Race to the Bottom:
Competition and Quality in Science
Appendices for Online Publication

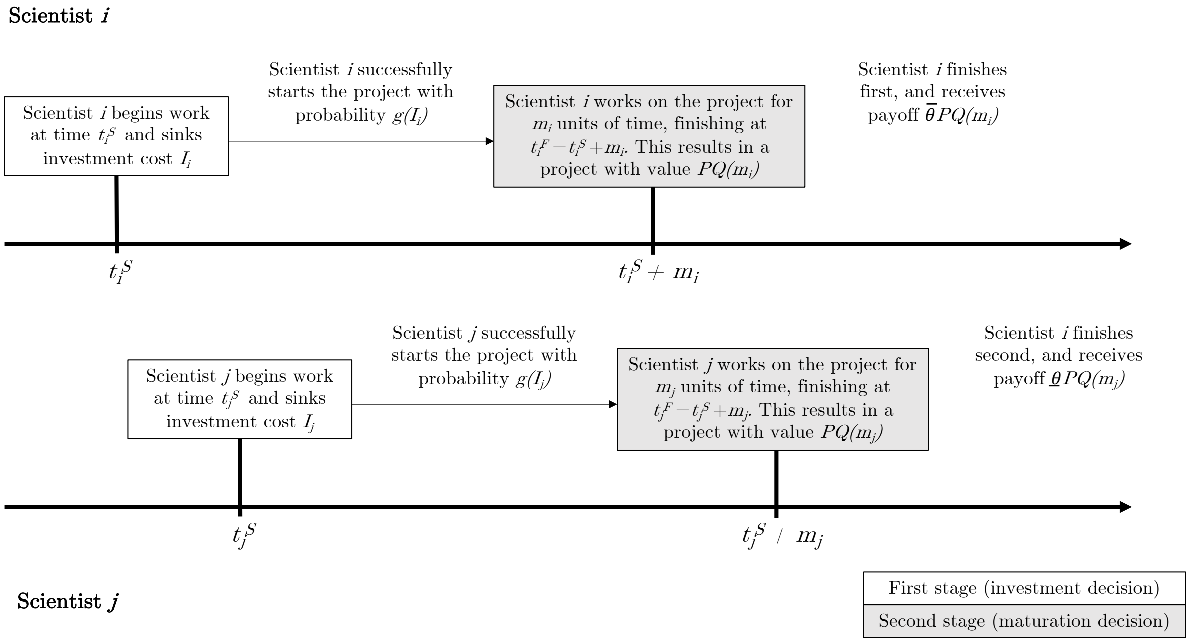
Ryan Hill and Carolyn Stein

January 22, 2025

1 Theoretical Appendix

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

Figure A1: Model Summary


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

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

1.1 Maturation

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

1.1.1 The No Competition Benchmark

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

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

$$m_i^{NC^*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \right\}. \tag{1}$$

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

$$\frac{Q'\left(m^{NC^*}\right)}{Q\left(m^{NC^*}\right)} = r. \tag{2}$$

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

1.1.2 Adding Competition

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

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

$$\pi(m_i, m_j) = (1-g) + gPr(t_i^F < t_j^F) = (1-g) + gPr(t_i^S + m_i < t_j^S + m_j). \tag{3}$$

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

$$m_i^{C*} \in \arg\max_{m_i} \left\{ e^{-rm_i} PQ\left(m_i\right) \left[\pi(m_i, m_j)\overline{\theta} + \left(1 - \pi(m_i, m_j)\right)\underline{\theta}\right]\right\}. \tag{4}$$

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

$$\frac{Q'\left(m_i^*\right)}{Q\left(m_i^*\right)} = r + \frac{1}{\Delta\left(\frac{2\overline{\theta} - g(\overline{\theta} - \underline{\theta})}{g(\overline{\theta} - \underline{\theta})}\right) + m_j - m_i^*}. \tag{5}$$

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

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

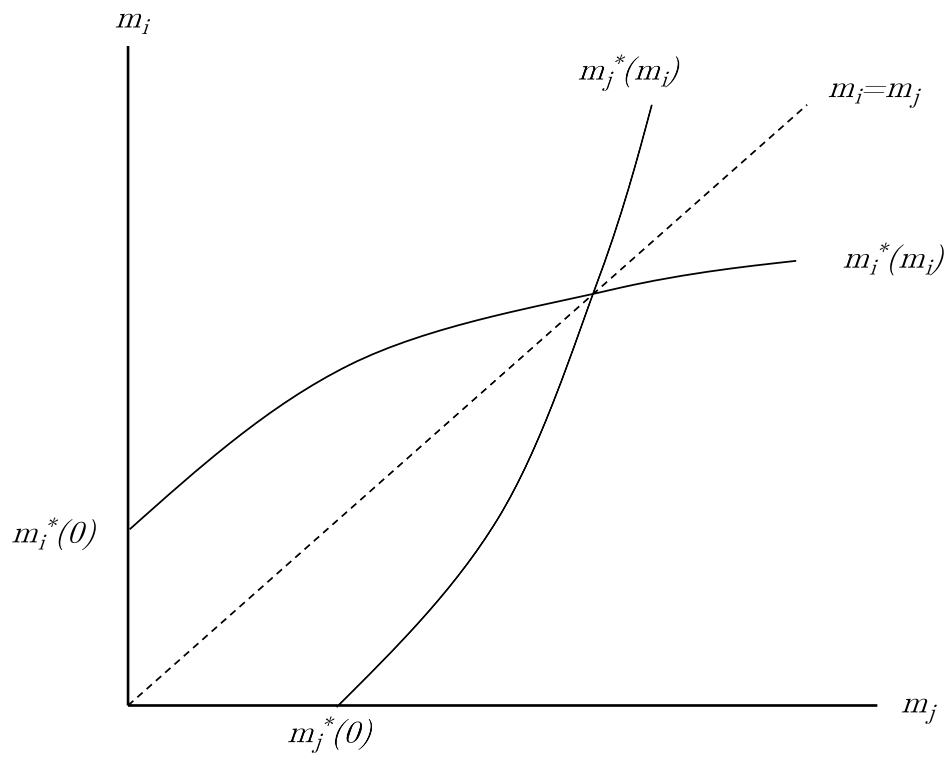
$$\frac{Q'\left(m^*\right)}{Q\left(m^*\right)} = r + \frac{g(I^*)(\overline{\theta} - \underline{\theta})}{\Delta\left(2\overline{\theta} - g(I^*)(\overline{\theta} - \underline{\theta})\right)}. \tag{6}$$

Proof. First, we will expand on how we derive the first-order condition for m_i^* (Equation 5). Taking the derivative of Equation 4 with respect to m_i and setting it equal to zero yields:

$$\frac{Q'(m_i^*)}{Q(m_i^*)} = r - \frac{\frac{\partial\pi}{dm_i}(\overline{\theta} - \underline{\theta})}{\pi(m_i, m_j)\overline{\theta} + (1 - \pi(m_i, m_j))\underline{\theta}}. \tag{7}$$

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 7 above yields Equation 5 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 7 collapses to the no-competition case, i.e., Equation 2. We will return to this caveat, but for now we will assume m_i is close to m_j. Equation 5 implicitly defines $m_i^*(m_j)$ as a function of m_j and parameters. If we can show that (i) $m_i^*(0) > 0$ and (ii) $\frac{dm_i^*}{dm_j} \in (0, 1)$, then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $m_i^*(m_j)$ and $m_j^*(m_i)$ will only cross the $m_i = m_j$ line once.

