MIT Labor and Public Finance lunches for their thoughtful comments and discussions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 (Hill and Stein) and the National Institute of Aging under Grant No. T32-AG000186 (Stein). All remaining errors are our own.

†Northwestern University, Kellogg School of Management, ryan.hill@kellogg.northwestern.edu.

‡MIT Economics Department, cstein@mit.edu. Job Market Paper.

1 Introduction

Credit for new ideas is the primary currency of scientific careers. Credit allows scientists to build reputations, which translate to grant funding, promotion, and prizes (Tuckman and Leahey, 1975; Diamond, 1986; Stephan, 1996). As described by Merton (1957), credit comes — at least in part — from disclosing one's findings first, thereby establishing priority. It is not surprising, then, that scientists compete intensely to publish important findings first. Indeed, scientific history has been punctuated with cutthroat races and fierce disputes over priority (Merton, 1961; Bikard, 2020).[1] This competition and fear of pre-emption or "getting scooped" is not uniquely felt by famous scientists, but rather permeates the field. Older survey evidence from Hagstrom (1974) suggests that nearly two thirds of scientists have been scooped at least once in their careers, and a third of scientists reported being moderately to very concerned about being scooped in their current work. Newer survey evidence focusing on experimental biologists (Hong and Walsh, 2009) and structural biologists more specifically (Hill and Stein, 2020) suggests that pre-emption remains common, and that the threat of pre-emption continues to be perceived as a serious concern.

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

Economists have long studied innovation races, often in the context of patent or commercial R&D races. There is a large theoretical literature which considers the strategic interaction between two teams racing to innovate. These models have varied and often contradictory conclusions, depending on how the innovative process is modeled. For example, in models where innovation is characterized

[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).

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

Along these same lines, we develop a model of how competition spurred by priority races impacts the quality of scientific research. In our model, there is a deterministic relationship between the time a scientist spends on a project and the project's ultimate scientific quality. The scientist will choose how long to work on a given project with this relationship in mind. However, multiple scientists may be working on any given project. Therefore, there is always a latent threat of being pre-empted. The scientist who finishes and publishes the project first receives more credit and acclaim than the scientist who finishes second. This implies that a scientist deciding how long to work on her project must trade off the returns to continued "polishing" against the threat of potentially being scooped. As a result, the threat of competition leads to lower quality projects than if the scientist know she was working in isolation.

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

While the model provides a helpful framework, the primary contribution of this paper is to provide empirical support for the its claims. The idea that competition may lead to lower quality

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

work is intuitive, and many scientists and journalists have speculated that this is the case (Fang and Casadevall, 2015; Vale and Hyman, 2016; Yong, 2018). However, systematically measuring the quality of scientific work is difficult. Consider the field of economics, for example — even with significant expertise, it is difficult to imagine "scoring" papers based on their quality of execution in a consistent, objective manner. Moreover, doing so at scale is infeasible.[3]

We make progress on the challenge of measuring scientific quality in the field of structural biology by using a unique data source called the Protein Data Bank (PDB). The PDB is a repository for structural coordinates of biological macromolecules (primarily proteins). The data are contributed by the worldwide research community, and then centralized and curated by the PDB. Importantly, every macromolecular structure is scored on a variety of quality metrics. At a high level, structural biologists are concerned with fitting three-dimensional structure models to experimental data, and so these quality metrics are measures of goodness of fit. They allow us to compare quality across different projects in an objective, science-based manner. To give an example of one of our quality metrics, consider refinement resolution, which measures the distance between crystal lattice planes. Nothing about this measure is subjective, nor can it be manipulated by the researcher.[4] Figure 1 shows the same protein structure solved at different refinement resolutions, to illustrate what a higher quality protein structure looks like.

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

The PDB has no obvious analog to project importance or potential, which is a pivotal variable in our model. Therefore, we use the rich meta-data in the PDB to construct our own measure. Rather than use ex-post citations from the linked publications as our measure of ex-ante potential (which might conflate potential with the ex-post quality of the work), we leverage the extensive structure-level covariates in the PDB to instead predict citations. These covariates include detailed characteristics of the protein known to the scientist before she begins working on the structure, such as the protein type, the protein's organism, the gene-protein linkage, and the prior number of papers written about the protein. Because the number of covariates is large relative to the number

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

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

of observations, overfitting is a concern. To avoid this, we implement Least Absolute Shrinkage and Selection Operator (LASSO) to select our covariates, and then impute the predicted values.

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

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

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

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

The fact that the most scientifically important structures are also the lowest quality intuitively seems suboptimal from a social welfare perspective. If project potential and project quality are

complements (as we assume in the model), then a lack of quality among high-potential projects is particularly costly from a welfare perspective. Indeed, relative to a first-best scenario in which a social planner could dictate both investment and maturation to each researcher, the negative relationship between potential and quality does imply a welfare loss.

However, the monitoring and coordination costs make this type of scheme unrealistic from a policy perspective. Instead, we consider a different policy lever: allowing the social planner to dictate the division of credit between the first- and second-place teams. We consider this policy response in part because some journals have recently enacted "scoop protection" policies[5] explicitly aimed at increasing the share of credit awarded to teams who lose priority races. We then ask: with this single policy lever, can the social planner jointly achieve the optimal level of investment *and* maturation? Our model suggests no. While making priority rewards more equal does increase maturation periods toward the socially optimal level, it simultaneously may reduce investment levels. If the social planner values the project more than the individual researcher (consistent with the notion of research generating positive spillovers), then this reduced investment may be costly from a social welfare perspective. The optimal choice of how to allocate credit depends on the balance of these two forces, but ultimately may lead to a credit split that is lopsided. This in turn will lead to the observed negative relationship between potential and quality. Therefore, while this negative relationship tells us we are not at an unconstrained first-best, it cannot rule out that we are at a constrained second-best.

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

2 A Model of Competition and Quality in Scientific Research

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

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

[5]These policies ask reviewers to treat recently scooped papers as if they are novel contributions; see Section 5.2.2 for more detail and examples.

project maturation involves a tradeoff between higher project quality and an increasing probability of getting scooped.

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

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

2.1 Preliminaries

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

Timing, Investment, and Maturation. Time is continuous and indexed by t. From the perspective of each scientist, the model consists of two stages. In the first stage, scientist i has an idea. We denote the moment the idea arrives as the start time, or t_i^S. However, the scientist must pay an upfront cost in order to pursue the idea. At t_i^S, scientist i must decide how much to invest in starting the project. If she invests I_i, she has probability $g(I_i) \in [0,1]$ 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.[6]

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

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

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

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

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

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

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, 2020) and anecdotal evidence (Ramakrishnan, 2018), we allow for the second-place team to share some of the credit.

Information Structure. The competing scientists have limited information about their competitor's progress in the race. Scientist i does not observe I_j, and so she doesn't know the probability her opponent enters, although she will have correct beliefs about this probability in equilibrium. In addition, she does not know her competitor's start time t_j^S. All she knows is that it is uniformly distributed around her own start time. In other words, she believes that $t_j^S \sim \text{Unif}\left[t_i^S - \Delta, t_i^S + \Delta\right]$ for some $\Delta > 0$. Figure 2 summarizes the model setup.

2.2 Maturation

We begin by solving the second stage problem of the optimal maturation delay, taking the first stage investment as given. In other words, we explore what the scientist does once she has successfully entered the project, and all her investment costs are already sunk. Our setup is similar to the approach of Bobtcheff et al. (2017), but an important distinction is that we only allow the project's

value to depend on the maturation time m, and not on calendar time t. This simplifies the second stage problem, and allows us to embed the solution into the first stage investment decision in a more tractable way.

2.2.1 The No Competition Benchmark

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

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

$$m_i^{NC^*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \right\}. \tag{3}$$

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

$$\frac{Q'\left(m^{NC^*}\right)}{Q\left(m^{NC^*}\right)} = r. \tag{4}$$

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

2.2.2 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, and denote it $m_i^{C^*}$. Scientist i believes that her competitor has also entered the project with some probability $g(I_j^{C^*})$, where $I_j^{C^*}$ is j's equilibrium first-stage investment. However, because investment is sunk in the first stage, we can treat $g(I_j^{C^*})$ as a parameter (simply g) in this part of the model to simplify the notation.

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

$$\pi(m_i, m_j) = (1-g) + gPr(t_i^F < t_j^F) = (1-g) + gPr(t_i^S + m_i < t_j^S + m_j). \tag{5}$$

The first term represents the probability that j fails to enter (and so i wins for sure), and the second

term is the probability that j enters, but i finishes first. The optimal maturation period is given by

$$m_i^{C*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \left[\pi(m_i, m_j)\overline{\theta} + \left(1 - \pi(m_i, m_j)\right) \underline{\theta} \right] \right\}. \tag{6}$$

The term outside the square brackets represents the full present discounted value of the project. The terms inside the brackets denote i's expected share of the credit, conditional on i successfully starting the project. The product of these two terms is scientist i's expected payoff conditional on successfully starting the project. Taking the first-order condition of Equation 6 implicitly defines scientist i's best-response function, which depends on m_j and other parameters:

$$\frac{Q'\left(m_i^{C*}\right)}{Q\left(m_i^{C*}\right)} = r + \frac{1}{\Delta\left(\frac{2\overline{\theta} - g(\overline{\theta} - \underline{\theta})}{g(\overline{\theta} - \underline{\theta})}\right) + m_j - m_i^{C*}}. \tag{7}$$

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

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

$$\frac{Q'\left(m^{C*}\right)}{Q\left(m^{C*}\right)} = r + \frac{g(I^{C*})(\overline{\theta} - \underline{\theta})}{\Delta\left(2\overline{\theta} - g(I^{C*})(\overline{\theta} - \underline{\theta})\right)}. \tag{8}$$

Proof. See Appendix A.1. □

Because $Q(m)$ is increasing and concave, we know Q'/Q is a decreasing function. Therefore, by comparing Equations 4 and 8, we can see that $m^{NC} > m^C$. In other words, competition leads to shorter maturation periods. This shortening is exacerbated when the difference between $\overline{\theta}$ and $\underline{\theta}$ is large (priority rewards are more lopsided), Δ is small (competitors start the projects close together, and so the "flow risk" of getting scooped is high), or when g is close to one (the entry of a competitor is likely). On the other hand, if $\overline{\theta} = \underline{\theta}$ (first and second place share the rewards evenly), $\Delta \to \infty$ (competition is very diffuse, so the "flow risk" of getting scooped is low), or $g = 0$ (the competitor doesn't enter), then we recover the no competition benchmark.

2.3 Investment

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

loses, her payoff is $\underline{\theta}PQ\left(m_i^C\right)$. Putting these pieces together (noting that in equilibrium, if both i and j enter, they are equally likely to win) and re-arranging, the optimal level of investment is

$$I_i^{C^*} \in \arg\max_{I_i} \left\{ g\left(I_i\right) e^{-rm_i^{C^*}} PQ\left(m_i^{C^*}\right) \left[\overline{\theta} - \frac{1}{2} g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right] - I_i \right\}. \tag{9}$$

Taking the first-order condition of Equation 9 implicitly defines scientist i's best-response function, which depends on I_j, $m_i^{C^*}$, and other parameters:

$$g'(I_i^{C^*}) = \frac{1}{e^{-rm_i^{C^*}} PQ\left(m_i^{C^*}\right)\left[\overline{\theta} - \frac{1}{2} g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{10}$$

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

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

$$g'(I^{C^*}) = \frac{1}{e^{-rm^{C^*}} PQ\left(m^{C^*}\right)\left[\overline{\theta} - \frac{1}{2} g\left(I^{C^*}\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{11}$$

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

Proof. See Appendix A.1. □

Equations 11 and 8 together define the optimal investment level and maturation period for scientists when entry into projects is endogenous. This allows us to prove three key results.

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

Proof. See Appendix A.1. Scientist i selects m_i^C by considering the probability that her competitor enters $g(I_j)$. If this probability goes up, she will choose a shorter maturation period which results in lower quality. □

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

Proof. See Appendix A.1. Scientist i will invest more to enter a high-potential project. Her competitor will do the same. In equilibrium, high-potential projects are more likely to result in priority races. □

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

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

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.[7] Therefore, structural biologists use experimental data to propose three-dimensional models of the protein shape to better understand biological function.

Structural biology has several unique features that make it amenable for our purposes (see Section 3.1.1 below), but it is also an important field of science. Proteins contribute to nearly every process inside the body, and understanding the shape and structure of proteins is critical to understanding how they function. Moreover, many heritable diseases — such as sickle-cell anemia, Alzheimer's disease, and Huntington's disease — are the direct result of protein mis-folding. Protein structures also play a critical role in drug development and vaccine design (Westbrook and Burley, 2018). Protease inhibitors, a type of antiretroviral drug used to treat HIV, are one important example of successful structure-based drug design (Wlodawer and Vondrasek, 1998). The rapid discovery and deposition of the SARS-CoV-2 spike protein structure has proven to be a key input in the ongoing development of COVID-19 vaccines and therapeutics (Wrapp et al., 2020). Over a dozen Nobel prizes have been awarded for advances in the field (Martz et al., 2019).

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

3.1.1 Why Structural Biology?

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

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

Third, the PDB contains rich descriptive data on each protein structure. For each structure, we observe covariates like the detailed protein classification, the taxonomy / organism, and the associated gene. Together, these characteristics allow us to develop measures of the protein's importance, based purely on ex-ante characteristics — a topic we discuss in more detail in Section 4.1.

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

Next, the scientist will bring her crystals to a synchrotron facility and subject the crystals to x-ray beams. The crystal's atom planes will diffract the x-rays, leading to a pattern of spots called a "diffraction pattern." Better (i.e., larger and more uniform) crystals yield superior diffraction

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

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

3.2.1 Measuring Quality

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

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

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

map exactly to Q in our model. Importantly, they score a project on its quality of execution, rather than on its importance or relevance.

