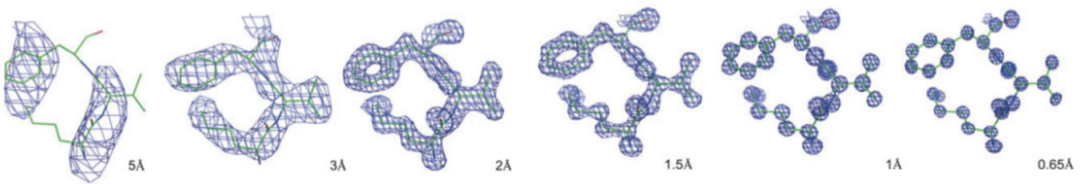
Race to the Bottom: Competition and Quality in Science

Ryan Hill
Brigham Young University
ryan_hill@byu.edu

Carolyn Stein
Stanford Institute for Economic Policy Research
cstein2@stanford.edu

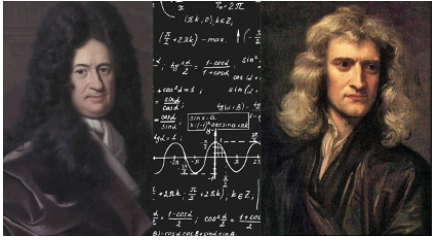


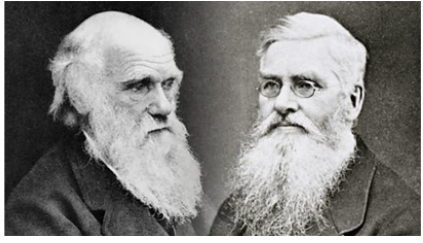
Incentives in Basic Science

- Basic scientific research advances our fundamental understanding of the world, but is not directly marketable
  - However, advances in basic research often serve as a key input in applied science (Nelson 1959, Arrow 1962)

Incentives in Basic Science

- Basic scientific research advances our fundamental understanding of the world, but is not directly marketable
  - However, advances in basic research often serve as a key input in applied science (Nelson 1959, Arrow 1962)
- Therefore, credit is the currency of scientific careers
  - Credit comes from disclosing findings first
  - Leads to priority races and fierce competition to be first





Competition in Science is a Double-Edged Sword

- Scientists compete to publish their findings first and establish priority. This competition can be good for science and society:
  - It can increase the pace of innovation
  - It induces scientists to disclose their work in order to get credit

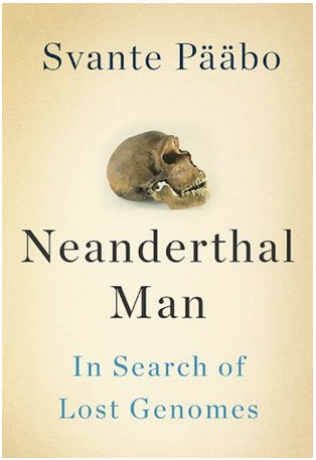
Competition in Science is a Double-Edged Sword

- Scientists compete to publish their findings first and establish priority. This competition can be good for science and society:
  - It can increase the pace of innovation
  - It induces scientists to disclose their work in order to get credit
- On the other hand, competition may have a dark side:
  - **Scientists may cut corners and reduce quality in their pursuit to publish first**

Example: Sequencing the Neanderthal Genome

"Hendrik's paper also illustrated a dilemma in science: doing all the analyses and experiments necessary to tell the complete story leaves you vulnerable to being beaten to the press...Even when you publish a better paper, you are seen as mopping up the details after someone who made the real breakthrough"

– Svante Pääbo, *Neanderthal Man: In Search of Lost Genomes*



This Project

Our goal is to answer two related questions:

1. Does competition in science lead to lower quality research?
2. If yes, what are the implications from a welfare and policy perspective?

This Project

Our goal is to answer two related questions:

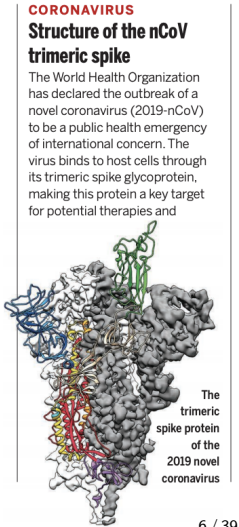
1. Does competition in science lead to lower quality research?
2. If yes, what are the implications from a welfare and policy perspective?

We do this by:

- Developing a model of competition and racing in science
- Testing the predictions of this model in the field of structural biology
- Exploring the welfare and policy implications of the priority premium in science

Why Structural Biology?