Figure A2: Maturation Best Response Functions



To show (i), plug $m_j = 0$ into Equation 5. This results in an equation that implicitly defines a unique $m_i^*(0) > 0$. To show (ii), we can totally differentiate equation 5 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

$$\frac{dm_i^*}{dm_j} = \left[\underbrace{\left(\frac{-Q(m_i^*)Q''(m_i^*) + Q'(m_i^*)^2}{Q(m_i^*)^2}\right)\left(\zeta + m_j - m_i^*\right)^2}_{>0} + 1\right]^{-1} \in (0,1). \tag{8}$$

Next, we confirm that the second-order conditions hold. Differentiating the objective function (Equation 4) twice with respect to m_i and evaluating at $m_i = m_j = m^*$ yields

$$Pe^{-rm_i}\left[Q''(m^*) - Q'(m^*)\left(r + \frac{1}{\zeta}\right)\right] < 0. \tag{9}$$

Therefore, $m_i^* = m_j^* = m^*$ is a local optimum. Plugging m^* in for both m_i and m_j (and assuming that $I_i = I_j = I^*$) in Equation 5 yields the expression in Proposition A1. However, as a final check, we need to confirm that this is also a global optimum. Note that Equation 6 tells us that as $\Delta \to 0$, $m_i^* \to 0$. This will yield a payoff of zero for researcher i.

This cannot be researcher i's best response, because there is always a $1-g$ probability that her competitor did not enter. Therefore, she would be better off selecting $m_i = m^{NC^*}$ and hoping that her competitor fails to enter the project. To map this intuition to the math, note that we are now considering a case where $m_i > m_j + \Delta$, and so the relevant first-order condition is now Equation 2. More generally, in order to ensure that $m_i^* = m_j^* = m^*$ is a global optimum we need the payoff from playing $m_i = m^*$ to be larger than the payoff to playing $m_i = m^{NC^*}$:

$$e^{-rm^*} PQ(m^*) \left[(1-\frac{g}{2})\overline{\theta} + \frac{g}{2}\underline{\theta}\right] > e^{-rm_i^{NC}} PQ(m_i^{NC^*}) \left((1-g)\,\overline{\theta} + g\underline{\theta}\right). \tag{10}$$

Because m^* is increasing in Δ, this defines a lower bound on Δ such that this equation will hold. Therefore, $m_i^* = m_j^* = m^*$ is a symmetric pure strategy Nash equilibrium as long as Δ is sufficiently large. Moreover, this is the only possible pure strategy Nash equilibrium. To see this, note that if $|m_i - m_j| < \Delta$, then the first-order condition in Equation 5 applies and we have the equilibrium defined by $m_i^* = m_j^* = m^*$. Alternatively, if $|m_i - m_j| \geq \Delta$, then the first-order condition defined by Equation 2 applies. But this implies that $m_i^* = m_j^* = m^{NC^*}$, which violates the assumption that $|m_i - m_j| \geq \Delta$. Therefore, if Δ is below some threshold, the Nash equilibrium must be mixed. We will focus on the pure strategy case throughout the remainder of the paper. □

Because $Q(m)$ is increasing and concave, we know Q'/Q is a decreasing function. Therefore, by comparing Equations 2 and 6, we can see that $m^{NC^*} > m^*$. In other words, competition leads to shorter maturation periods. This shortening is exacerbated when the difference between $\overline{\theta}$ and $\underline{\theta}$ is large (priority rewards are more lopsided), Δ is small (competitors start the projects close together, and so the "flow risk" of getting scooped is high), or when g is close to one (the entry of a competitor is likely). On the other hand, if $\overline{\theta} = \underline{\theta}$ (first and second place share the rewards evenly), $\Delta \to \infty$ (competition is very diffuse, so the "flow risk" of getting scooped is low), or $g = 0$ (the competitor doesn't enter), then we recover the no competition benchmark.

1.2 Investment

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

$$I_i^* \in \arg\max_{I_i} \left\{ g(I_i)\, e^{-rm_i^*} PQ\left(m_i^*\right) \left[\overline{\theta} - \frac{1}{2} g(I_j)\left(\overline{\theta} - \underline{\theta}\right)\right] - I_i \right\}. \tag{11}$$

Taking the first-order condition of Equation 11 implicitly defines scientist i's best-response function, which depends on I_j, m_i^*, and other parameters:

$$g'(I_i^*) = \frac{1}{e^{-rm_i^*} PQ\left(m_i^*\right)\left[\overline{\theta} - \frac{1}{2} g\left(I_j\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{12}$$

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

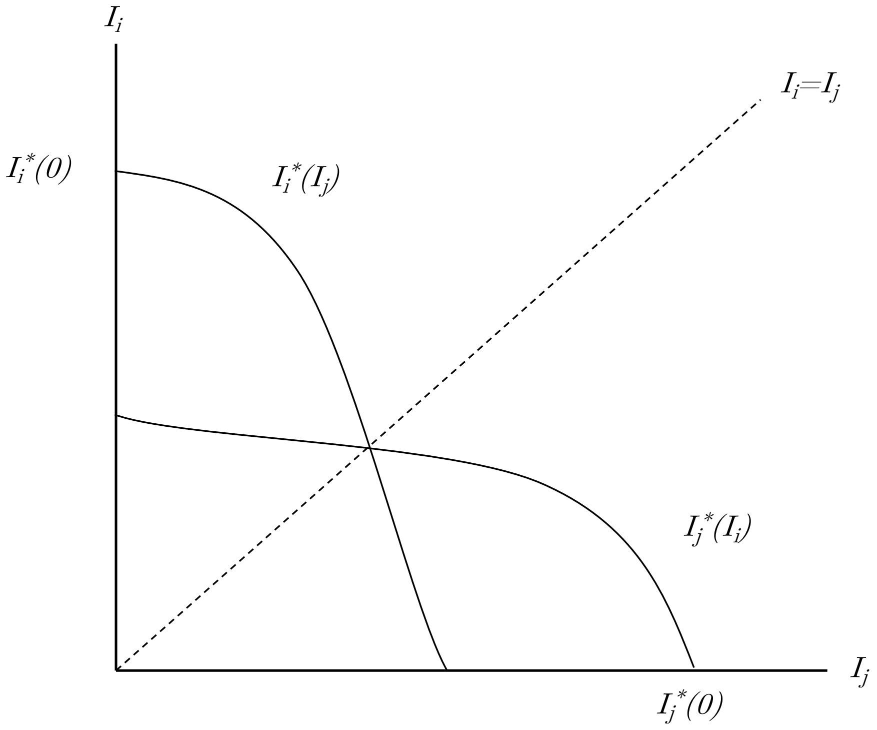
Proposition A2. *Assume that researchers are playing a symmetric pure strategy Nash equilibrium when selecting m in the second stage. Then, in the first stage, there is a unique symmetric pure strategy Nash equilibrium where $I_i^* = I_j^* = I^C$ and I_i^* is implicitly defined by*

$$g'(I^*) = \frac{1}{e^{-rm^*} PQ\left(m^*\right)\left[\overline{\theta} - \frac{1}{2} g\left(I^*\right)\left(\overline{\theta} - \underline{\theta}\right)\right]}. \tag{13}$$

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

Proof. Equation 12 implicitly defines $I_i^*(I_j)$ as a function of I_j, m_i^* (which depends on I_j), and parameters. If we can show that (i) $I_i^*(0) > 0$ and (ii) $\frac{dI_i^*}{dI_j} < 0$ then we will know that there is a unique and symmetric pure strategy Nash equilibrium, because $I_i^*(I_j)$ and $I_j^*(I_i)$ will only cross the $I_i = I_j$ line once.

Figure A3: Investment Best Response Functions



To show (i), imagine that j invests zero. Then i should surely invest some positive amount, because the marginal return will be be proportional to $g'(I_i)$. Due to the Inada conditions assumption on $g(\cdot)$, $g'(I_i)$ will be quite large for small values of I_i. To show (ii), we can totally differentiate Equation 12 with respect to I_j. Gathering terms and rearranging, we have that

$$\frac{dI_i^*}{dI_j} = -\frac{e^{-rm_i^*}P\left[\left(\overline{\theta} - \frac{1}{2}g(I_j)(\overline{\theta}-\underline{\theta})\right)\left(Q'(m_i^*) - rQ(m_i^*)\right)\frac{dm_i^*}{dI_j} - Q(m_i^*)\left(\frac{1}{2}g'(I_j)(\overline{\theta}-\underline{\theta})\right)\right]}{g''(I_j)\left[e^{-rm_i^*}PQ(m_i^*)\left(\overline{\theta}-\frac{1}{2}g(I_j)(\overline{\theta}-\underline{\theta})\right)\right]^2} < 0 \tag{14}$$

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

Equations 13 and 6 together define the optimal investment level and maturation period for scientists when entry into projects is endogenous. This allows us to prove the three key results described in the main text.