An important feature of these measures is that they are all either calculated or independently validated by the PDB, leaving no scope for misreporting or manipulation by authors. Since 2013, the PDB has required that x-ray structures undergo automatic validation reports prior to deposition. These reports take the researcher's proposed model and experimental data as inputs, and use a suite of software programs to produce and validate various quality measures. In 2014, the PDB ran the same validation reports retrospectively on all structures that were already in the PDB (Worldwide Protein Data Bank, 2013), so we have full historical coverage for these quality measures. Appendix Figure C.1 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. As described in Section 3.2.1, scientists can improve resolution by spending time improving the quality of the protein crystals and by fine-tuning the experimental conditions during x-ray exposure. In our main analysis, we will standardize refinement resolution so that the units are in standard deviations and higher values represent better quality.

R-free. The R-free is one of several residual factors (i.e., R-factors) reported by the PDB. In general, R-factors are a measure of agreement between a scientist's structure model and experimental data. Similar to resolution, lower values are better. An R-factor of zero means that the model fits the experimental data perfectly; a random arrangement of atoms would give an R-factor of about 0.63. Two R-factors are worth discussing in more detail: R-work and R-free. When fitting a model, the scientist will set aside about ten percent of the data for cross-validation. R-work measures the goodness of fit in the non-cross-validation sample. R-free measures the goodness of fit in the cross-validation sample. R-free is our preferred R-factor, because it is less likely to suffer from overfitting (Goodsell, 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 C.1). We create the index by adding all three standardized quality measures and then standardizing the sum.

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.

For most structures, the PDB collects two key dates which allow us to infer the maturation period: the collection date and the deposition date. The collection date is self-reported and date corresponds to the date that the scientist subjected her crystal to x-rays and collected her experimental data. The deposition date corresponds to the date that the scientist deposited (i.e., uploaded) her structure to the PDB. Because journals require evidence of deposition before publishing articles, the deposition date corresponds roughly to when the scientist submitted her paper for peer review.[11] The timespan between these two dates represents the time it takes the scientist to go from the raw diffraction data to a completed draft (the "diffraction pattern" stage to the "completed structure" stage in Figure 3). In other words, it is the time spent determining the protein's structure, refining the structure, and writing the paper. However, note that this maturation period only includes time spent working on the structure once the protein was successfully crystallized and taken to a synchrotron. Anecdotally, crystallizing the protein (the first step in Figure 3) can be the most time-consuming step. Because 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.

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

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 the text or performing complementary analyses. Appendix Figure C.2 shows the histogram of the difference between the number of paper authors and structure authors. While we view the number of structure authors as a cleaner measure of investment, because these authors contributed directly to solving the protein structure, we will use both in our analysis.

3.2.4 Measuring Competition

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

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

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

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

will lead to attenuation bias if this proxy is used as an independent variable in a regression.

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. However, as a general rule, larger and "floppier" proteins are more difficult to crystallize than their smaller and more rigid counterparts (Rhodes, 2006). Moreover, since these larger proteins are more complex, with more folds, they are harder to model once the experimental data are in hand. Therefore, despite the general uncertainty of protein crystallization, size is a predictor of difficulty.

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

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.[14] Our version of these data start in 1990 and end in 2018. Broadly, we are able to link the Web of Science citations data to the PDB using PubMed identifiers, which are unique IDs assigned to research

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

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

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,[15] and exclude any self-citations.[16] By restricting to citations in the three years since publication (rather than total cumulative citations) we avoid the problem that older papers have had more time to accumulate citations. Note that these citation variables are unique at the *paper* level, rather than at the structure level. Structures are linked to papers in a many-to-one fashion. In other words, while some papers only have one affiliated structure, other papers may have multiple affiliated structures. We discuss how we handle multiple matching of structures to a single paper in Section 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. 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 drug industry (Wishart et al., 2018). The current release contains over 11,000 drugs, including about 2,600 approved drugs (approved by the FDA, Health Canada, EMA, etc.), 6,000 experimental

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

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

(i.e., pre-clinical) drugs, and about 4,000 investigational drugs (in Phase I/II/III human trials).[17] Importantly for us, beyond just linking to the target protein, DrugBank provides the PDB ID(s) for any target structure that has been deposited in the PDB. This allows us to link structures to the drugs that target them.

3.4 Sample Construction

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

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

Table 1 provides summary statistics for both the full sample and our analysis sample. Panel A presents structure-level statistics and Panel B presents paper-level statistics. Although our analysis sample comprises a small subset of the total structures, it appears fairly representative of the full sample. There are a few exceptions to this claim. The maturation period (years between collection and deposit) is shorter in the analysis sample, likely because we focus on the first deposit of a given protein, and so racing is more likely. Competition (deposits per similarity cluster within two years) is smaller in the analysis sample, but this occurs mechanically because we drop all deposits after the first structure deposition.[20] Similarly, the number of UniPROT papers (i.e., papers published prior to the first structure discovery) is lower in the analysis sample because there are more UniPROT papers for structures in crowded clusters. For more detail on the full distributions of our key outcome variables, see the histograms in Appendix Figure C.4.

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

[18] For structures without an associated publication, we attempt to predict whether the structure would have have been the only structure in a paper *had it been published*. See Appendix B for details. Appendix Figure C.3 suggests that we are able to correctly classify these structures the majority of the time.

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

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

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 4 and 5, which rely on cross-sectional variation in potential. Propositions 4 states that high-potential projects should generate more investment and therefore more competition. Proposition 5 states that high-potential projects should therefore be more rushed and lower quality. We provide a variety of evidence which points to increased competition and rushing — rather than other omitted factors — as the primary channel.

Finally, we return to Proposition 3, which states that more competitive projects (projects at higher risk of having multiple teams competing simultaneously) are more likely to be rushed and lower quality. We do not have a clean measure of ex-ante competition — as discussed in Section 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 4 functions as the first stage, while Proposition 5 is the reduced form.

4.1 Defining Project Potential

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

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

This kind of prediction is possible due to extremely detailed data describing and categorizing every structure in the PDB. Each structure is given a detailed classification (over 500 different classifications, such as "transcription protein" or "signaling protein"), a taxonomy (over 1,000 different organisms, such "homo sapiens" (human) or "mus musculus" (mouse)), and a link to the gene which codes for the protein (over 2,500 different genes). We also take advantage of the UniPROT prior paper measure (described in Section 3.3.2) as an additional predictor. For each structure, we com-

pute the number of citations that the associated publication accrued over the first three years since publication (excluding self-citations). Since the citation counts are heavily right-skewed, we transform these counts into percentiles. We then use these detailed data to predict citation percentiles for each structure. It is worth pointing out that we explicitly *exclude* our complexity covariates from this prediction, in an effort to create a measure of potential that is uncorrelated with project complexity.

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

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

Figure 4 compares actual versus predicted citation percentiles, to help assess the prediction quality. Panel A shows a histogram of actual versus predicted percentiles. While the predicted values are more clustered toward the middle percentiles, we are able to generate fairly good dispersion. Panel B shows the binned scatterplot of actual percentiles on the y-axis versus predicted percentiles on the x-axis. The fit along the $y = x$ line appears quite good throughout the distribution. Taken together, these figures suggest our prediction exercise is reasonably successful. Appendix Table C.2 shows the LASSO-selected covariates and the post-LASSO ordinary least squares coefficients. While many of the coefficients are difficult to interpret, it is reassuring to see some common-sense coefficients — for example, proteins that had more prior papers written before the structure discovery tend to be more highly cited. The R^2 from the post-LASSO ordinary least squares regression suggests that we are able to capture about 17 percent of the variation in actual citation percentile with our predictions.

4.2 The Relationship between Potential and Competition

Proposition 4 predicts that scientists will invest more in starting high-potential projects, which will generate more competition for completing these projects. We measure investment using the number of structure authors and paper authors, as discussed in Section 3.2.3. We proxy for competition by counting the number of times the structure was deposited in the PDB within two years of the initial deposit, as discussed in Section 3.2.4. Because this variable is heavily right-skewed, we take the log.

Figure 5 shows the relationship between investment and potential. We illustrate the relationship using a binned scatterplot. To construct this binned scatterplot, we first residualize investment and

potential with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of investment against the mean of potential in each group. Finally, we add back the mean investment period to make the scale easier to interpret after residualizing. As Figure 5 demonstrates, high-potential projects have both more structure authors and more paper authors, suggesting that researchers allocate more scarce personnel to more important projects. The highest-potential structures have about 4.8 structure authors and 7.5 paper authors on average, while the lowest-potential structures have about 4.5 structure authors and 6.3 paper authors on average.

Figure 6 is similar to Figure 5, but shows the relationship between potential and competition. The highest-potential structures have about 1.5 deposits per similarity cluster,[21] while the lowest-potential structures have about 1.1 deposits in the similarity cluster.

Table 2 formalizes these relationships. For structure i deposited in year t, we estimate:

$$Y_{it} = \alpha + \beta P_{it} + X'_{it}\gamma + \tau_t + \epsilon_{it} \tag{12}$$

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

Panel A presents the estimates of β with deposition year fixed effects, which corresponds to the plots shown in Figures 5 and 6. Throughout the remainder of this paper, we will find it convenient to benchmark effect sizes by comparing structures in the 90^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 31^{st} percentile of the citation distribution, as shown in Panel A of Figure 4) to structures in the 10^{th} percentile of the potential distribution (corresponding to structures *predicted* to fall in the 63^{rd} percentile of the citation distribution). We will term these "high-potential structures" and "low-potential structures" respectively. Columns (1) and (2) focus on the effect of potential on investment. The coefficient of 0.008 in column (1) implies that high-potential structures have 0.25 more structure authors than low-potential structures.[22] Similarly, column (2) implies that high-potential structures also have about one additional author compared to low-potential structures. Both coefficients are statistically significant at the one percent level.

Columns (3) turns to the effect of potential on competition. The coefficient of 0.009 in column (3) suggests that high-potential structures have about 30 percent more deposits in their similarity cluster than low-potential structures.[23] Again, this effect is statistically significant at the one percent level. Appendix Table C.3 provides similar estimates for alternative measures of competition.

Collectively, these results suggest that researchers are interested in maximizing their citations, and rationally choose which projects to invest in and pursue with citations in mind. In other words,

[21] We arrive at this by noting that $e^{0.4} = 1.5$.
[22] We calculate this by taking $0.008 \times (63 - 31) = 0.25$.
[23] We calculate this by taking $e^{0.009 \times (63-31)} = 1.3$.

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 5 predicts that high-potential projects will be completed more quickly, as scientists internalize the fact that they are more likely to face competition for these projects. The second part of Proposition 5 predicts that this decrease in maturation will lead to lower quality among the high-potential projects. Figure 7 shows the relationship between maturation and potential, controlling for deposition year. The highest-potential projects have maturation periods of about 1.7 years, while the lowest-potential projects have maturation periods of about 1.9 years — a difference of just over two months. Figure 8 illustrates the relationship between potential and quality. Across all four quality measures, we see that higher potential is associated with lower quality. The magnitude of these correlations is notable. In Panel A, for example, we see that the highest-potential projects have resolution measures that are nearly a full standard deviation lower than the lowest-potential projects. These trends are fairly consistent across the different quality measures.

Table 3 presents these relationships in regression form. We estimate the same regression as in Equation 12, but replace the dependent variable Y with our measures of maturation and quality. β remains the coefficient of interest, because it describes the relationship between potential and maturation or potential and quality. Focusing on Panel A, column (1) shows that higher-potential projects have shorter maturation periods. The coefficient of -0.005 implies that high-potential structures are completed about 0.17 years (or just over two months) faster than low-potential structures. Since the typical low-potential structure takes has a maturation period of about 1.9 years, this represents a decline of about nine percent. This effect is statistically significant at the one percent level.[24]

Columns (2) to (5) of Table 3 measure the effect of potential on quality. Again looking at Panel A and focusing on the aggregate quality index in column (5), the coefficient of -0.021 implies that high-potential structures have quality index scores that are about 0.7 standard deviations below their low-potential counterparts. The magnitudes are similar across the other quality measures in columns (2) to (4), and all the coefficients are statistically significant at the one percent level.

Together, these results suggest that high-potential projects are more likely to be finished quickly, which translates to lower quality on average. However, as discussed in Section 4.6, this negative relationship could be driven by omitted variables bias. In this setting, we are particularly concerned that high-potential structures are more complicated, and this complexity — not rushing — is what

[24] As discussed in Section 3.2.2, our measure of maturation is imperfect. For one, it measures elapsed time, but not necessarily the hours spent working on any particular project. In addition, it only measures the time between when the scientist collects her experimental data and when she submits a draft. It does not include the time spent isolating and crystallizing the protein. Anecdotally, crystallization can be the most difficult and lengthly part of the process. Therefore, the estimates above represent the shortening of a *part* of the project lifespan.

drives the lower quality. This motivates our work in the following two sections.

4.4 Competition or Complexity?

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

In this and the following section, we will provide three distinct pieces of evidence which together suggest that complexity alone cannot explain the negative relationship we observe. We start by pointing out the negative relationship between potential and maturation shown in Figure 7. If scientists are agnostic toward priority rewards, but high-potential structures are more complex, then we would expect that scientists spend *longer* on these complex structures. In fact, we find the exact opposite, as discussed in Section 4.3. Researchers spend *less* time on the high-potential structures. This suggests that complexity alone cannot explain the negative relationship between potential and quality.

In general, our estimates of β in Equation 12 will be biased if the conditional independence assumption fails. In this context, the conditional independence assumption requires that our outcome of interest (maturation or quality) is independent of potential, conditional on controls. Therefore, our next strategy is to include controls for structure complexity, in an effort to achieve conditional independence. These controls, which are outlined in Section 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.