- Structural biology is the study of the three-dimensional structure of biological macromolecules (proteins)
- Important field of science!
- Uniquely detailed project-level data in the Protein Data Bank (PDB)
  - Objective measures of project quality
  - Project timelines
  - Links to publications
  - Other project details



Preview of Results

- Model predicts:
  - Most (ex-ante) important projects are more competitive, rushed, and lower quality

Preview of Results

- Model predicts:
  - Most (ex-ante) important projects are more competitive, rushed, and lower quality
- Empirical results:
  - High-potential projects are more competitive (multiple researchers working simultaneously)
  - High-potential projects are completed faster and are lower quality
  - Follow-on work ameliorates but does not eliminate the negative relationship between potential and quality
  - Quality magnitudes large enough to impact usefulness of projects for drug development

Preview of Results

- Model predicts:
  - Most (ex-ante) important projects are more competitive, rushed, and lower quality
- Empirical results:
  - High-potential projects are more competitive (multiple researchers working simultaneously)
  - High-potential projects are completed faster and are lower quality
  - Follow-on work ameliorates but does not eliminate the negative relationship between potential and quality
  - Quality magnitudes large enough to impact usefulness of projects for drug development
- Welfare implications:
  - Negative relationship between potential and quality is inconsistent with idealized first best
  - Reducing competition by reducing the priority premium does not necessarily improve welfare

Contributions to the Literature

- Sociology and economics of science
  - Merton (1957); Merton (1961); Hagstrom (1974); Dasgupta and Maskin (1987); Dasgupta and David (1994); Stephan (1996)
- Strategic behavior in patent and R&D races
  - Loury (1979); Lee and Wilde (1980); Dasgupta and Stiglitz (1980); Reinganum (1982); Fudenberg et al. (1983); Harris and Vickers (1985); Harris and Vickers (1987); Grossman and Shapiro (1987); Hopenhayn and Squintani (2016); Bobtcheff, Bolte, and Mariotti (2017)
- Scientific literature / concern about the impact of competition on science
  - Brown and Ramaswamy (2007); Fang and Casadevall (2005); Alberts et al. (2014)
- **Our (primary) contribution:** bring empirics to a largely theoretical literature

Agenda

1. Introduction
2. A Model of Competition and Quality in Science
3. Structural Biology and the PDB
4. Empirical Results
5. Welfare Considerations

Summary of the Model

- Projects vary in their ex-ante potential (P)

Summary of the Model

- Projects vary in their ex-ante potential (P)
- Scientists decide how long to work on a project (m), trading off improving the quality of their work (increasing $Q(m)$) against the threat of being scooped

Summary of the Model

- Projects vary in their ex-ante potential (P)
- Scientists decide how long to work on a project (m), trading off improving the quality of their work (increasing $Q(m)$) against the threat of being scooped
- **Key ingredient:** entry into projects is endogenous $\rightarrow$ there is more likely to be competition in high potential projects
  - Operationalize this by letting scientists choose costly I, probability of entry is $g(I)$

Summary of the Model

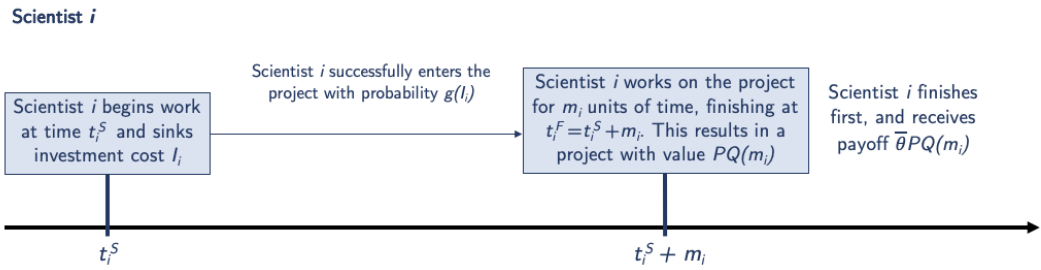
- Projects vary in their ex-ante potential (P)
- Scientists decide how long to work on a project (m), trading off improving the quality of their work (increasing $Q(m)$) against the threat of being scooped
- **Key ingredient:** entry into projects is endogenous $\rightarrow$ there is more likely to be competition in high potential projects
  - Operationalize this by letting scientists choose costly I, probability of entry is $g(I)$
- **Key result:** high potential projects will be executed with lower quality