Proof of Proposition 1. *Consider an exogenous increase in the probability of project entry, g. This corresponds to an increase in competition, because it makes racing more likely. When projects become more competitive, the maturation period becomes shorter and projects become lower quality. In other words,* $\frac{dm^*}{dg} < 0$ *and* $\frac{dQ(m^*)}{dg} < 0$.

Proof. Looking at Equation 6, the left hand side is decreasing in m^*. Looking at the right hand side, we see it is increasing in $g(I^*)$. For the equality to hold as $g(I^*)$ increases, it must be the case that m^* decreases, i.e., that $\frac{dm^*}{dg} < 0$. Because $Q(m)$ is increasing, this also implies that $\frac{dQ(m^*)}{dg} < 0$. □

Proof of Proposition 2. *Higher potential projects generate more investment and are therefore more competitive. In other words,* $\frac{dI^*}{dP} > 0$ *and* $\frac{dg(I^*)}{dP} > 0$.

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

$$\underbrace{e^{-rm_1^*}P_1Q(m_1^*)\left[\overline{\theta} - \frac{1}{2}g(I_1^*)(\overline{\theta} - \underline{\theta})\right]}_{\text{PDV of project 1}} > \underbrace{e^{-rm_2^*}P_2Q(m_2^*)\left[\overline{\theta} - \frac{1}{2}g(I_2^*)(\overline{\theta} - \underline{\theta})\right]}_{\text{PDV of project 2}}.$$

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

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

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

1.3 The Social Planner's Problem

The social planner does not care which researcher finishes the project first, and therefore has a different objective function:

$$\max_{m,I} \left\{ \underbrace{\left(1 - (1 - g(I))^2\right)}_{\text{probability at least one researcher successfully starts}} \cdot \underbrace{e^{-rm}kPQ(m)}_{\text{social PDV of project}} - \underbrace{2I}_{\text{investment costs}} \right\}. \tag{15}$$

The socially optimal value of m, denoted m^{SP^*}, is defined by the first-order condition of Equation 15 with respect to m:

$$\frac{Q'(m^{SP^*})}{Q(m^{SP^*})} = r. \tag{16}$$

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 2). 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.

2 Data Appendix

2.1 Description of the Protein Data Bank Data

The first iteration of the Protein Data Bank (PDB) started in 1971. Today, a non-profit organization called the World Wide Protein Data Bank (wwPDB) curates and manages the database. The wwPDB is a collaboration of four existing data banks from around the world: Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB), Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), and Biological Magnetic Resonance Data Bank (BMRB). The data has been standardized and currently represents the universe of discoveries deposited in each of these archives. All new discoveries deposited to any database are transferred to, processed, standardized, and archived by the RCSB (Berman et al. 2006) at Rutgers University. Details about the PDB data can be found on their website.[1]

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

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

Data about Ramachandran outliers, one of the quality metrics, was not available through RCSB custom reports. Instead, we accessed validation reports data from the PDBe REST API[3] provided

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

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

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

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.

2.2 Description of the Web of Science Data

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

2.3 Description of the UniPROT Knowledgebase Data

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

2.4 Description of DrugBank Data

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

2.5 Harmonizing the Data

Variables in our data are reported at three different levels: the entity level, the structure level, and the paper level. Entity is the smallest level, as some protein structures are comprised of multiple

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

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

entities. Structure is the middle level, as some papers contain multiple structures. Paper is the largest level. The levels fully nest (there is a many-to-one correspondence between entities and structures, and a many-to-one correspondence between structures and papers). Below, we report the variables we use at the level they are uniquely indexed:

Variables at the Entity Level:

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

Variables at the Structure Level:

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

Variables at the Paper Level:

- PubMed ID
- Paper authors
- Citations

Throughout our analysis, we use a protein structure as our unit of analysis. However, some of the variables we need are indexed at either the entity or the paper level. To create a one-to-one link between papers and structures, we drop all instances where papers are linked to multiple structures (20% of PDB-linked papers). Moreover, since about 30% of deposits are never published, we make a similar restriction for groups of structure deposits that appear to have been part of the same unpublished project. We group unpublished structures into the same "project" if the deposits have the same first and last PDB structure author and share the same release date. Unpublished projects with more than one structure are dropped to mirror the single-structure paper restriction. Appendix Figure E2 assesses this heuristic among the set of published structures.

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

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

- Priority race. As an additional measure of competition, we have an indicator for whether a protein was involved in a priority race. Following Hill and Stein (2025), we code a structure as being involved in a priority race if any of the entities were involved, but drop instances where structures contain more than 15 entities (less than one percent of the sample).

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

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

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

- Quality improvement. As part of our effort to measure the cost of improving structures, we develop an indicator which codes for whether a protein structure represents an improvement over a prior structure. Quality measures are at the structure level, but again protein similarity is computed at the entity level. For every entity in the protein structure, we measure whether it represents an improvement over the "priority entity." We then aggregate up to the structure level. A structure with any improved entity is coded as having "any improvement" and a structure with all entities being better than the priority entity is coded as having "full improvement."

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

3 Survey Experiment Details

3.1 Selection of Comparison Fields

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

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

3.2 Selection of Structural Biologists

Structural biology is not listed as a specific field in the MAG taxonomy. Therefore, we use two approaches to constructing the structural biology group. First, we find all email addresses that are listed on papers directly linked to the PDB, giving us 3,038 addresses. We call this the "structural biology - PDB" group. This is the sample that most directly matches the authors in our main analysis, but we were concerned that the sample size might be too small. Therefore, we also used a second approach to supplement this sample. In this approach, we calculate the share of all level-0/level-1 fields that contain a link to a PDB publication, and select fields that have the largest share of PDB-linked papers. The final combinations we chose for this broader category are: biology-stereochemistry, biology-crystallography, biology-biophysics, and chemistry-stereochemistry. Not including the email addresses directly linked to the PDB, this broader category consists of 7,195 email addresses. We combine our PDB group with this group to create a larger sample that we call "structural biology - all."

3.3 Survey Implementation and Text

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

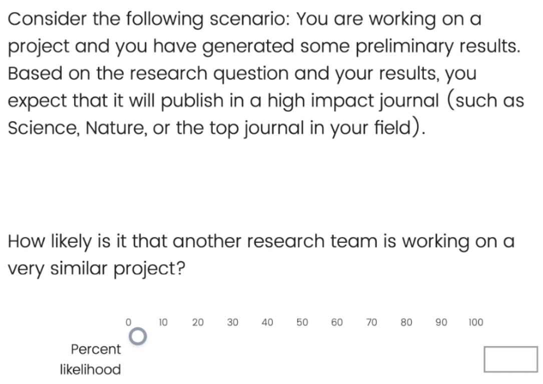
We ran the survey using Qualtrics, and sent the initial email on May 15, 2023 to 99,282 email addresses. We sent a reminder to anyone who had not filled out the survey on May 18 and May 24. We closed the survey on June 5. In total, we received 10,557 complete responses. We dropped all responses that Qualtrics coded as likely spam, leaving us with 9,211 responses (9.3% response rate). 8,237 respondents answered all three of our survey experiment questions, so these comprised our sample.

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

Figure C1: Qualtrics Survey Experiment Questions

(a) Question 1: The Effect of Potential on Competition (Randomized)

High potential



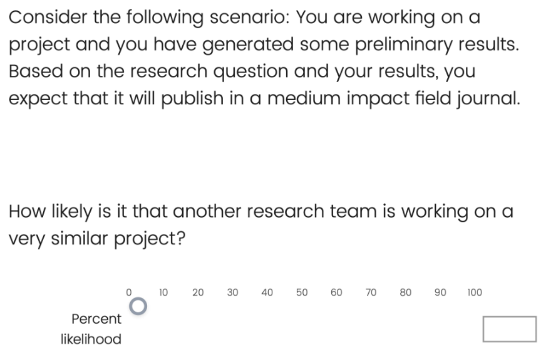
Consider the following scenario: You are working on a project and you have generated some preliminary results. Based on the research question and your results, you expect that it will publish in a high impact journal (such as Science, Nature, or the top journal in your field).