Panel B of Table 3 illustrates the effect of adding these complexity controls in Equation 12 when quality is the dependent variable. To start, we note that these controls are powerful predictors of project quality. The R^2 dramatically increases in columns (2) through (5) with the inclusion of these controls. For example, in column (5), the R^2 increases by over a factor of three (going from 0.065 in Panel A to 0.215 in Panel B).

At the same time, the inclusion of these controls does not have a large effect on our estimated coefficients. Comparing Panels A and B in Table 3, we observe that the coefficients remain stable. For example, looking at our quality index outcome in column (5), we see that complexity controls reduce the magnitude of our estimate by just ten percent. Across all four quality outcomes, the coefficients remain negative and statistically significant at the one percent level.

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

4.5 Investigating Structural Genomics Groups

In this section, we contrast structures deposited by structural genomics (SG) groups and those deposited by other researchers, in order to separate the effect of researcher rushing from other omitted factors (in particular, project complexity). As we discuss below, researchers in SG groups are less focused on competing for priority. Therefore, the optimization problem these researchers face in selecting the maturation period is similar to the no competition benchmark of the model, presented in Section 2.2.1. The model predicts that in this case, Proposition 5 should no longer hold. In other words, without competitive incentives, we no longer expect to see a negative relationship between potential and maturation or quality.[25] 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. Inspired by the success of the Human Genome Project, SG groups have a different mission than university and private-sector labs. These groups focus on achieving comprehensive coverage of the protein folding space, and eventually full coverage of the human "proteome," the catalog of all human proteins (Grabowski et al., 2016). Even without solving the structure of every protein, SG groups have achieved broader coverage of the "protein folding space," which has allowed subsequent structures to be solved more easily. For a more complete history of these structural genomics consortia, see Burley et al. 2008; Grabowski et al. 2016. All told, these initiatives have produced nearly 15,000 PDB deposits.

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

We are able to identify SG deposits in our data by looking at the structure authors in the PDB. If the structure was solved by an SG group, that group name will be listed as the last structure

[25] 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).

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

Given these facts, it is not surprising that SG structures are lower-potential on average. This is in line with mission of the SG groups, which seek to provide coverage for less-studied proteins. However, Figure 9 plots the potential distributions for SG and non-SG structures. Here we see that despite the difference in means, the histograms show that the two distributions have overlapping supports. This suggests that we can draw reasonable comparisons between how SG and non-SG structures are impacted by competition and potential.

4.5.2 Analysis of Structural Genomics Consortia

Figure 10 compares the relationship between potential and maturation for both SG and non-SG structures. The two binned scatterplots are constructed separately and overlaid on the same set of axes. Because we bin each series separately, there are the same number of observations in each marker within the same series (but not across series). The fact that the markers do not line up vertically over the x-axis reflects the fact that the two series have different supports.

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

However, the key takeaway from Figure 10 is that there is also a visible difference in slopes. As previously illustrated, the higher-potential non-SG structures are have shorter maturation periods (are completed more quickly). By contrast, the higher-potential SG structures appear to have have slightly *longer* maturation periods.

Figure 11 is isomorphic, but presents the the effects on quality. Across all four quality measures,

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

We formalize the trends shown in Figures 10 and 11 using a differences-in-differences framework. For structure i deposited in year t, we estimate the following regression:

$$Y_{it} = \alpha + \beta P_{it} + \lambda NonSG_{it} + \delta(P_{it} \times NonSG_{it}) + \tau_t + X'_{it}\gamma + \epsilon_{it} \tag{13}$$

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

Table 5 presents the results. Focusing first on column (1) of Panel A, we see that our estimate of β (the coefficient on potential) is positive, reflecting the fact that SG groups spend *longer* in high-potential projects. We also see that our estimate λ (the coefficient on the non-SG indicator) is positive, reflecting the fact that non-SG structures are completed more slowly on average (due to higher rates of associated paper publication). However, our estimate of δ, the interaction between potential and non-SG, is negative and statistically significant. The negative estimate of the δ coefficient suggests that relationship between potential and maturation is more negative for non-SG structures relative to SG structures. In fact, it is large enough to more than offset β, implying that non-SG researchers spend less time on high-potential structures, in contrast with their SG counterparts.

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

Columns (2) to (5) focus on the quality outcomes. Starting with Panel A, the negative estimates of β imply that even among the SG structures, there is a negative relationship between potential and

quality. The positive estimates of λ reflect the fact that they y-intercept of the non-SG structures lies above the SG structures. However, more relevant is where the two series intersect at the minimum value of P (which recall is at about $P = 30$, rather than $P = 0$). If we rescaled our measure of P, the main effect of non-SG would in fact be close to zero, suggesting that quality is similar across two groups at the lowest level of potential.[26]

The estimates of the primary coefficient of interest, δ, are negative across all four quality measures and statistically significant at the one percent level. This implies that the negative relationship between potential and quality is stronger for the non-SG (i.e. more competitive) researchers. Focusing on column (5), we can interpret the the estimated δ coefficient as implying that among the non-SG structures, competition causes high-potential structures to be 0.4 standard deviations lower quality than low-potential structures, relative to SG structures. The magnitudes of the estimates are consistent across all of our quality measures. The inclusion of complexity controls in Panel B does not alter the estimates meaningfully.

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

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

4.6 The Relationship between Competition and Quality

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

However, the model also suggests a solution: we can instrument for competition using project

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

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.

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

$$Y_{it} = \alpha + \beta C_{it} + X'_{it}\gamma + \tau_t + \epsilon_{it} \tag{14}$$

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

However, we also estimate a separate specification, using two-stage least squares and instrumenting for competition using project potential. The first stage regression is identical to Equation 12, with competition (measured as the log number of structures deposited in the same cluster within two years) as the dependent variable. The second stage regression for structure i deposited in year t is given by:

$$Y_{it} = \tilde{\alpha} + \tilde{\beta}\hat{C}_{it} + X'_{it}\tilde{\gamma} + \tilde{\tau}_t + \eta_{it} \tag{15}$$

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

Table 6 shows the results from both of these specifications. Comparing the coefficients of β (in Panel A) and $\tilde{\beta}$ (in Panel B), we see that competition is correlated with shorter maturation periods and lower quality in both specifications. However, as perhaps expected, we see that the estimates in Panel A are attenuated. To interpret the coefficients in Panel B, consider one structure where the expected number of researchers working is 1.25 and another more competitive structure where the expected number of researchers working is 1.5. This can roughly be interpreted as a 25 percentage point increase in the probability of a competitor. The coefficient in column (1) implies this second structure would be completed one to two months faster.[27] The coefficient in column (5) implies the second structure would score 0.4 standard deviations lower using our quality index.

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

[27] $-0.610 \times (\ln 1.5 - \ln 1.25) = 0.11$ years or 1.33 months.

the structure. If the researcher simply wants to understand the protein's function, a lower-quality structural model may be sufficient. However, if a scientist hopes to use a protein structure for structure-based drug design, then a high-quality structure is required. Anderson (2003) suggests that in order to be useful for structure-based drug design, the structures must have a resolution of 2.5Å or lower, and an R-free of 0.25 or lower.[28] While these cutoffs may not be hard-and-fast, they tell us something about the usefulness of a structure given its quality. It is not uncommon for structures to fall below these thresholds. About 35 percent of the non-SG structures in our analysis sample lie below this resolution cutoff. About 45 percent of these same structures lie below the R-free cutoff.

Drugs typically work by binding to proteins, changing the protein's function. The protein that the drug binds to is known as the "target." In an effort to empirically validate these claims, we use DrugBank to link drugs to their protein targets, and these targets to their PDB ID(s). For every structure in the PDB, this allows us to count the number of drugs that target that particular structure. 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 12 shows the relationship between drug development and resolution in a binned scatterplot.[29] Here we plot unstandardized resolution, so recall that lower values correspond to higher quality. We also plot the 2.5Å cutoff for reference. There is a clear positive relationship between higher levels of drug development and lower (i.e., better) resolution. The relationship is nonlinear, with a sharp drop off at around 2.0Å, which is slightly lower (i.e., better) than the 2.5Å cutoff. Panel B repeats this procedure with R-free (again, lower values unstandardized R-free correspond to higher quality). We again see a sharp drop off in drug development at lower quality. Here that drop off occurs at an R-free of about 0.23, which is slightly lower (i.e., better) than the 0.25 threshold proposed by Anderson (2003). Still, taken together with the conventional wisdom from the literature, these figures suggest that a certain level of quality is necessary for drug development. Moreover, this threshold is stringent enough that many of the structures in our data do not meet or surpass it. This suggests that the negative quality effects we measure are large enough to impact downstream drug development.

5 Welfare Implications

Thus far, we have been focused entirely on the positive predictions of the model. Normative conclusions are more difficult to draw. Nevertheless, in the first part of this section, we make the case that researchers cannot easily "fix" low-quality structures, and so the quality effects we measure capture a real inefficiency in the generation of new scientific knowledge. While many low-quality

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

[29] If a structure has been deposited multiple times, we use resolution form the best (i.e., highest-quality) structure. The idea is that a pharmaceutical firm would always use the best structure available. We discuss this in more detail in Section 5.1.

structures are improved over time, offsetting some of the detrimental effects of racing, this comes at a substantial cost. Next, we turn to the question of optimal policy. We show that the current allocation of investment and maturation chosen by racing teams falls short of idealized first-best, but it may represent a constrained second-best allocation. We discuss alternative policies that might improve quality and investment levels in science.

5.1 Will Follow-On Work Fix the Problem?

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

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

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

marginal quality improvement. Therefore, the present discounted value of this improvement is

$$e^{-rm^{IMP}} P \left[Q(m^{IMP}) - Q(m^{C^*}) \right]. \tag{16}$$

The optimal maturation period for the improved structure, m^{IMP^*}, is given by[30]

$$m^{IMP^*} \in \arg\max_{m^{IMP}} \left\{ e^{-rm^{IMP}} P \left[Q(m^{IMP}) - Q(m^{C^*}) \right] \right\} \tag{17}$$

which yields the first-order condition

$$\frac{Q'(m^{IMP^*})}{[Q(m^{IMP^*}) - Q(m^{C^*})]} = r. \tag{18}$$

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

Proof. See Appendix A.1. The intuition is that the present discounted value of improving a project depends primarily on the project's potential (P) and the quality improvement ($Q(m^{IMP^*}) - Q(m^{C^*})$). Both of these are increasing in P, so the effect on the present discounted value is positive. □

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

$$I^{IMP^*} \in \arg\max_{I^{IMP}} \left\{ g(I^{IMP}) e^{-rm^{IMP^*}} P \left[Q(m^{IMP^*}) - Q(m^{C^*}) \right] - I^{IMP} \right\} \tag{19}$$

which yields the first-order condition

$$g'(I^{IMP^*}) = \frac{1}{e^{-rm^{IMP^*}} P \left[Q(m^{IMP^*}) - Q(m^{C^*}) \right]}. \tag{20}$$

This immediately gives us Proposition 6.

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

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

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

To document whether Proposition 6 is true empirically, we need to identify when a project in our analysis sample is re-solved.[31] We are once again able to use the PDB's cluster classification. If we see that a structure in our analysis sample has another structure in its same similarity cluster that was deposited two years or later than the initial structure, we say that structure was re-solved.[32] We use this "two year" rule in an effort to separate contemporaneous work from replications or re-deposits. Panel A of Figure 13 plots the probability a structure is re-solved as a function of project potential. We observe exactly what Proposition 6 predicts — higher P structures are more likely to be re-solved. Scientists are more willing to invest in re-solving these structures because (a) they are more valuable and (b) there is more room for improvement.

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

$$\overline{Q}_{max}(P) = Q(m^{C^*}) + g(I^{IMP^*}) \left[Q(m^{IMP^*}) - Q(m^{C^*}) \right]. \tag{21}$$

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

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

Proof. See Appendix A.1. Intuitively, both $g(I^{IMP^*})$ and $Q(m^{IMP^*}) - Q(m^{C^*})$ are increasing in P. This must at least partially offset the negative relationship between $Q(m^{C^*})$ and P. □

Panel B of Figure 13 tests this proposition. The first series on the plot (the dots) shows the relationship between potential and a structure's initial quality, as in Figure 8. However, the second series (the diamonds) shows the relationship between potential and the structure's *maximum* quality, when looking across all structures within a similarity cluster. The vertical distance between the red and blue series represents the average quality improvement. As predicted by Proposition 7, the relationship between potential and maximum quality is less negative than the relationship between potential and initial quality. In fact, the relationship between potential and maximum quality is U-shaped. The intuition is that at low values of P, the incentives to re-solve are low, but the initial quality is high. At high values of P, the incentives to re-solve are high. This leads to high maximum quality at the extremes of the potential distribution, and lower maximum quality in the middle of the distribution.

Returning to our concerns about project complexity in Section 4.4, it is comforting to see that the maximum quality values at the top end of the potential distribution are nearly as high as the

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

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

maximum quality values at the bottom of the potential distribution, because it suggests that high quality is possible for these high-potential structures. If the negative relationship between potential and initial quality were driven purely by structure complexity, we might expect that it is simply impossible to solve these high-potential structures at the same level of quality.[33]

Together, Panels A and B of Figure 13 suggest that there are three distinct sources of welfare loss associated with rushing in structural biology. First, there is the loss of structure quality, which translates to lost downstream innovation. However, Panel B shows that without taking into account the subsequent re-deposits, we will overestimate the magnitude of this lost quality as much of it (particularly for the highest potential structures) is made up in future work. Second, there is the time cost associated with the re-deposits. While much of the lost structure quality is eventually reclaimed via follow-on work, this takes additional time. Finally, there is the monetary cost associated with re-solving the same structures. The PDB estimates that the average cost to replicate a structure is about $100,000 (Sullivan et al., 2017).