Summary of the Model

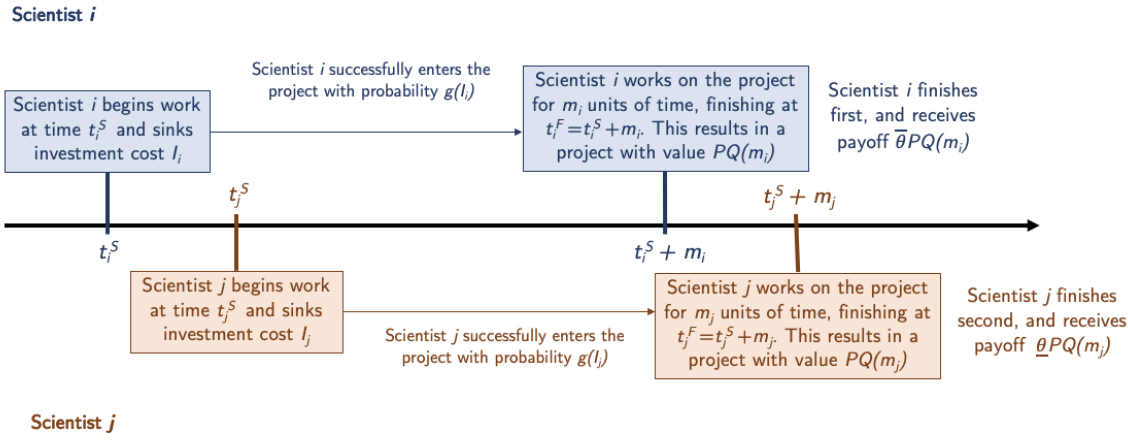
- Projects vary in their ex-ante potential (P)
- Scientists decide how long to work on a project (m), trading off improving the quality of their work (increasing $Q(m)$) against the threat of being scooped
- **Key ingredient:** entry into projects is endogenous $\rightarrow$ there is more likely to be competition in high potential projects
  - Operationalize this by letting scientists choose costly I, probability of entry is $g(I)$
- **Key result:** high potential projects will be executed with lower quality
- **Key policy lever:** the credit split between first and second-place finisher $(\overline{\theta}$ vs. $\underline{\theta})$

more detail

Timing


Scientist *i*
Scientist *i* begins work at time t_i^S and sinks investment cost I_i
Scientist *i* successfully enters 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
$t_i^S + m_i$

Timing


Scientist i
Scientist i begins work at time t_i^S and sinks investment cost I_i
Scientist i successfully enters 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_j^S
$t_i^S + m_i$
$t_j^S + m_j$
Scientist j begins work at time t_j^S and sinks investment cost I_j
Scientist j successfully enters 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 j finishes second, and receives payoff $\underline{\theta} PQ(m_j)$
Scientist j

Information

What does scientist i know about scientist j?

Information

What does scientist i know about scientist j?

- Knows that j entered with probability $g(I_j)$ (known in equilibrium)
- Believes that j's start time is uniformly distributed around her own start time:

$$t_j^S \sim U[t_i^S - \Delta, t_i^S + \Delta]$$

- Implication: the value of i's start time is not informative about whether she is ahead or behind

maturation FOC investment FOC

Key Propositions

- **Proposition 1.** $\frac{dI^*}{dP} > 0$ and $\frac{dg(I^*)}{dP} > 0$
  "high-potential projects generate more investment $\rightarrow$ are more competitive"

Key Propositions

- **Proposition 1.** $\frac{dI^*}{dP} > 0$ and $\frac{dg(I^*)}{dP} > 0$
  "high-potential projects generate more investment $\rightarrow$ are more competitive"
- **Proposition 2.** $\frac{dm^*}{dg} < 0$ and $\frac{dQ(m^*)}{dg} < 0$
  "competitive projects completed faster $\rightarrow$ are lower quality"

Key Propositions

- **Proposition 1.** $\frac{dI^*}{dP} > 0$ and $\frac{dg(I^*)}{dP} > 0$
  "high-potential projects generate more investment $\rightarrow$ are more competitive"
- **Proposition 2.** $\frac{dm^*}{dg} < 0$ and $\frac{dQ(m^*)}{dg} < 0$
  "competitive projects completed faster $\rightarrow$ are lower quality"
- **Proposition 3.** $\frac{dm^*}{dP} < 0$ and $\frac{dQ(m^*)}{dP} < 0$
  key model prediction: "high-potential projects completed faster $\rightarrow$ are lower quality"
  (comes directly from the chain rule)

1 Introduction

2 A Model of Competition and Quality in Science

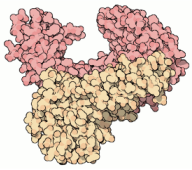
3 Structural Biology and the PDB

4 Empirical Results

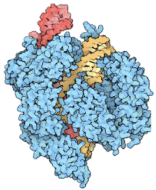
5 Welfare Considerations

What is Structural Biology?