How likely is it that another research team is working on a very similar project?

0 10 20 30 40 50 60 70 80 90 100

Percent likelihood ○ □

Low potential



Consider the following scenario: You are working on a project and you have generated some preliminary results. Based on the research question and your results, you expect that it will publish in a medium impact field journal.

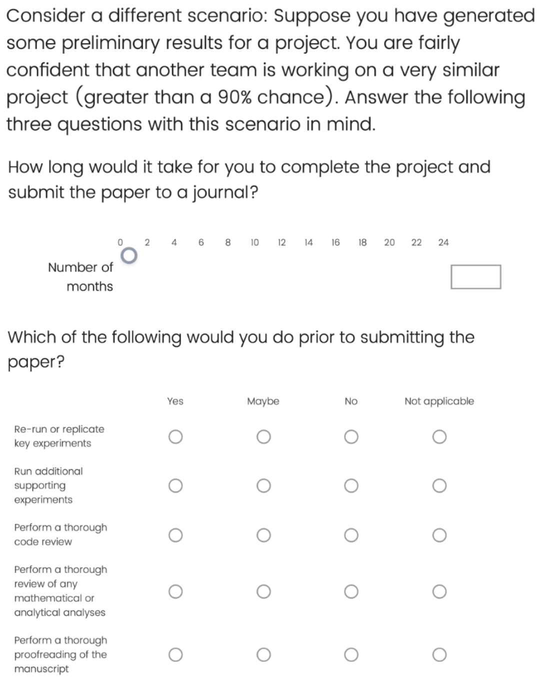
How likely is it that another research team is working on a very similar project?

0 10 20 30 40 50 60 70 80 90 100

Percent likelihood ○ □

(b) Questions 2 and 3: The Effect of Competition on Maturation and Quality (Randomized)

High competition



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

How long would it take for you to complete the project and submit the paper to a journal?

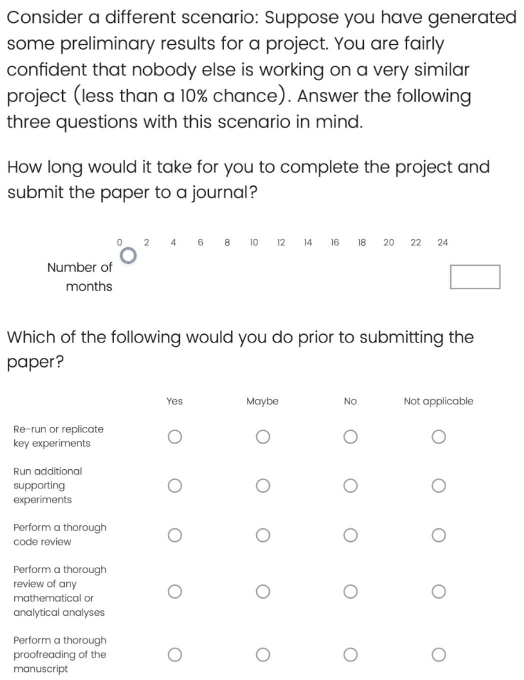
0 2 4 6 8 10 12 14 16 18 20 22 24

Number of months ○ □

Which of the following would you do prior to submitting the paper?

	Yes	Maybe	No	Not applicable
Re-run or replicate key experiments	○	○	○	○
Run additional supporting experiments	○	○	○	○
Perform a thorough code review	○	○	○	○
Perform a thorough review of any mathematical or analytical analyses	○	○	○	○
Perform a thorough proofreading of the manuscript	○	○	○	○

Low competition



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

How long would it take for you to complete the project and submit the paper to a journal?

0 2 4 6 8 10 12 14 16 18 20 22 24

Number of months ○ □

Which of the following would you do prior to submitting the paper?

	Yes	Maybe	No	Not applicable
Re-run or replicate key experiments	○	○	○	○
Run additional supporting experiments	○	○	○	○
Perform a thorough code review	○	○	○	○
Perform a thorough review of any mathematical or analytical analyses	○	○	○	○
Perform a thorough proofreading of the manuscript	○	○	○	○

4 Welfare Calculations

4.1 Imputing Missing Quality

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

$$Y_{it} = \alpha + \beta P_i + \lambda NonSG_i + \delta(P_i \times NonSG_i) + \tau_t + X_i' \gamma + \varepsilon_{it} \tag{17}$$

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

$$Y_{it}^{CF} = \alpha + \beta P_i + \tau_t + X_i' \gamma + \varepsilon_{it} = Y_{it} - \lambda - \delta P_i. \tag{18}$$

4.2 Costs of Improved Deposits

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

The first challenge arises because, as discussed in Appendix 2, different variables are defined at different levels. In particular, quality is defined at the structure level. However, similarity is defined at the entity level, and some structures have several entities. We use both of these variables to define structure improvements, as detailed below.

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

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

Definition 2. We take our subsample of 54,816 structures from definition 1 but we drop any structures that were involved in a priority race (see Hill and Stein (2025) for more details on how priority races are defined). The goal is to exclude unintentional re-deposits, i.e., projects that would have been produced anyway because the racing teams were working contemporaneously. This leaves us with 54,172 structures (42% of the x-ray sample).

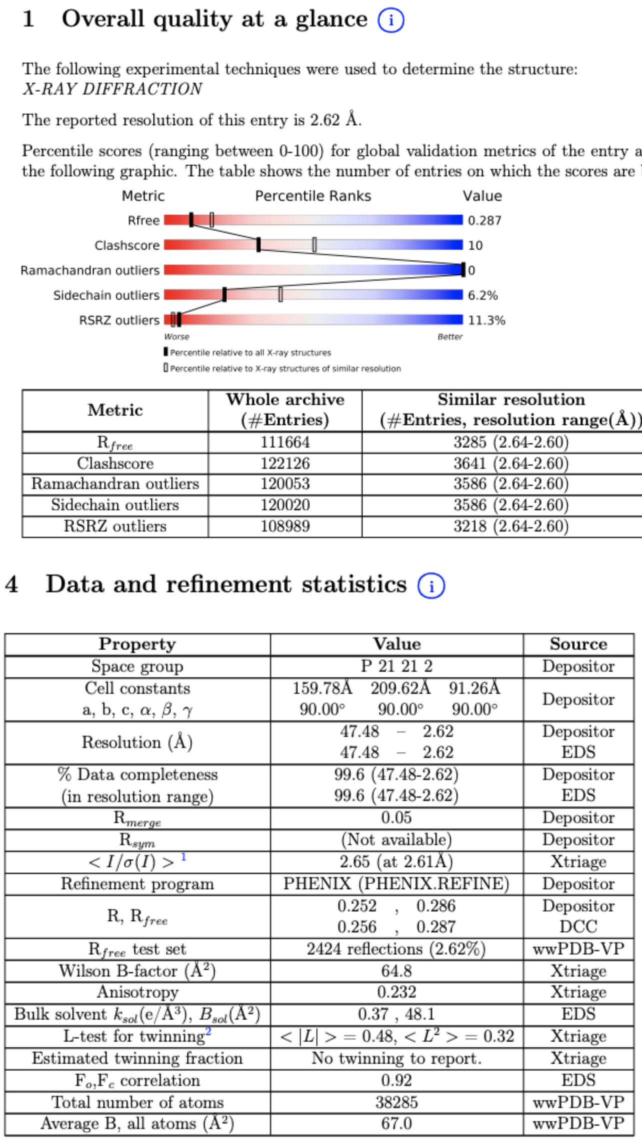
Definition 3. So far, we have not imposed any restrictions that these re-deposits represent improvements over the initial deposits. In our model, the sole purpose of re-deposits is to improve the quality. Thus, we might argue that only these should count in our calculation of the costs of improved deposits.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

Here we run into the issue of quality being defined at the structure level whereas similarity is defined at the entity level, as discussed in Appendix 2. Aggregating up to the structure level, we call any re-deposit an "improved re-deposit" if at least one entity is an improvement over the priority entity. This leaves us with 23,318 structures (18% of the x-ray sample).