5.2 Optimal Policy

5.2.1 The Infeasible First Best

We start our optimal policy analysis by considering how equilibrium maturation and investment that arises from researchers competing for priority (i.e., m^{C^*} and I^{C^*}) compares to the outcome preferred by an unconstrained social planner. In this setting, an unconstrained social planner would like to dictate both investment (I) and maturation (m) to researchers. The social planner's objective differs from an individual researcher's objective in two ways: first, the social planner only cares that at least one researcher successfully starts the project. If both researchers start the project, the planner is indifferent as to which researcher completes the project first, and the second (replicated) structure adds no additional social value. This wedge is similar to the inefficiency identified by Dasgupta and Maskin (1987). Second, consistent with the notion of research generating positive spillovers, the social value of a given project is greater than the private value. We operationalize this by assuming that the social planner's PDV of the project at completion is $e^{-rm}kPQ(m)$, rather than $e^{-rm}\overline{\theta}PQ(m)$ or $e^{-rm}\underline{\theta}PQ(m)$ (the first- and second-place researcher's private PDV, respectively). We further assume that k is large relative to $\overline{\theta}$ and $\underline{\theta}$ (we put more formal bounds on k in the analysis below). Putting these facts together, we have the social planner's objective function:

$$\max_{m,I} \left\{ \underbrace{\left(1-(1-g(I))^2\right)}_{\text{probability at least one researcher successfully starts}} \cdot \underbrace{e^{-rm}kPQ(m)}_{\text{social PDV of project}} - \underbrace{2I}_{\text{investment costs}} \right\}. \tag{22}$$

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

Contrast this with the individual researcher's objective function (Equation 9, reproduced and slightly re-arranged below):

$$\max_{m_i, I_i} \left\{ \underbrace{g(I_i)}_{\text{probability } i \text{ successfully starts}} \cdot \underbrace{e^{-rm_i}\left[\overline{\theta} - \frac{1}{2}g(I_j)(\overline{\theta} - \underline{\theta})\right] PQ(m_i)}_{i\text{'s expected private PDV of project}} - \underbrace{I_i}_{i\text{'s investment cost}} \right\}. \quad (23)$$

The socially optimal value of m, denoted m^{SP^*}, is defined by the first-order condition of Equation 22 with respect to m:

$$\frac{Q'(m^{SP^*})}{Q(m^{SP^*})} = r. \quad (24)$$

Notice that this is identical to the first-order condition which defines the optimal value of m in the absence of competition (m^{NC^*}, see Equation 4). Therefore, we know that $m^{SP^*} > m^{C^*}$. In other words, the social planner wants projects to mature for longer than researchers will allow them to in a competitive environment. This happens precisely because the social planner — unlike the individual researcher — does not care who finishes the project first. Concerns over priority distort the individual researcher's choice of m away from the social optimum.

The socially optimal value of I, denoted I^{SP^*}, is defined by the first-order condition of Equation 22 with respect to I:

$$g'(I^{SP^*}) = \frac{1}{e^{-rm^{SP^*}} kPQ(m^{SP^*})(1 - g(I^{SP^*}))}. \quad (25)$$

Comparing this equation with the first-order condition that defines I^{C^*} (Equation 10), we can see that if k is sufficiently large,[34] then $I^{SP^*} > I^{C^*}$. Intuitively, if the social planner values the project sufficiently more than the researcher, the social planner will want the researcher to invest more than the privately optimal level.

The empirical evidence supports the theoretical argument that individual researchers distort their behavior away from the social optimum. More specifically, Equation 24 implies that if we were at the first best, then the relationship between potential and quality should be flat. Instead, we observe a negative relationship between potential and quality, consistent with researchers distorting their behavior in an effort to complete their projects first.

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

The social planner cannot realistically dictate I and m for each project. Monitoring the progress of every scientific team as they work on their projects requires too much information to be feasible. Instead, a more reasonable lever for the social planner might be $\overline{\theta}$ or $\underline{\theta}$, the share of credit allocated to the first and second-place team, respectively. While the literature has often assumed that priority races are winner-take-all, implying that $\underline{\theta} = 0$ (for example, Merton (1957); Fudenberg et al. (1983);

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

Bobtcheff et al. (2017)) empirical evidence suggests that this is not the case. Hill and Stein (2020) find that in structural biology, winning teams involved in priority races receive about 55 percent of the credit (as measured by citations) — a far cry from 100 percent. While that same paper provides survey evidence to suggest that structural biologists are more pessimistic about the costs of being scooped (the surveyed authors estimated the winning paper would accrue about 70 percent of the total citations), the 100 percent benchmark does not appear to be correct in this setting.

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

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

Other journals have released similar policies.[35] In light of these changes, the distribution of credit is a particularly interesting and relevant policy tool to study. However, the precise way in which we allow the social planner to manipulate the distribution of credit will have different implications for optimal policy. We consider two cases in turn.

[35]For example, in January 2018, *PLOS Biology* released a statement reading, "scientific research can be a cutthroat business, with undue pressure to publish quickly, first, and frequently. The resulting race to publish ahead of competitors is intense and to the detriment of the scientific endeavor. Just as summiting Everest second is still an incredible achievement, so too, we believe, is the scientific research resulting from a group who have (perhaps inadvertently) replicated the important findings of another group. To recognize this, we are formalizing a policy whereby manuscripts that confirm or extend a recently published study ("scooped" manuscripts, also referred to as complementary) are eligible for consideration at *PLOS Biology* (The PLOS Biology Staff Editors, 2018). In November 2018 the editor of *Cell Systems* released a statement saying "*Cell Systems* thinks it is valuable — as well as simply humane — to welcome strong experimental studies that are "scooped" (Justman, 2018).

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

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

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

Proof. Recall that the social planner would like the researcher to behave as if there is no competition. In other words, $m^{SP^*} = m^{NC^*}$. Intuitively, if we equate the rewards for the first- and second-place researcher, we have eliminated competition, and so researchers will let their projects mature optimally. However, this results in investment below the socially optimal level. See Appendix A.1 for more detail. □

By setting $\overline{\theta} = V - \overline{\theta} = \frac{V}{2}$, the social planner is able to select the optimal maturation period, but investment is too low. Next, we will show that as the social planner raises $\overline{\theta}$ — making priority rewards more lopsided — maturation periods become shorter, but investment may increase. This sets up a tradeoff for the social planner: more unequal priority rewards lead to shorter maturation periods (moving us away from the optimal maturation level), but potentially higher investment levels (moving us closer to the optimal investment level). This implies that optimal priority rewards may be unequal. Proposition 8 below formalizes this logic.

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

Proof. See Appendix A.1. □

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

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

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

Proof. Setting $\overline{\theta} = \underline{\theta}$ ensures that we achieve the socially optimal maturation, as shown in Lemma 2. Allowing $\overline{\theta} + \underline{\theta}$ to be unconstrained means we can induce the appropriate amount of investment. Intuitively, if the social value of a project is high, then $\overline{\theta} + \underline{\theta}$ will be larger. See Appendix A.1 for details. □

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

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

5.2.3 An Alternative Policy: Ending Races Early

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

higher than in the standard competitive case, because the projects are more valuable when allowed to fully mature.

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

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

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

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. While this negative relationship is inconsistent with an idealized first best, where a social planner can dictate how much investment researchers dedicate to projects and how long they let these projects mature, it *not inconsistent* with a more realistic constrained second best, where the social planner can only dictate how credit is shared between first- and second-place researchers.

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

References

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

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.

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 Econommetric Models: An Introduction," 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.

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.

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 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.

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

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

Goodsell, David S., "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.

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

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

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

Hill, Ryan and Carolyn Stein, "Scooped! Estimating Rewards for Priority in Science," *Working Paper*, 2020.

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

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

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

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

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

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

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

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

Marder, Eve, "Scientific Publishing: Beyond Scoops to Best Practices," *eLife*, 2017, *6*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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.

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

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

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

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

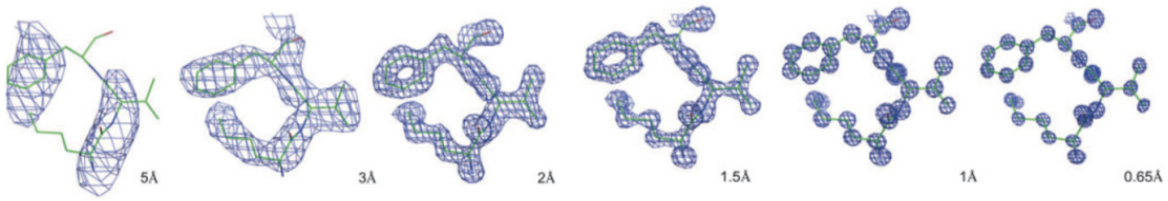
Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan, "Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation," *Science*, 2020, *367* (6483), 1260–1263.

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

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

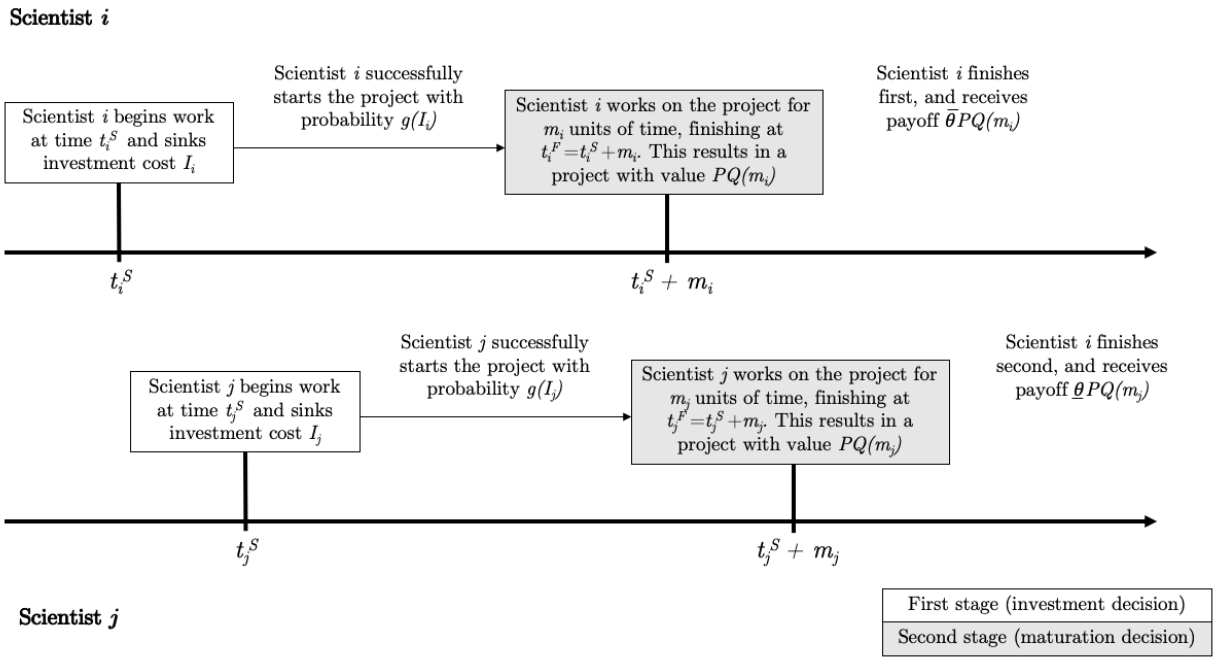
Figures and Tables

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



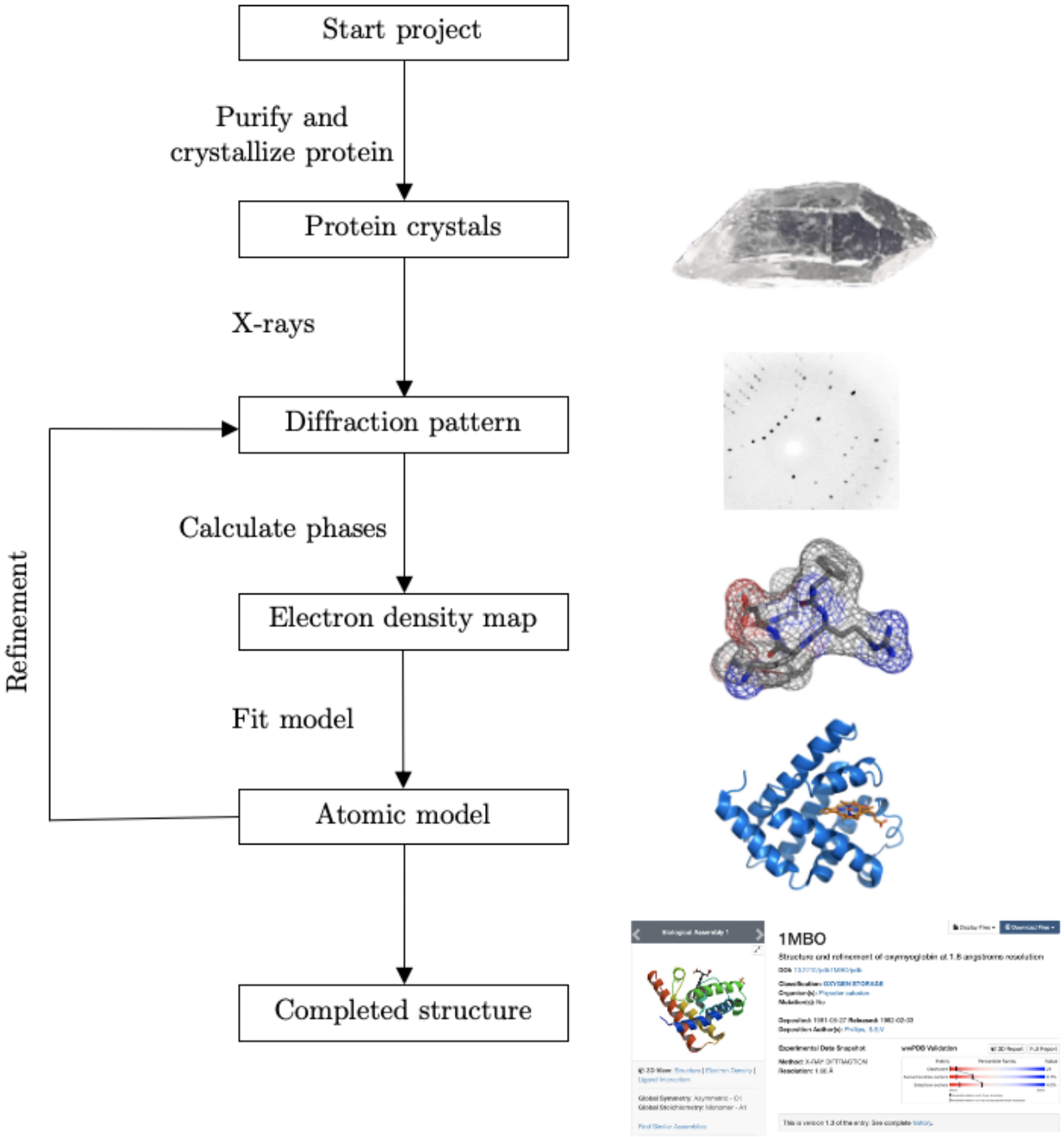
Notes: This figure shows the electron density maps from a fragment of the triclinic lysozyme (PDB ID 2VB1) at different refinement resolutions. The Angstrom (Å) values measure the smallest distance between crystal lattice planes that can be detected in the experimental data. Lower values correspond to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