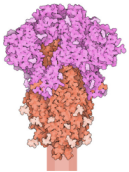
- The study of the molecular structure of macromolecules, especially proteins



HIV reverse transcriptase



CRISPR Cas9 protein



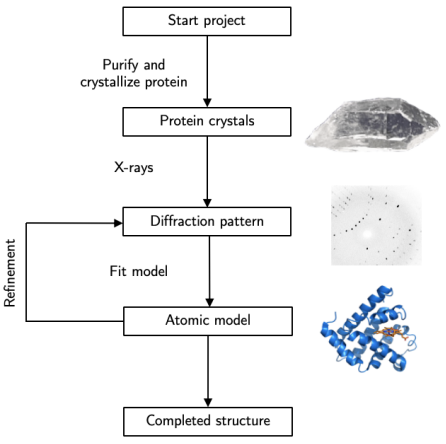
SARS-CoV-2 spike protein

- An important field of science, with applications in genetic diseases and drug development

How do Scientists Solve Protein Structures?

About 90% of proteins are solved using X-ray crystallography. This involves three steps:

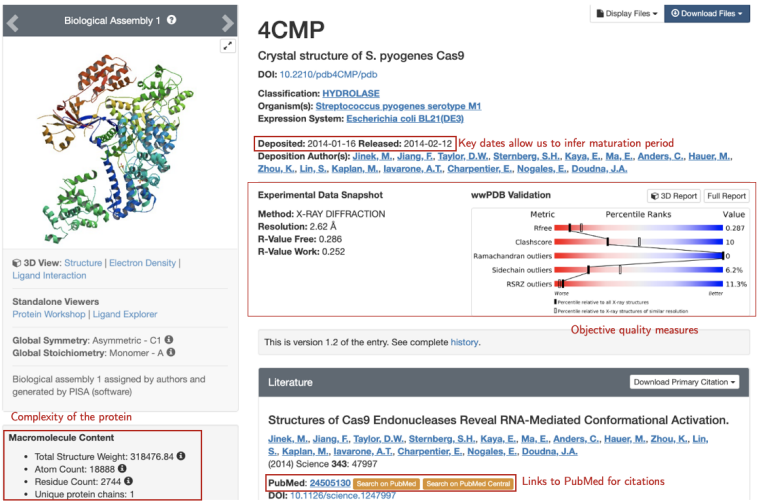
1. First, proteins are purified and crystallized
2. Next, the crystals are placed in an x-ray beam, which produces a diffraction pattern
3. Finally, the diffraction data is used to infer the structure. Biologists will "refine" their structure by comparing their model to the diffraction data, trying to minimize any discrepancies. Process is more "art than science" and luck plays a role



What is the Protein Data Bank?

- Established in 1971, the Protein Data Bank (PDB) is a database for 3D structural data of large biological molecules (proteins and nucleic acids)
- Most scientific journals and some funding agencies require scientists to submit their structure data to the PDB
- Today, the PDB contains 100,000+ structures, and is growing ~10% annually

Example PDB Entry - CRISPR-Associated Protein 9 (Cas9)

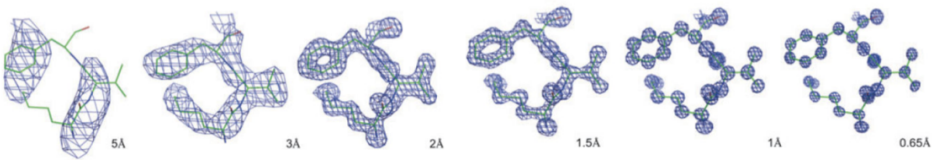


sample construction

Mapping to the Model: Quality

A unique feature of structural biology is the objective, ex-ante measures of project quality:

1. Refinement resolution: similar to resolution of a photograph

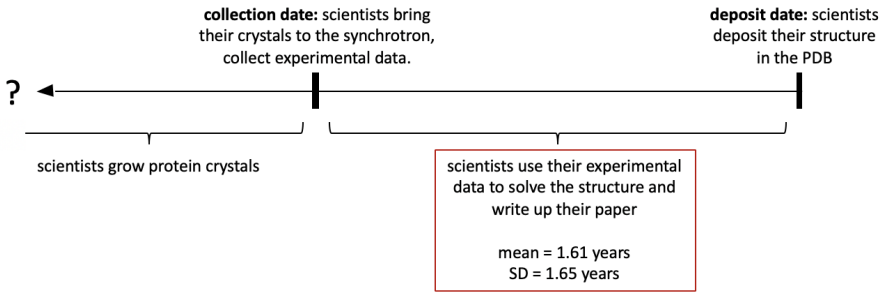


2. R-free: model fit, estimated on a holdout sample of the experimental data
3. Outliers: errors in the model based on chemical properties