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

5 Appendix Figures and Tables

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



1 Overall quality at a glance ⓘ

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

The reported resolution of this entry is 2.62 Å.

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

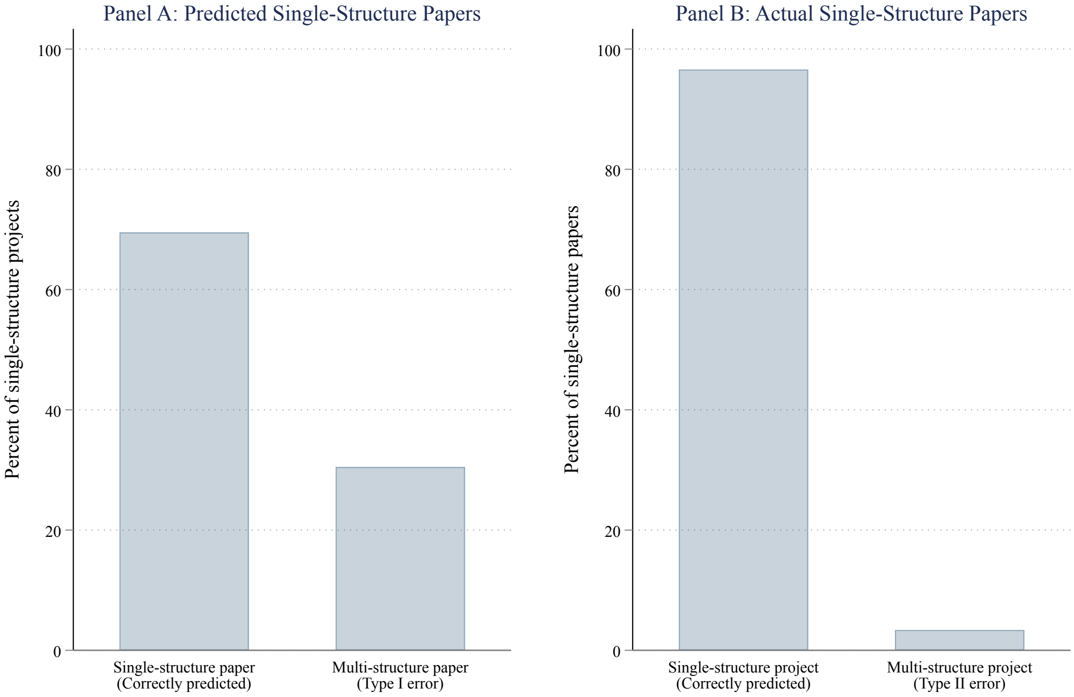
Metric	**Whole archive (#Entries)**	**Similar resolution (#Entries, resolution range(Å))**
R_{free}	111664	3285 (2.64-2.60)
Clashscore	122126	3641 (2.64-2.60)
Ramachandran outliers	120053	3586 (2.64-2.60)
Sidechain outliers	120020	3586 (2.64-2.60)
RSRZ outliers	108989	3218 (2.64-2.60)

4 Data and refinement statistics ⓘ

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

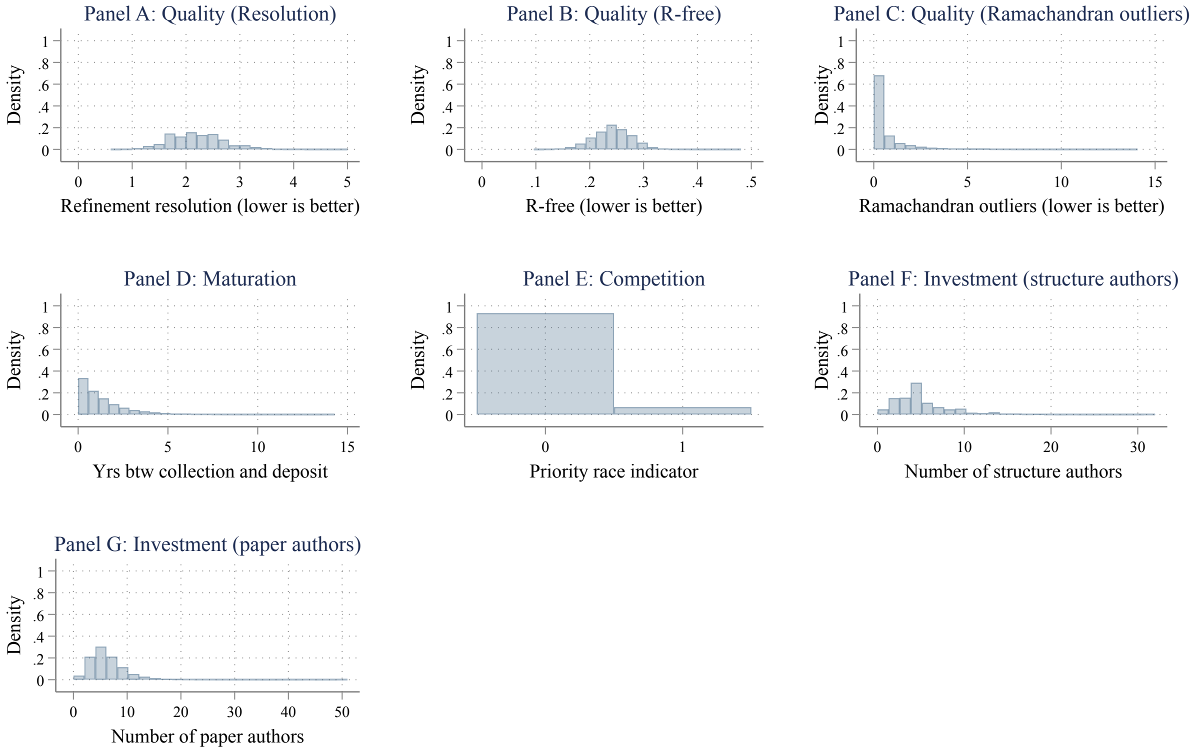
Notes: This figure presents some snapshots from the PDB x-ray structure validation report for PDB ID 4CMP. The “Source” column describes the software package (if applicable) that calculated the quality measure / property.

Figure E2: Predicting Single-Structure Projects



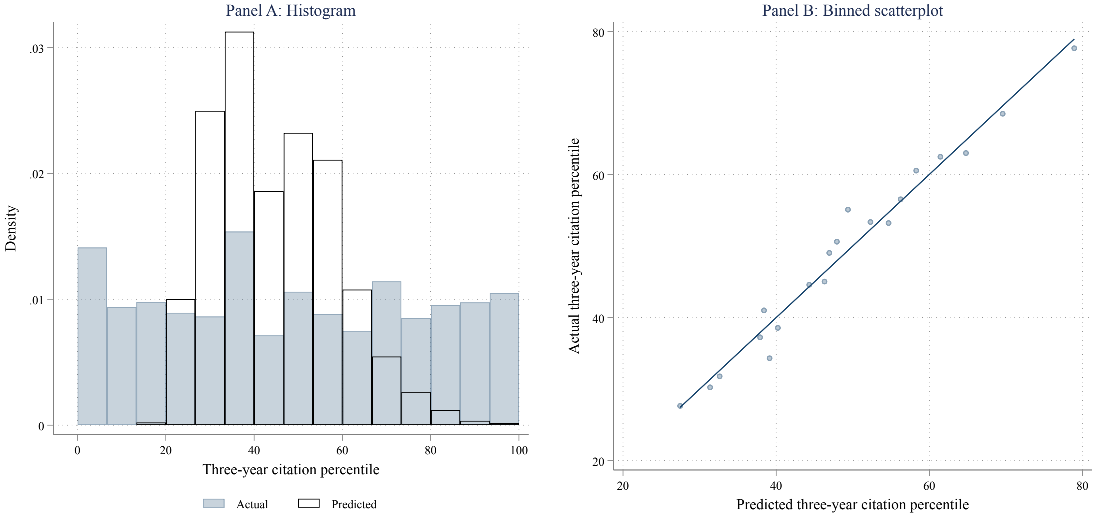
Notes: This figure assesses how well we predict whether a structure will be the only structure in a paper. Panel A looks at the set of structures we predict will fall in single-structure papers ("single structure projects"). About 70% of these are indeed single-structure papers, implying a 30% false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