Figure 2: Model Summary



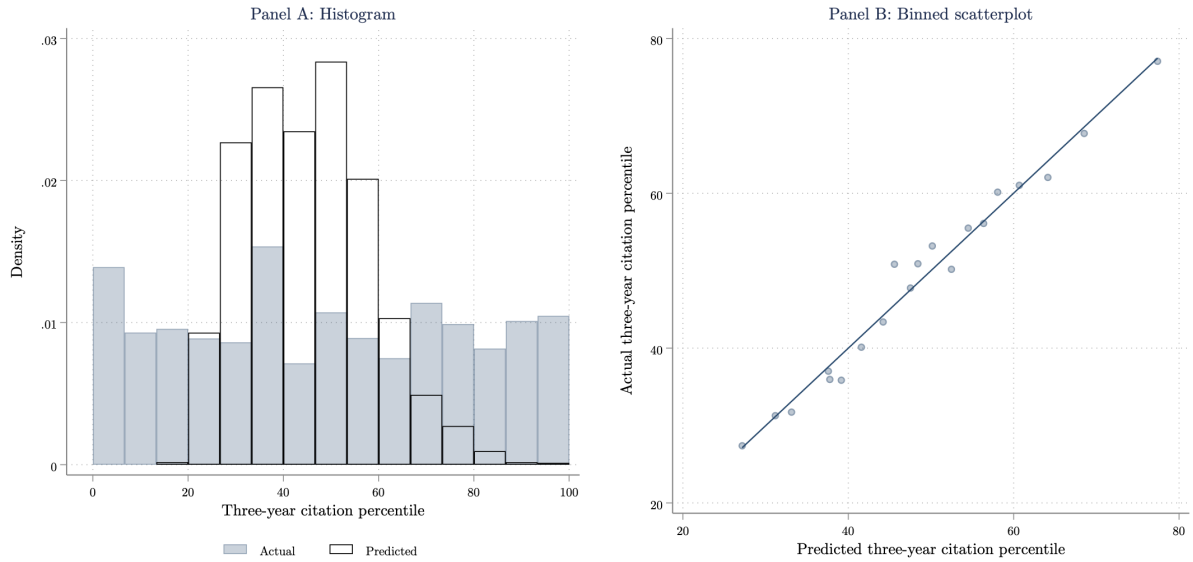
Notes: This figure summarizes the setup of the model described in the text.

Figure 3: Summary of the X-Ray Crystallography Process



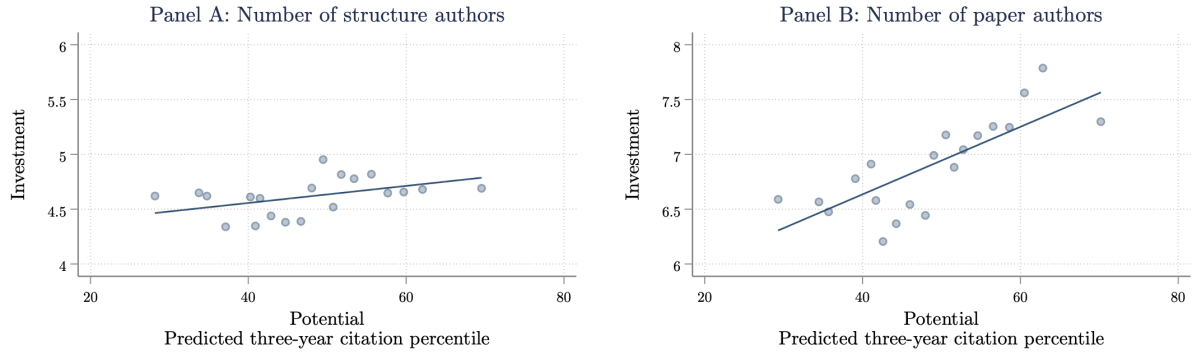
Notes: This figure summarizes the process of solving a protein structure via x-ray crystallography. The images in this figure were taken from Thomas Splettstoesser (www.scistyle.com) and rendered with PyMol based on PDB ID 1MBO.

Figure 4: LASSO Validation



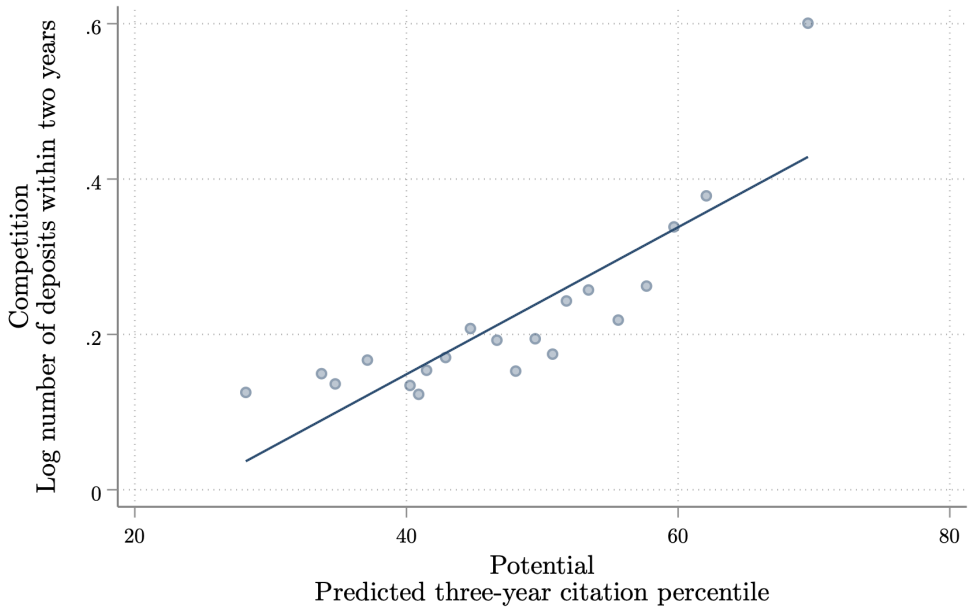
Notes: Panel A of this figure plots the distribution of actual and predicted potential. Panel B presents a graph of actual versus predicted potential as a binned scatterplot. In both panels, potential is measured by the percentile of the structure's three-year citation count. To construct this binned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

Figure 5: The Effect of Potential on Investment



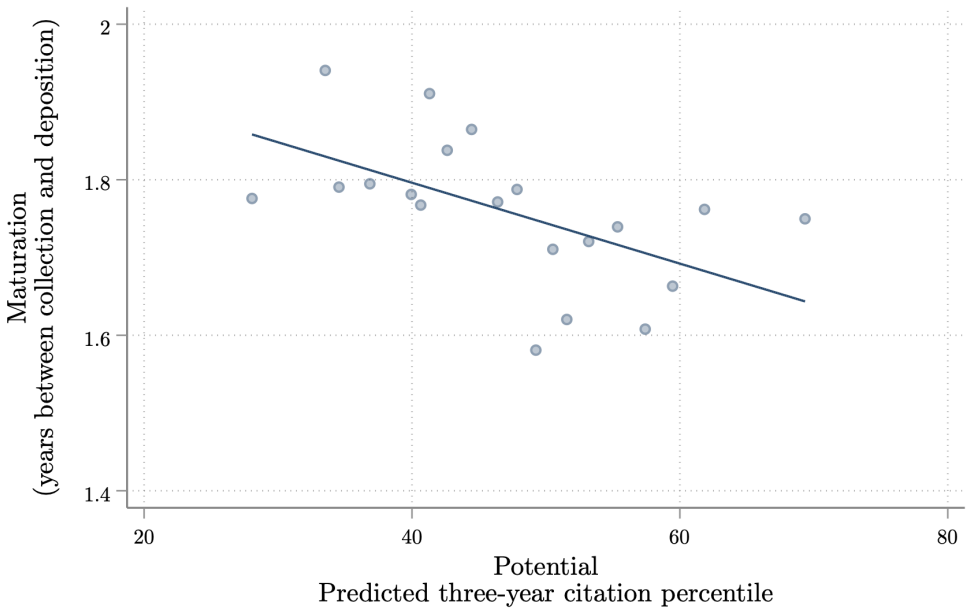
Notes: This figure plots the relationship between potential and investment, testing Proposition 4 of the model. Potential is measured as the predicted three-year citation percentile. Investment is measured as either the number of structure authors or number of paper authors. The plot is presented as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

Figure 6: The Effect of Potential on Competition



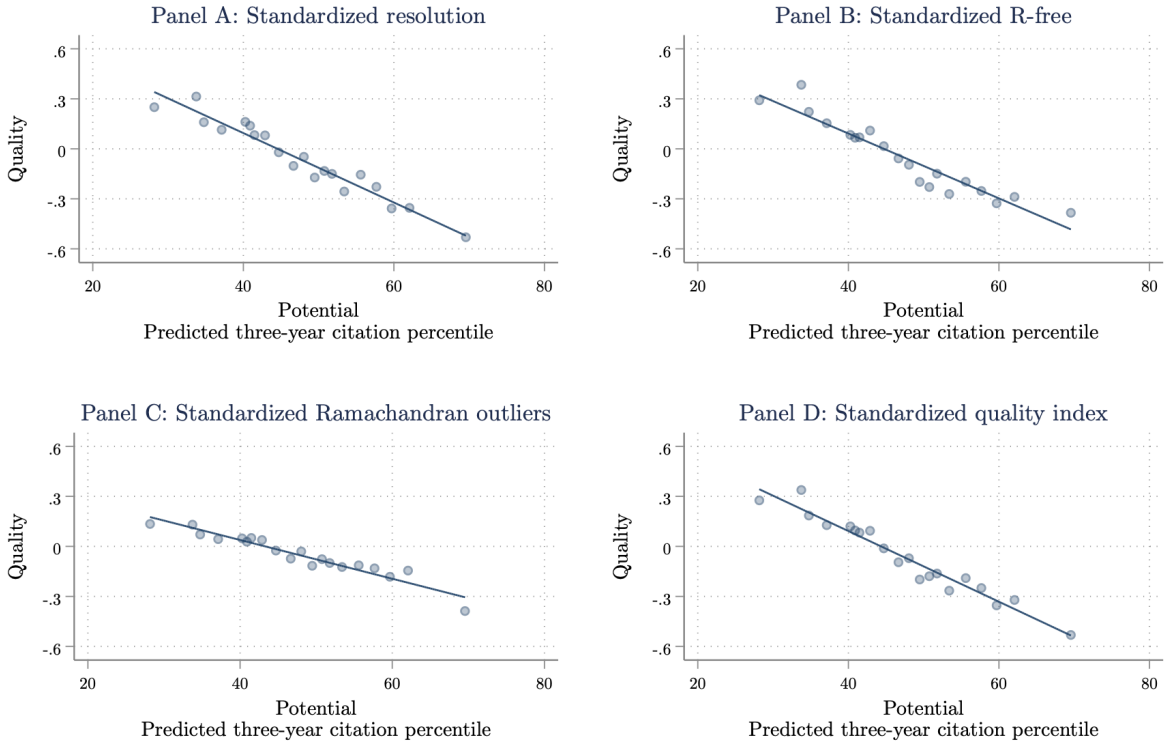
Notes: This figure plots the relationship between potential and competition, testing Proposition 4. Potential is measured as the predicted three-year citation percentile. Competition is measured as the log number of deposits that appear in the 100 percent similarity cluster within two years of the first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

Figure 7: The Effect of Potential on Maturation



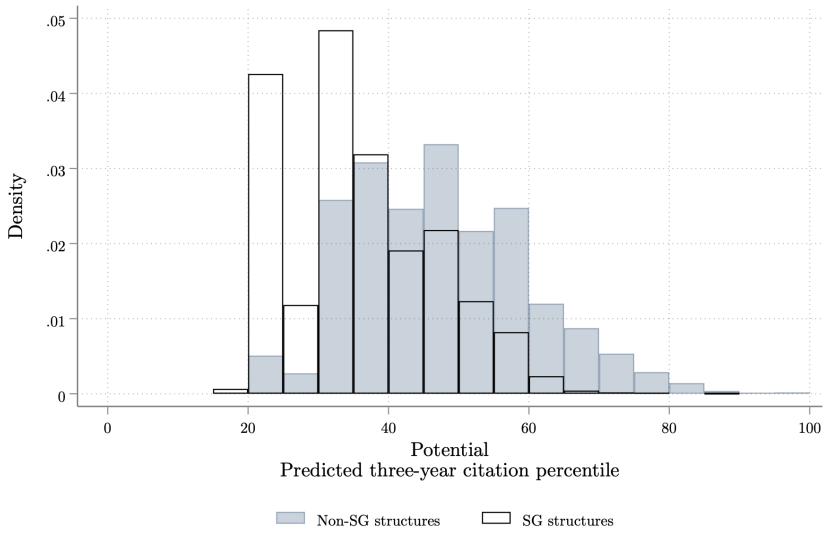
Notes: This figure plots the relationship between potential and maturation, testing Proposition 5. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plot is presented as a binned scatterplot.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