Figure E3: Distributions of Key Outcome Variables



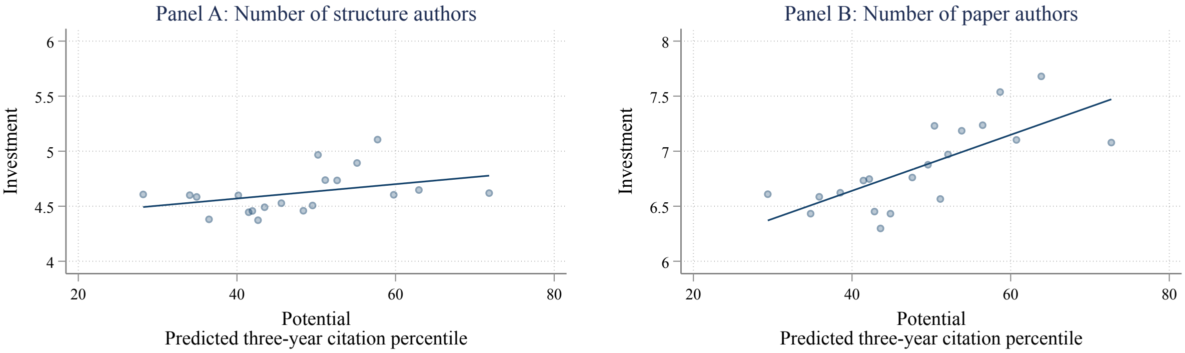
Notes: This figure provides histograms of the distributions of our key outcome variables. All variables have been winsorized at the 99.9^{th} percentile to make the figures easier to read. The sample is the full analysis sample.

Figure E4: LASSO Validation



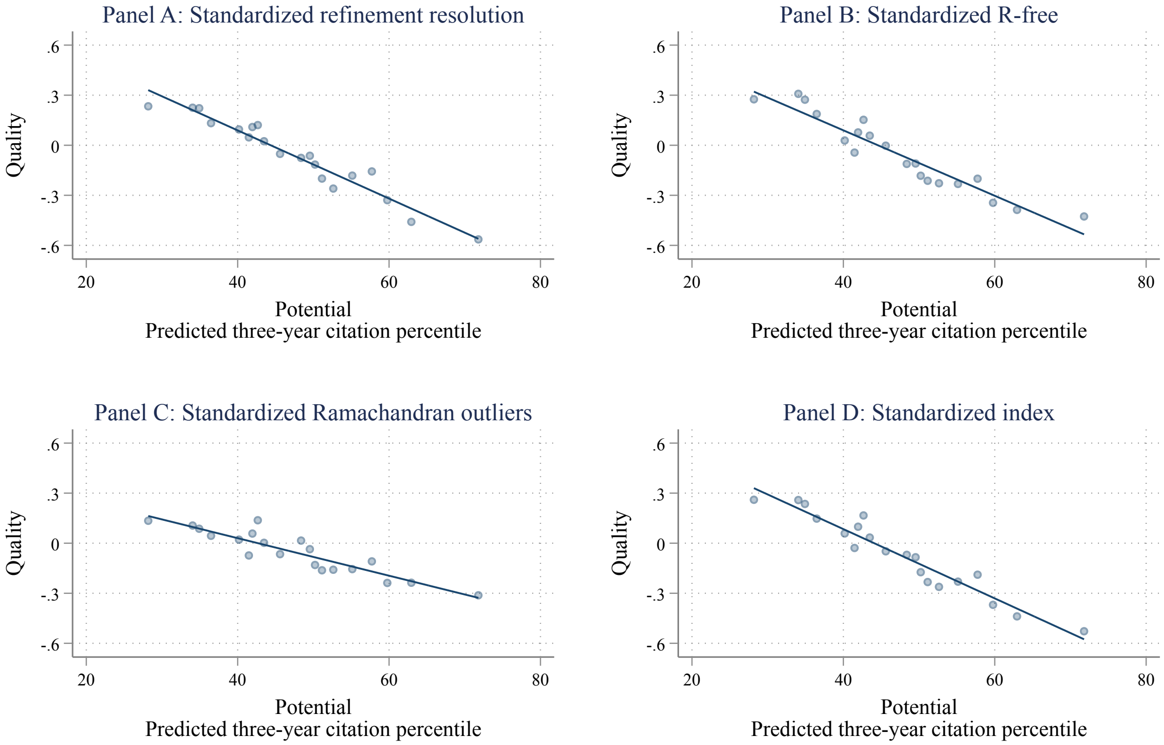
Notes: Panel A of this figure plots the distribution of actual and predicted potential. Panel B presents a graph of actual versus predicted potential as a binned scatterplot. In both panels, potential is measured by the percentile of the structure's three-year citation count. The sample is all structures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

Figure E5: The Effect of Potential on Investment



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

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



Panel A: Standardized refinement resolution

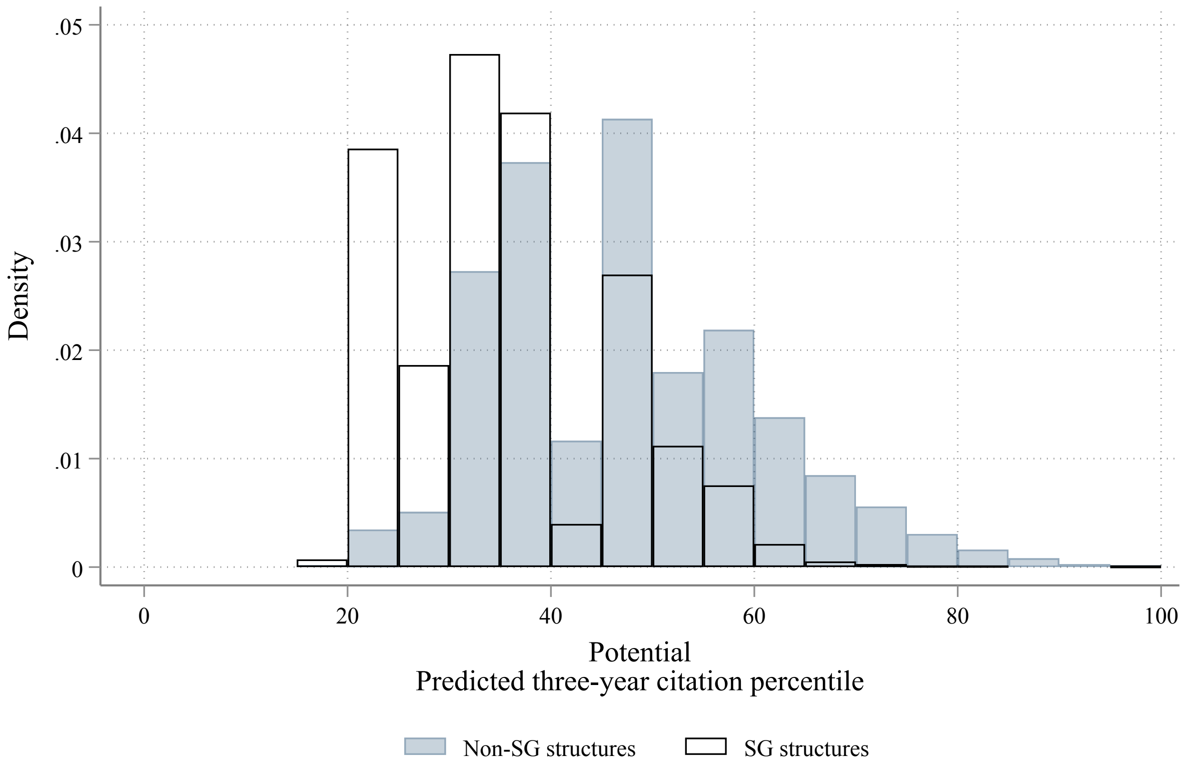
Panel B: Standardized R-free

Panel C: Standardized Ramachandran outliers

Panel D: Standardized index

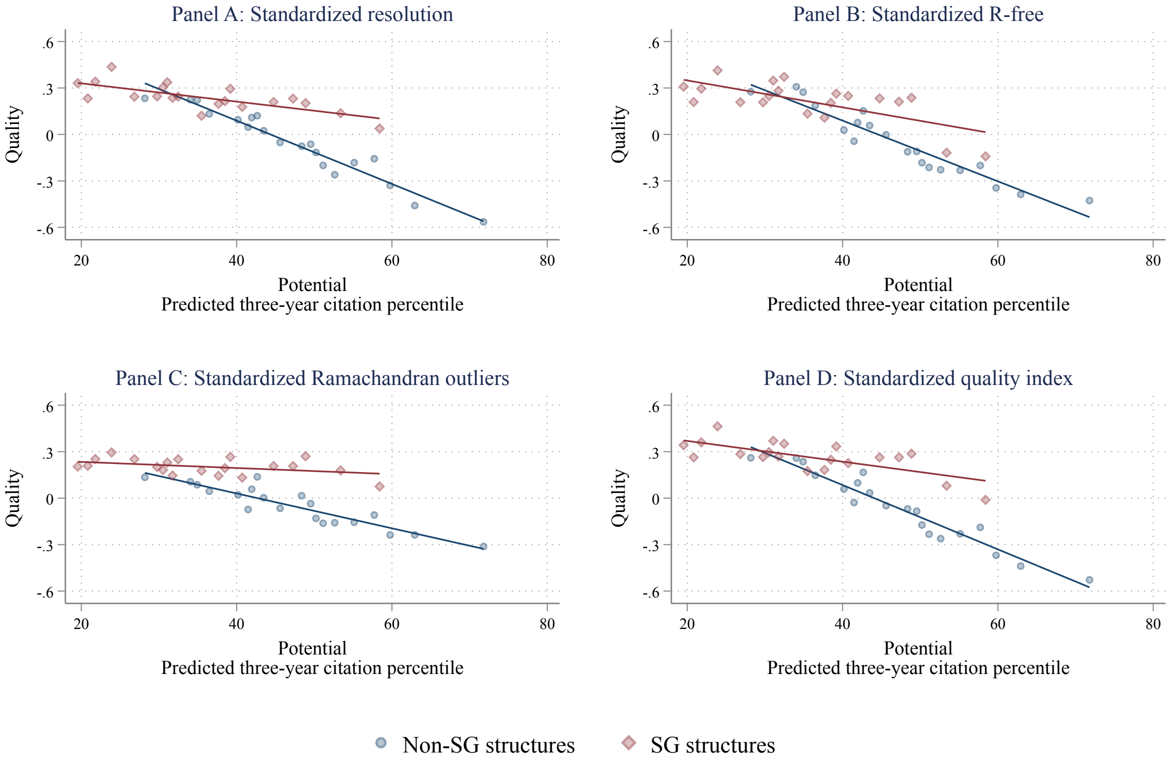
Notes: This figure plots the relationship between potential and additional quality measures, testing Proposition 3. Potential is measured as the predicted three-year citation percentile. Quality measures are described in the text. The plots are presented as a binned scatterplot, constructed as described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

Figure E7: Potential Distributions by Structural Genomics Status



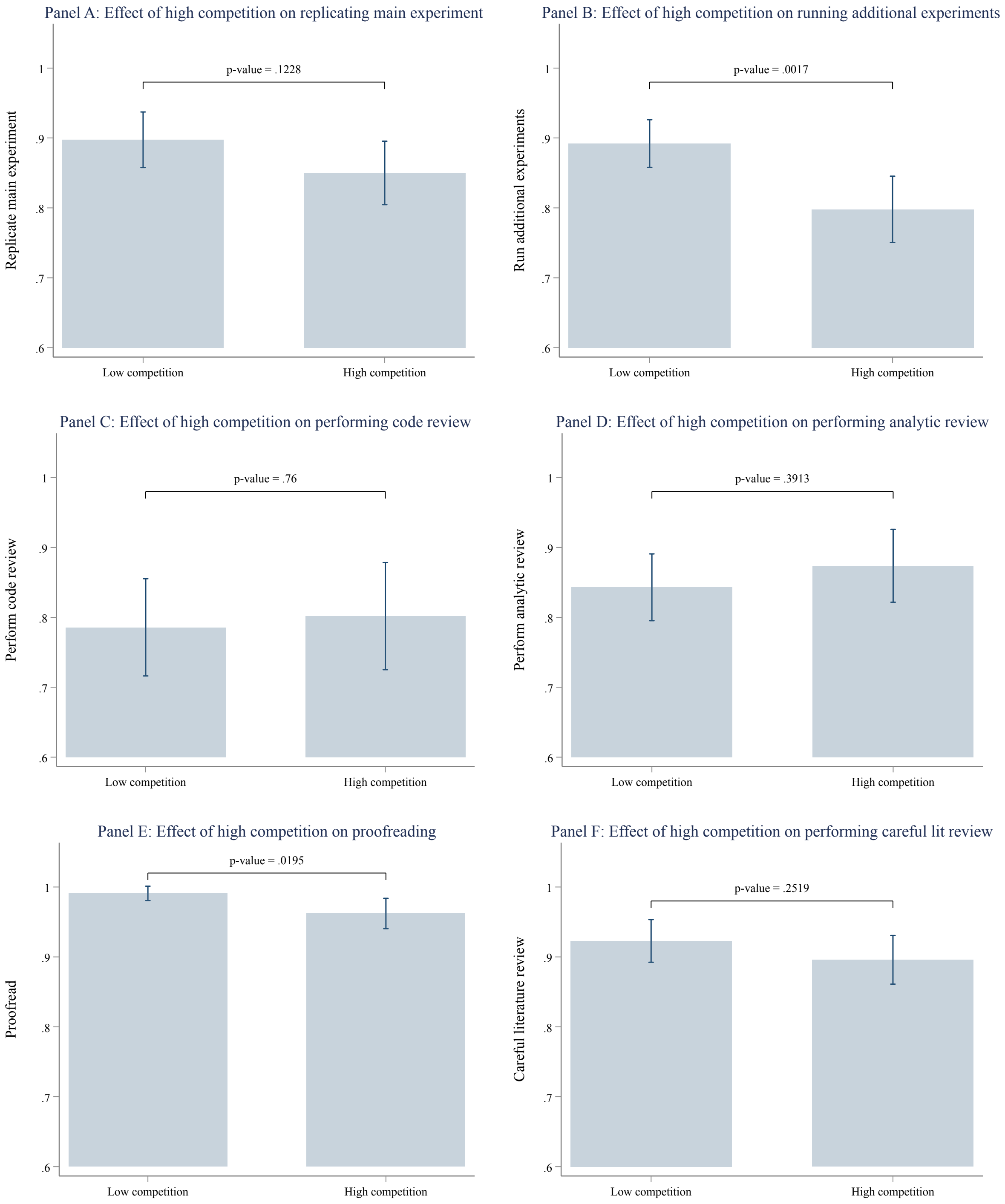
Notes: This figure plots the distribution of potential (as measured by predicted three-year citation percentile) for both non-SG and SG structures. The sample is all structures in the analysis sample.

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



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

Figure E9: Survey Experiment: Individual Quality Measures


Panel A: Effect of high competition on replicating main experiment

Panel B: Effect of high competition on running additional experiments

Panel C: Effect of high competition on performing code review

Panel D: Effect of high competition on performing analytic review

Panel E: Effect of high competition on proofreading

Panel F: Effect of high competition on performing careful lit review

Notes: This figure shows the results from our survey experiment of 309 active structural biologists. Details of the survey experiment can be found in Section 5 and Appendix C. Respondents were randomly assigned to receive a prompt about a low-competition or high-competition project and asked which quality control items they would complete before submitting the project. Each panel corresponds to a different measure of quality.