Figure 8: The Effect of Potential on Quality



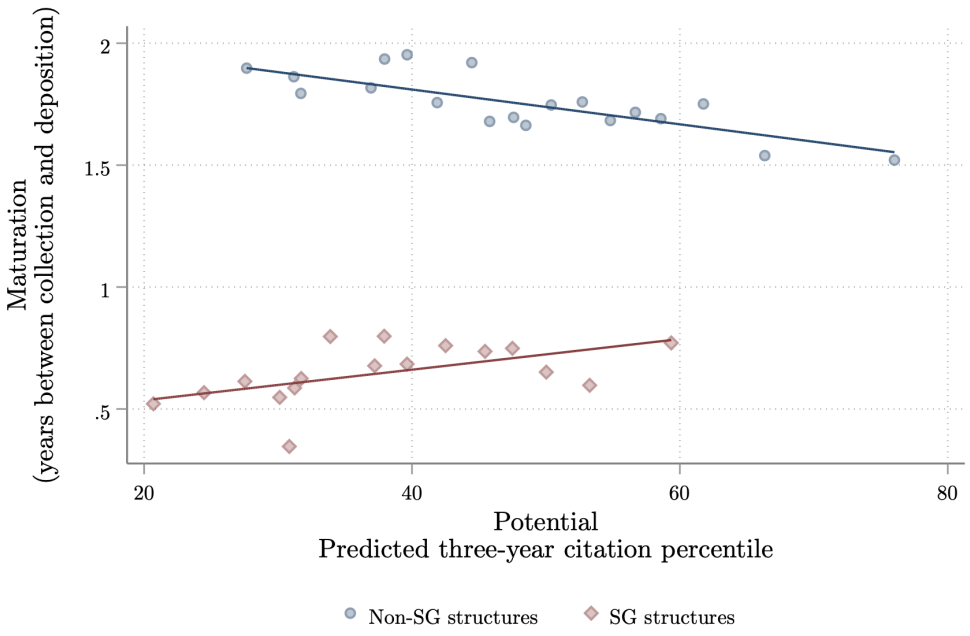
Notes: This figure plots the relationship between potential and quality, testing Proposition 5. Potential is measured as the predicted three-year citation percentile. Quality is measured by our four standardized quality measures described in detail in Section 3.2.1. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we first residualize potential and quality with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of quality against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. The sample is the full analysis sample as defined in the text, excluding SG deposits.

Figure 9: Potential Distributions by Structural Genomics Status



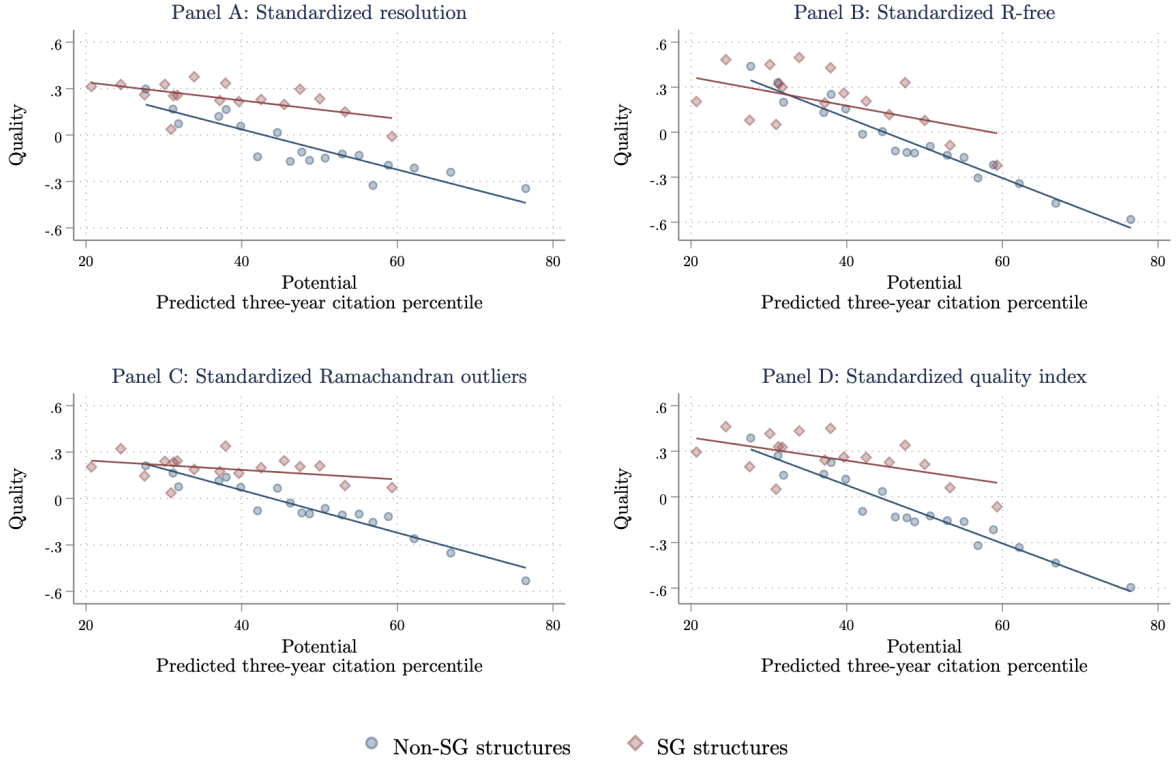
Notes: This figure plots the distribution of potential (measured by predicted three-year citation percentile) for both non-SG and SG structures. The sample is all structures in the analysis sample.

Figure 10: The Effect of Potential on Maturation by Structural Genomics Status



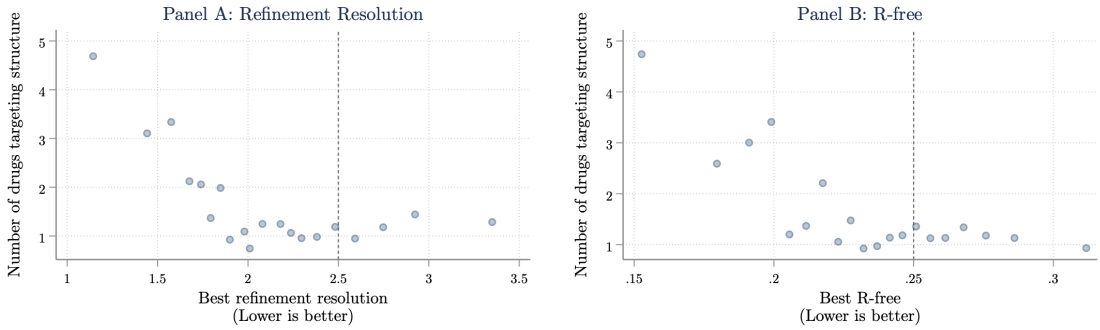
Notes: This figure plots the relationship between potential and maturation, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Maturation is measured by the number of years between the deposition and collection dates. The plots are presented as two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

Figure 11: The Effect of Potential on Quality by Structural Genomics Status



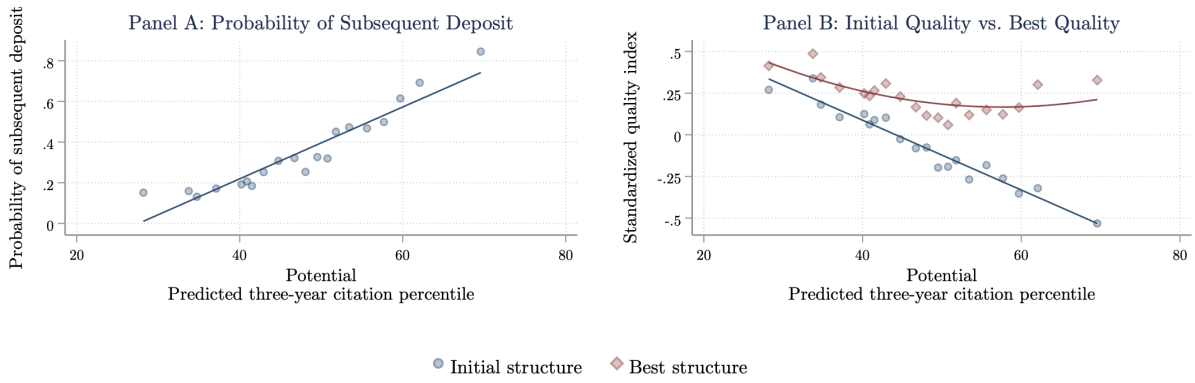
Notes: This figure plots the relationship between potential and quality, split by non-SG and SG structures. Potential is measured as the predicted three-year citation percentile. Quality is measured by our four standardized quality measures described in detail in Section 3.2.1. The plots are presented as two separate binned scatterplots, overlaid on the same axes. To construct these binned scatterplots, we first residualize potential and quality with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of quality against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. We repeat this procedure separately for the SG and non-SG structures, but plot the resulting series on the same axes. As a result, there are the same number of observations within each point in the same series. The sample is the full analysis sample.

Figure 12: Relationship between Structure Quality and Drug Development



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

Figure 13: Subsequent Structure Deposits and Maximum Structure Quality



Notes: This figure plots the relationship between potential and probability of subsequent deposition (Panel A) and the relationship between potential and initial quality and best quality (Panel B). A subsequent deposit is defined as a deposit in the same 100 percent cluster that is deposited in the PDB more than two years after the first deposit. Quality is measured using our quality index described in detail in Section 3.2.1. The plots are presented as binned scatterplots. To construct these binned scatterplots, we first residualize the dependent variable (indicator for subsequent deposit, the initial quality, or the best quality) and potential with respect to a set of deposition year indicators. We then divide the sample into 20 equal-sized groups based on the ventiles of the potential measure, and plot the mean of the dependent variable against the mean of potential in each bin. Finally, we add back the mean quality to make the scale easier to interpret after residualizing. The sample is the full analysis sample.

Table 1: Summary Statistics: Full Sample versus Analysis Sample

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

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

Table 2: The Effect of Potential on Investment and Competition

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

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

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

Table 3: The Effect of Potential on Maturation and Quality

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

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

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

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

Table 5: The Effect of Potential on Maturation and Quality, by Structural Genomics Status

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

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

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

Table 6: The Effect of Competition on Maturation and Quality

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

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

A Theory Appendix

A.1 Proofs of Propositions

Proof of Proposition 1.

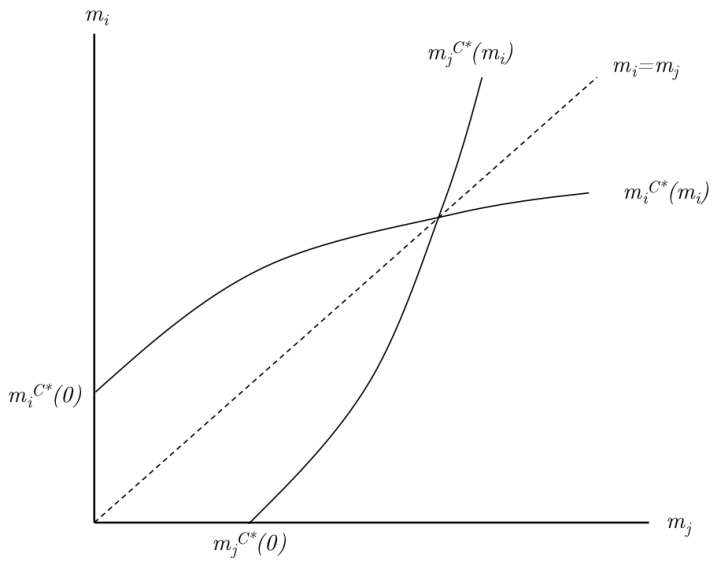
First, we will expand on how we derive the first-order condition for $m_i^{C^*}$ (Equation 7). Taking the derivative of Equation 6 with respect to m_i and setting it equal to zero yields:

$$\frac{Q'(m_i^{C^*})}{Q(m_i^{C^*})} = r - \frac{\frac{\partial \pi}{dm_i}(\overline{\theta} - \underline{\theta})}{\pi(m_i, m_j)\overline{\theta} + (1 - \pi(m_i, m_j))\underline{\theta}}. \tag{26}$$

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 26 above yields Equation 7 in the text. However, if m_i is much larger than m_j (i.e., if $m_i > m_j + \Delta$), then $\frac{\partial \pi}{\partial m_i} = 0$ and Equation 26 collapses to the no-competition case, i.e., Equation 4. We will return to this caveat, but for now we will assume m_i is close to m_j.

Equation 7 implicitly defines $m_i^{C^*}(m_j)$ as a function of m_j and parameters. If we can show that (i) $m_i^{C^*}(0) > 0$ and (ii) $\frac{dm_i^{C^*}}{dm_j} \in (0, 1)$, then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $m_i^{C^*}(m_j)$ and $m_j^{C^*}(m_i)$ will only cross the $m_i = m_j$ line once.