Table E1: Correlation Between Quality Outcomes

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

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

Table E2: LASSO-Selected Covariates

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

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

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

	Competition	Maturation		Quality	
	Priority Race	Years	Years	Std. index	Std. index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	0.0012***	-0.0063***	-0.0039	-0.0208***	-0.0153***
OLS SE	(0.0002)	(0.0013)	(0.0025)	(0.0008)	(0.0015)
Bootstrapped SE	(0.0002)	(0.0014)	(0.0022)	(0.0009)	(0.0014)
Principal investigator FEs?			Y		Y
Panel B. With complexity controls					
Potential	0.0012***	-0.0060***	-0.0034	-0.0190***	-0.0141***
OLS SE	(0.0002)	(0.0014)	(0.0026)	(0.0008)	(0.0014)
Bootstrapped SE	(0.0002)	(0.0014)	(0.0023)	(0.0010)	(0.0013)
Principal investigator FEs?			Y		Y

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

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

	Std. refinement resolution	Std. R-free	Std. Rama outliers
Dependent variable	(1)	(2)	(3)
Panel A. Without complexity controls			
Potential	-0.020***	-0.020***	-0.011***
	(0.001)	(0.001)	(0.001)
R-squared	0.049	0.082	0.064
Panel B. With complexity controls			
Potential	-0.019***	-0.018***	-0.010***
	(0.001)	(0.001)	(0.001)
R-squared	0.273	0.160	0.101
Mean of dependent variable	-0.062	-0.056	-0.053
Observations	16,215	16,215	16,215

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

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

	Std. refinement resolution	Std. R-free	Std. Rama outliers	Std. index
Dependent variable	(1)	(2)	(3)	(4)
Panel A. Without complexity controls				
Potential	-0.014***	-0.015***	-0.009***	-0.015***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.128	0.133	0.098	0.124
Panel B. With complexity controls				
Potential	-0.015***	-0.015***	-0.009***	-0.016***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.321	0.198	0.130	0.244
Mean of dependent variable	-0.062	-0.056	-0.053	-0.069
Observations	16,215	16,215	16,215	16,215

Notes: This table shows the relationship between additional quality measures and potential, controlling for the journal of publication. The level of observation is a structure-paper. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Complexity controls include log molecular weight, log residue count, and log atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample that have been published. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table E6: The Effect of Potential on Quality, Complexity Control Robustness

	Competition	Maturation		Quality	
	Priority race	Years	Years	Std. index	Std. index
Dependent variable	(1)	(2)	(3)	(4)	(5)
Panel A. Without complexity controls					
Potential	0.0012***	-0.0063***	-0.0039	-0.0208***	-0.0153***
	(0.0002)	(0.0013)	(0.0025)	(0.0008)	(0.0015)
Principal investigator FEs?			Y		Y
R-squared	0.010	0.017	0.493	0.068	0.480
Panel B. With non-parametric complexity controls					
Potential	0.0012***	-0.0061***	-0.0040	-0.0206***	-0.0149***
	(0.0002)	(0.0014)	(0.0025)	(0.0007)	(0.0014)
Principal investigator FEs?			Y		Y
R-squared	0.012	0.022	0.496	0.172	0.536
Mean of dependent variable	0.077	1.746	1.723	-0.069	-0.118
Observations	16,215	14,638	12,088	16,215	13,505

Notes: This table shows the relationship between competition / maturation / quality and potential, estimating equation (5) in the text. The level of observation is a structure-paper. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Non-parametric complexity controls take molecular weight, residue count, and atom site count and split each variable into bins based on quintiles. These quintiles are then all fully interacted. All regressions control for deposition year. The number of observations corresponds to the number of non-structural genomics structures in the analysis sample. The number of observations in column (2) is lower because maturation is missing for a subset of observations. The number of observations in columns (3) and (5) are lower because we drop singleton-PI observations when adding PI fixed effects. The mean of the standardized quality variables is not zero because we exclude SG structures which are part of the standardization sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table E7: The Effect of Potential on Quality by Structural Genomics Status (Additional Outcomes)

	Std. refinement resolution	Std. R-free	Std. Rama outliers
Dependent variable	(1)	(2)	(3)
Panel A. Without complexity controls			
Potential	-0.007***	-0.010***	-0.003***
	(0.001)	(0.001)	(0.001)
Non-structural genomics	0.358***	0.215***	0.086*
	(0.054)	(0.056)	(0.048)
Potential * Non-structural genomics	-0.013***	-0.009***	-0.008***
	(0.001)	(0.001)	(0.001)
R-squared	0.056	0.090	0.073
Panel B. With complexity controls			
Potential	-0.006***	-0.008***	-0.003***
	(0.001)	(0.001)	(0.001)
Non-structural genomics	0.361***	0.217***	0.080*
	(0.049)	(0.054)	(0.048)
Potential * Non-structural genomics	-0.012***	-0.009***	-0.007***
	(0.001)	(0.001)	(0.001)
R-squared	0.265	0.170	0.106
Mean of dependent variable	0.000	0.000	0.000
Observations	20,434	20,434	20,434

Notes: This table shows the relationship between additional quality measures and potential, interacted with structural genomics status, estimating equation (6) in the text. The level of observation is a structure-paper. Potential is measured as the predicted three-year citation percentile, following the LASSO prediction method described in the text. Structural genomics deposits are defined as described in the text. Complexity controls include log molecular weight, log residue count, and log atom site count and their squares. All regressions control for deposition year. The number of observations corresponds to the number of structures in the analysis sample. Heteroskedasticity-robust standard errors are in parentheses.

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

Table E8: First Stage Results from Most Common Species

Dependent Variable: Competition	(1)	(2)	(3)	(4)	(5)
Taxonomy	Human	E. coli	Mouse	Yeast	Hay bacillus
Taxonomy indicator	0.033***	0.005	-0.003	0.015	-0.020
	(0.005)	(0.009)	(0.009)	(0.012)	(0.016)
Complexity controls?	Y	Y	Y	Y	Y
First-stage F statistic	38.0	0.3	0.1	1.6	1.5
Count of taxonomy observations	4,005	992	826	616	221
Total observations	16,215	16,215	16,215	16,215	16,215

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

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

Table E9: Assessing Balance Between Non-Human and Human Structures

	Non-human structures	Human structures	Difference
Molecular weight	11.01	10.94	-0.065***
Residue count	6.26	6.20	-0.062***
Atom site count	8.27	8.20	-0.069***
Observations	12,210	4,005	

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

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

Table E10: Cost Estimates of Solving or Replicating an X-Ray Crystallography Structure

Source	**Cost Estimate (Nominal)**	**Cost Estimate (2024 Dollars)**	**Text from Source**
Sullivan, Kevin, Peggy Brennan-Tonetta, and Lucas J. Marxen. 2017. "Economic Impacts of the RCSB Protein Data Bank." *Mimeo.*	$100,000	$119,600	"While the costs of data creation and deposition are unknown, a reasonable estimate to replicate the RCSB-PDB data archive is $12 billion (assuming $100,000 avg. cost to replicate each entry)."
Darnell, Steve. 2015. "Why Structure Prediction Matters." *DNAStar Blog.*	$100,000	$131,900	"Solving structures using crystallography and NMR requires extremely specialized training, a high degree of skill, and a lot of luck. The cost of solving a new, unique structure is on the order of $100,000."
Ledford, Heidi. 2010. "Consortium Solves its 1,000th Protein Structure." *Nature.*	$150,000	$213,600	"An international team of 180 scientists, united in their goal to rapidly determine the structure of proteins related to human health, has solved its 1,000th structure. It has taken the researchers of the Structural Genomics Consortium (SGC) six years and US$150 million to achieve...Edwards says that the team still tries to keep the price per structure below $150,000."
Stevens, Raymond C. 2003. "The Cost and Value of Three-Dimensional Protein Structure." *Drug Discovery World.*	$140,000 to $450,000 (for non-membrane proteins)	$237,000 to $763,000	See Table 5 of Stevens (2003).

Notes: This table shows different cost estimates of solving or replicating an x-ray crystallography structure. We use the Bureau of Labor Statistics' CPI estimates to transform the nominal dollar amounts to 2024 dollar amounts, assuming that the nominal amounts are from the year of the source's publication.

References

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.

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

Hill, Ryan and Carolyn Stein, "Scooped! Estimating Rewards for Priority in Science," *Journal of Political Economy*, 2025, *133* (3).

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.