Figure A.1: Maturation Best Response Functions



To show (i), plug $m_j = 0$ into Equation 7. This results in an equation that implicitly defines a unique $m_i^{C^*}(0) > 0$. To show (ii), we can totally differentiate equation 7 with respect to m_j. For notational ease, define $\zeta \equiv \Delta\left(\frac{2\overline{\theta} - g(\overline{\theta} - \underline{\theta})}{g(\overline{\theta} - \underline{\theta})}\right)$, and note that $\zeta > 0$. Gathering terms and rearranging,

we have that

$$\frac{dm_i^{C^*}}{dm_j} = \left[\underbrace{\left(\frac{-Q(m_i^{C^*})Q''(m_i^{C^*}) + Q'(m_i^{C^*})^2}{Q(m_i^{C^*})^2} \right) \left(\zeta + m_j - m_i^{C^*} \right)^2}_{>0} + 1 \right]^{-1} \in (0,1). \tag{27}$$

Next, we confirm that the second-order conditions hold. Differentiating the objective function (Equation 6) twice with respect to m_i and evaluating at $m_i = m_j = m^{C^*}$ yields

$$Pe^{-rm_i} \left[Q''(m^{C^*}) - Q'(m^{C^*}) \left(r + \frac{1}{\zeta} \right) \right] < 0. \tag{28}$$

Therefore, $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ is a local optimum. Plugging m^{C^*} in for both m_i and m_j (and assuming that $I_i = I_j = I^{C^*}$) in Equation 7 yields the expression in Proposition 1.

However, as a final check, we need to confirm that this is also a global optimum. Note that Equation 8 tells us that as $\Delta \to 0$, $m_i^{C^*} \to 0$. This will yield a payoff of zero for researcher i. This cannot be researcher i's best response, because there is always a $1-g$ probability that her competitor did not enter. Therefore, she would be better off selecting $m_i = m^{NC^*}$ and hoping that her competitor fails to enter the project. To map this intuition to the math, note that we are now considering a case where $m_i > m_j + \Delta$, and so we the relevant first-order condition is now Equation 4.

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

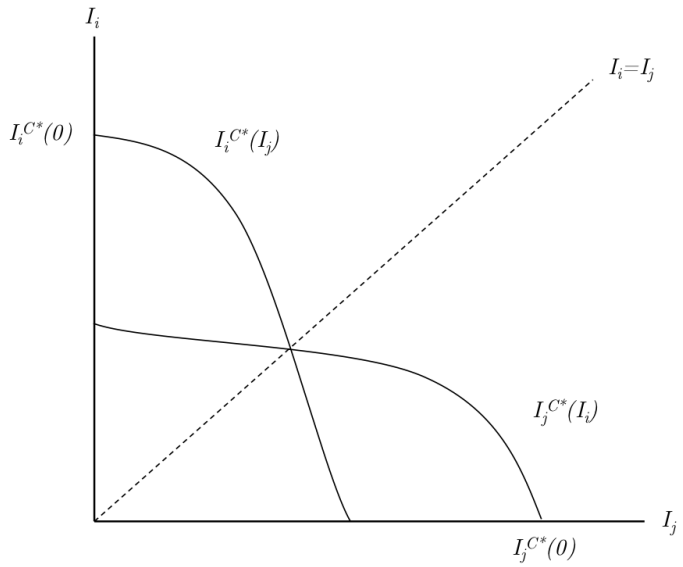
$$e^{-rm^{C^*}} PQ(m^{C^*}) \left[(1 - \frac{g}{2})\overline{\theta} + \frac{g}{2}\underline{\theta} \right] > e^{-rm_i^{NC}} PQ(m_i^{NC^*}) \left((1-g)\,\overline{\theta} + g\underline{\theta} \right). \tag{29}$$

Because m^{C^*} is increasing in Δ, this defines a lower bound on Δ such that this equation will hold. Therefore, $m_i^{C^*} = m_j^{C^*} = m^{C^*}$ is a symmetric pure strategy Nash equilibrium as long as Δ is sufficiently large. Moreover, this is the only possible pure strategy Nash equilibrium. To see this, note that if $|m_i - m_j| < \Delta$, then the first-order condition in Equation 7 applies and we have the equilibrium defined by $m_i^{C^*} = m_j^{C^*} = m^{C^*}$. Alternatively, if $|m_i - m_j| \geq \Delta$, then the first-order condition defined by Equation 4 applies. But this implies that $m_i^* = m_j^* = m^{NC^*}$, which violates the assumption that $|m_i - m_j| \geq \Delta$. Therefore, if Δ is below some threshold, the Nash equilibrium must be mixed. We will focus on the pure strategy case throughout the remainder of the paper.

Proof of Proposition 2.

Equation 10 implicitly defines $I_i^{C^*}(I_j)$ as a function of I_j, $m_i^{C^*}$ (which depends on I_j), and parameters. If we can show that (i) $I_i^{C^*}(0) > 0$ and (ii) $\frac{dI_i^{C^*}}{dI_j} < 0$ then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $I_i^{C^*}(I_j)$ and $I_j^{C^*}(I_i)$ will only cross the $I_i = I_j$ line once.

Figure A.2: 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 10 with respect to I_j. Gathering terms and rearranging, we have that

$$\frac{dI_i^{C^*}}{dI_j} = \frac{e^{-rm_i^{C^*}} P \left[\left(rQ(m_i^{C^*}) - Q'(m_i^{C^*}) \right) \frac{dm_i^{C^*}}{dI_j} + Q(m_i^{C^*}) g'(I_j)(\overline{\theta} - \underline{\theta}) \right]}{g''(I_j) \left[e^{-rm_i^{C^*}} PQ(m_i^{C^*}) \left(\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta}) \right) \right]^2} < 0 \tag{30}$$

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

Proof of Proposition 3.

Looking at Equation 7, the left hand side is decreasing in m^{C^*}. Looking at the right hand side, we see it is increasing in $g(I^{C^*})$. For the equality to hold as $g(I^{C^*})$ increases, it must be the case that m^{C^*} decreases, i.e., that $\frac{dm^{C^*}}{dg(I^{C^*})} < 0$. Because $Q(m)$ is increasing, this also implies that $\frac{dQ(m^{C^*})}{dg(I^{C^*})} < 0$.

Proof of Proposition 4.

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

$$e^{-rm^{C^*}} PQ(m^{C^*}) \left[\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right]. \tag{31}$$

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

Proof of Proposition 5.

See main text.

Proof of Lemma 1.

Let $\Delta Q = Q(m^{IMP^*}) - Q(m^{C^*})$ denote the realized quality improvement. The derivative of the present discounted value of a project improvement (Equation 16) with respect to project potential P is given by:

$$-re^{-rm^{IMP^*}} \frac{dm^{IMP^*}}{dP} P\Delta Q + e^{-rm^{IMP^*}} \Delta Q + e^{-rm^{IMP^*}} P \frac{d\Delta Q}{dP}. \tag{32}$$

The first term represents the change in discounting due to the effect of P on m^{IMP^*}, the second term represents the direct effect of shifting P, and the final term represents the change in the quality improvement, via the effect of P on m^{IMP^*} and m^{C^*}. Totally differentiating Equation 18 with respect to P and rearranging yields:

$$\frac{dm^{IMP^*}}{dP} = \frac{rQ'(m^{C^*}) \frac{dm^{C^*}}{dP}}{rQ'(m^{IMP^*}) - Q''(m^{IMP^*})} < 0 \tag{33}$$

where we can sign the expression by noting that $\frac{dm^{C^*}}{dP}$ is negative, as shown in Proposition 5. Next, we can re-write Equation 18 as

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

Taking the derivative of this equation with respect to P yields

$$\frac{d\Delta Q}{dP} = \frac{Q''(m^{IMP^*})}{r} \cdot \frac{dm^{IMP^*}}{dP} > 0 \tag{34}$$

due to the concavity of $Q(\cdot)$. Together, these two derivatives allow us to unambiguously show that the expression in Equation 32 is positive.

Proof of Proposition 6.

See main text.

Proof of Proposition 7.

Taking the derivative of Equation 21 with respect to P yields

$$\frac{d\overline{Q}_{max}}{dP} = \frac{dQ(m^{C^*})}{dP} + g'(I^{IMP^*})\frac{dI^{IMP^*}}{dP}\Delta Q + g(I^{IMP^*})\frac{d\Delta Q}{dP}. \tag{35}$$

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

Proof of Lemma 2.

Plugging $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ into Equation 7, we recover Equation 4, which defines both the no-competition maturation period and the social planner's optimal maturation period. Plugging $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ and $m = m^{SP^*}$ into Equation 10, we have

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

Comparing this to Equation 25, we see that as long as k is sufficiently large (in this case, as long as $k > \frac{V/2}{(1-g(I^{SP^*}))}$), then $I^{SP^*} > I^{C^*}$.

Proof of Proposition 8.

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

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

and

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

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

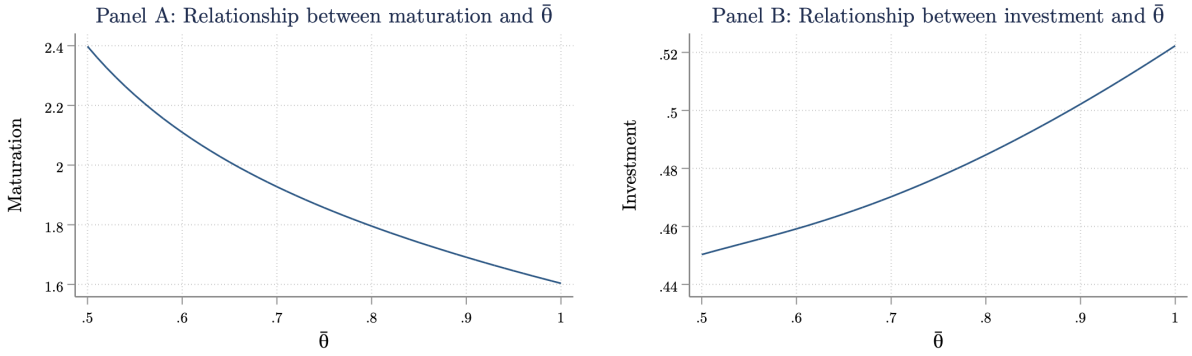
the indirect effect via m ($\frac{\partial I^{C^*}}{\partial m^{C^*}} \cdot \frac{dm^{C^*}}{d\bar{\theta}}$) dominates.

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

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

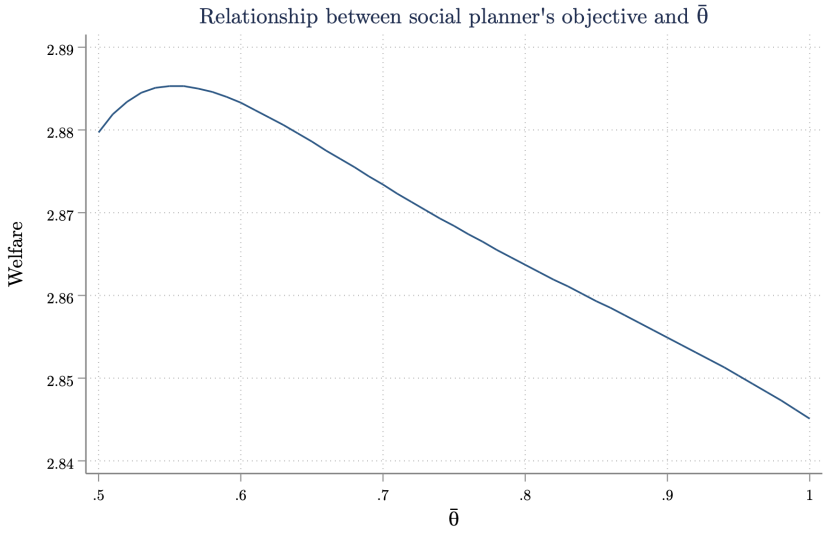
Then, we can numerically compute $\frac{dm^{C^*}}{d\bar{\theta}}$ and $\frac{dI^{C^*}}{d\bar{\theta}}$. We show these below. This results in $\frac{dm^{C^*}}{d\bar{\theta}} < 0$ and $\frac{dI^{C^*}}{d\bar{\theta}} > 0$.

Figure A.3: Numerically calculated $\frac{dm^{C^*}}{d\bar{\theta}}$ and $\frac{dI^{C^*}}{d\bar{\theta}}$



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

Figure A.4: Welfare as a function of $\overline{\theta}$



Proof of Proposition 9.

As long as $\overline{\theta} = \underline{\theta}$, then $m^{C^*} = m^{SP^*}$, as shown in the proof of Proposition 8. To achieve $I^{C^*} = I^{SP^*}$, we plug $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ and $m = m^{SP^*}$ into Equation 2, and equate this with Equation 25:

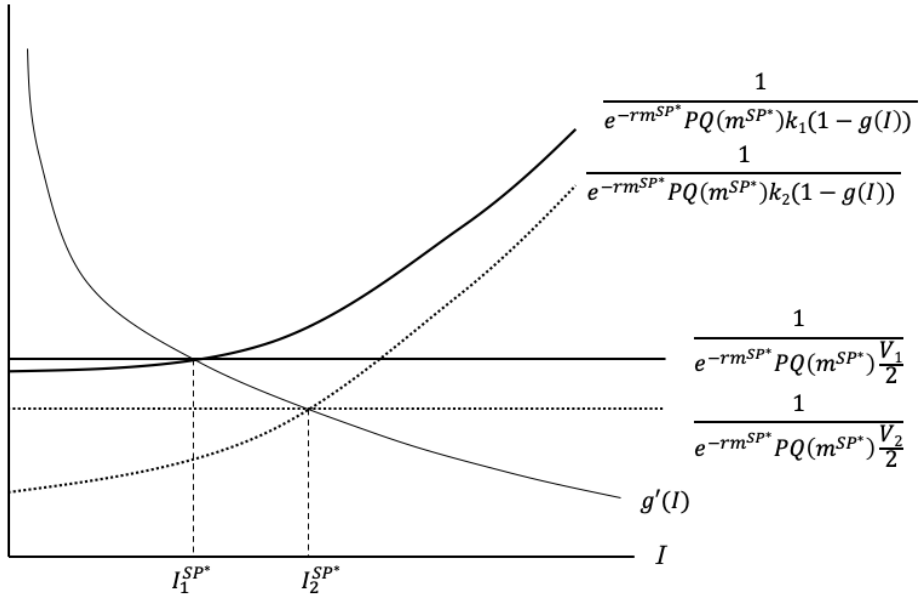
$$\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V/2)} = \frac{1}{e^{-rm^{SP^*}} kPQ(m^{SP^*})(1 - g(I^{SP^*}))}.$$

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

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

So we can recover the first best if $\overline{\theta} = \underline{\theta} = k(1-g(I^{SP^*}))$. Figure A.5 below helps illustrate that $\overline{\theta} = \underline{\theta} = \frac{V}{2}$ is increasing in k. Suppose $k = k_1$. To achieve $I^{C^*} = I^{SP^*}$, we need $\frac{1}{e^{-rm^{SP^*}} k_1 PQ(m^{SP^*})(1-g(I))}$ to intersect both $g'(I)$ and $\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V_1/2)}$, which occurs at $I = I_1^{SP}$ in Figure A.5. However, if we increase k from k_1 to k_2, then $\frac{1}{e^{-rm^{SP^*}} k_2 PQ(m^{SP^*})(1-g(I))}$ shifts down (shown by a dotted line). To maintain this intersection, then $\frac{1}{e^{-rm^{SP^*}} PQ(m^{SP^*})(V_2/2)}$ must also shift down (again shown by a dotted line), which implies that $V_2 > V_1$.

Figure A.5: Achieving Optimal Investment



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.[36]

We access the data directly from the RCSB Custom Report Web Service.[37] The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

- Structure Summary: structure ID, structure title, structure authors, deposit date, release date, experimental technique, classification, macromolecule type, molecular weight, residue count, and atom site count.
- Citation: PubMed ID, publication year, and journal name.

[36] http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction
[37] https://www.rcsb.org/pdb/results/reportField.do

- 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[38] provided by the European Bioinformatics Institute (EMBL-EPI). Data for this study was downloaded on October 25, 2019 and merged using the standard PDB structure identifiers.

Many of the variables we use in the analysis, such as predicted citations, are calculated at the paper level. However 20 percent of PDB-linked papers have more than one structure, with an average of 1.5 structures per paper. Because each linked structure has a unique set of quality metrics, it is difficult to ascribe paper-level characteristics to any one of the individual structures. Our main analysis sample therefore drops all structures linked to multi-structure papers. Since about 30% of deposits are never published, we make a similar restriction for groups of structure deposits that appear to have been part of the same unpublished project. We group unpublished structures into the same "project" if the deposits have the same first and last PDB structure author and share the same release date. Unpublished projects with more than one structure are dropped to mirror the single-structure paper restriction.

A further complication of the PDB data is that cluster groupings are defined at a level of granularity that is smaller than the structure or article level. Proteins are composed of "chains" of amino acids, and large proteins are often characterized in the PDB as a set of distinct chains. Further, chains of amino acids are often grouped as "entities", and many proteins are combinations of two or more entities. This is relevant to our sample construction because the BLAST similarity algorithm clusters at the entity level rather than the protein level. In particular, our main analysis sample includes only "priority" structure deposits, meaning that the PDB entry was the first to produce a structure for a given entity. In practice, we keep any structure that has at least one entity that is the first deposit among all other entities that are 100 percent similar according to the BLAST algorithm. This means that in some cases, only one part of the structure is truly a novel discovery, but these deposits still represent important contributions for which scientists often compete to publish first.

Some relevant protein characteristics are assigned at the entity, rather than the structure level. For example, we use gene-protein linkages as an input to the predicted citation LASSO model described in Section 4.1. The PDB data assigns gene linkages at the entity level, meaning some proteins (9.4 percent) have multiple gene linkages. To simplify the citation prediction model, we assign a single gene-linkage to the full protein by taking the modal gene name amongst the protein entities and breaking ties alphabetically. Similarly, some structures are complexes of entities from

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

different organisms (e.g. a human protein bound to a virus), so we assign the modal taxonomy to the 5.9 percent of proteins with multiple taxonomies.

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,[39] 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.[40]

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

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

C Appendix Figures and Tables

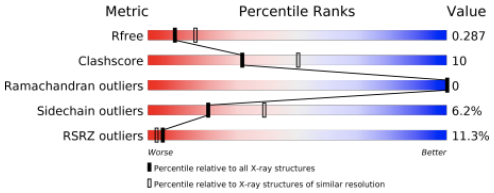
Figure C.1: Validation Report for PDB ID 4CMP — Crystal Structure of S. pyogenes Cas9

1 Overall quality at a glance (i)

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.



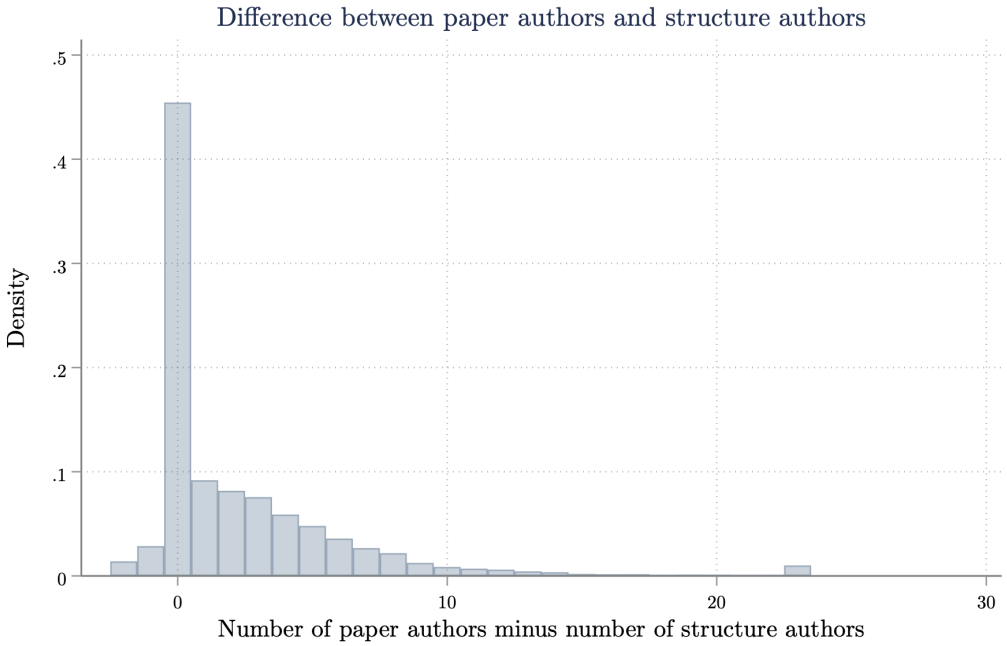
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	3285 (2.64-2.60)
Clashscore	122126	3641 (2.64-2.60)
Ramachandran outliers	120053	3586 (2.64-2.60)
Sidechain outliers	120020	3586 (2.64-2.60)
RSRZ outliers	108989	3218 (2.64-2.60)

4 Data and refinement statistics (i)

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

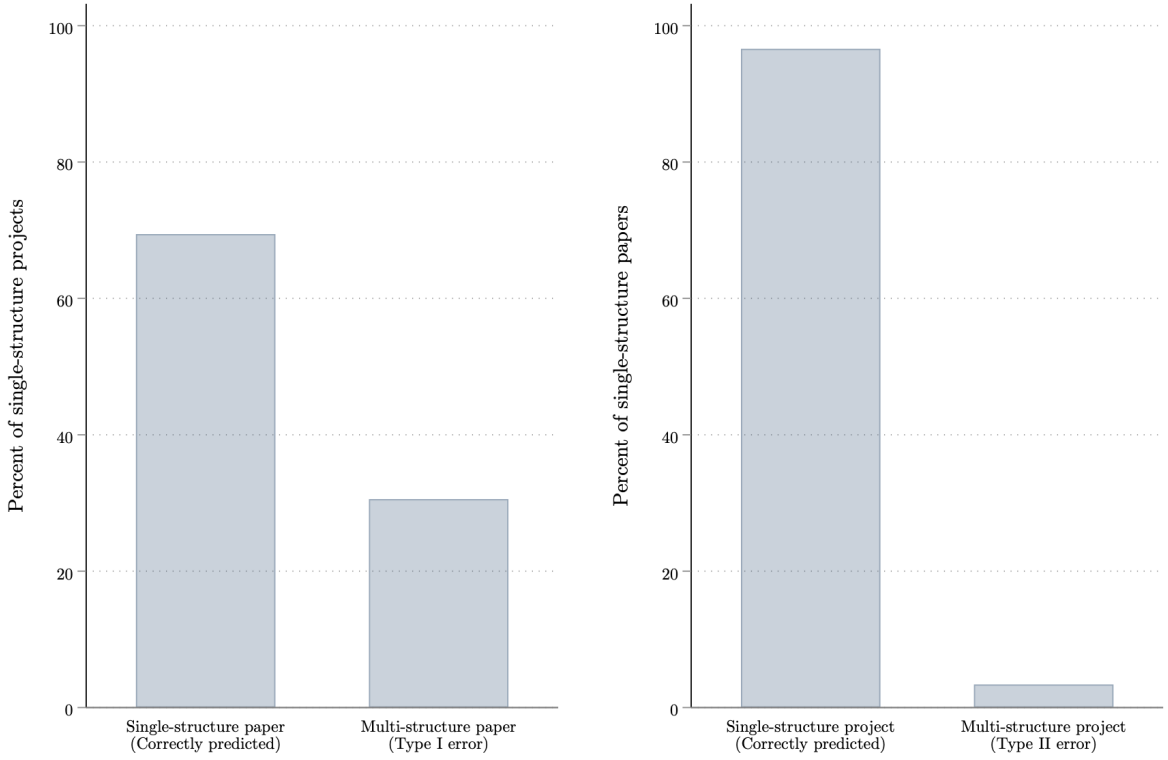
Notes: This figure presents some snapshots from the PDB x-ray structure validation report for PDB ID 4CMP. The "Source" column describes the software package (if applicable) that calculated the quality measure / property.

Figure C.2: Difference between Number of Structure Authors versus Number of Paper Authors



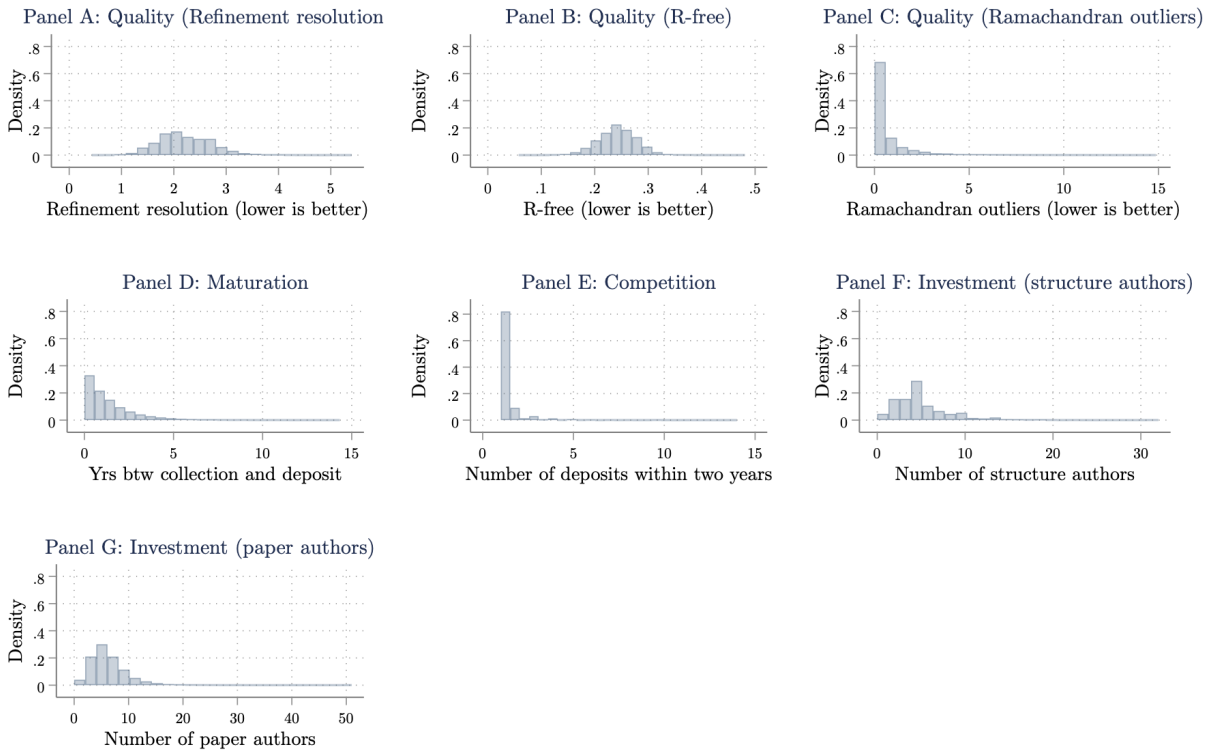
Notes: This figure the difference between the number of paper authors and the number of structure authors. The difference variable has been winsorized at the 1^{st} and 99^{th} percentile. The sample is the full analysis sample, excluding unpublished papers (which lack a paper author count).

Figure C.3: 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 C.4: Distributions of Key Outcome Variables



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

Table C.1: Correlation Between Quality Outcomes

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

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

Table C.2: LASSO-Selected Covariates

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

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

Table C.3: The Effect of Potential on Alternative Competition Measures

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

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

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