Combine these outcomes into a standardized quality index (higher is better)

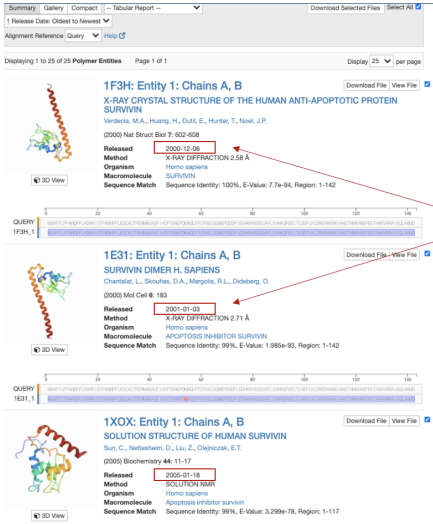
Mapping to the Model: Maturation

- We can actually observe time spent on project (maturation period):



Mapping to the Model: Competition

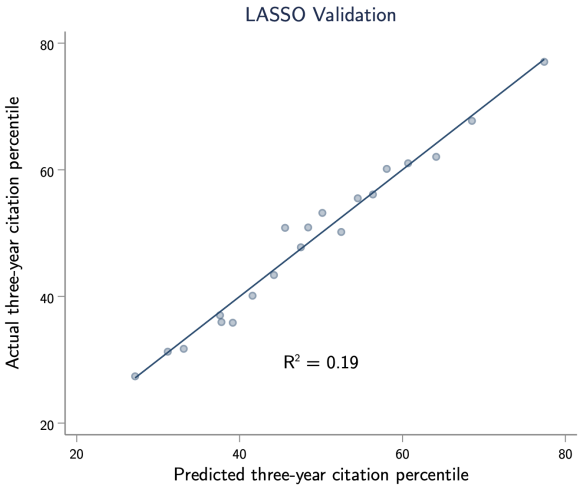
- The PDB uses amino acid sequence similarity to flag proteins that are identical
- Number of times the same protein is deposited (within two years) can proxy for competition
- Note that we are measuring ex-post realized competition, a noisy proxy for ex-ante competition



Mapping to the Model: Measuring and Predicting Potential in the PDB

- One way to measure potential: use ex-post citations (over some time window)
  - Problems: ex-post citations different than ex-ante potential, conflates potential and quality
- Alternatively: predict citations using only ex-ante characteristics of the structure
  - To avoid over-fitting, we use LASSO to select the model

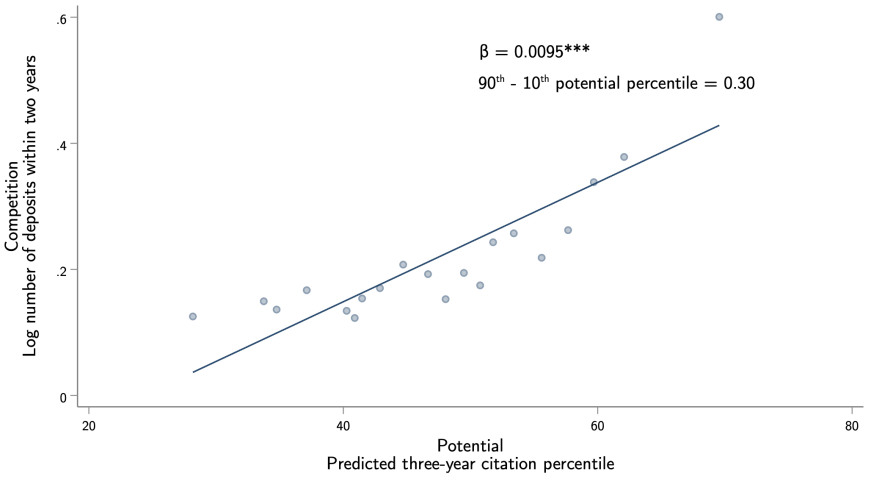
LASSO details



Agenda

1. Introduction
2. A Model of Competition and Quality in Science
3. Structural Biology and the PDB
4. Empirical Results
5. Welfare Considerations

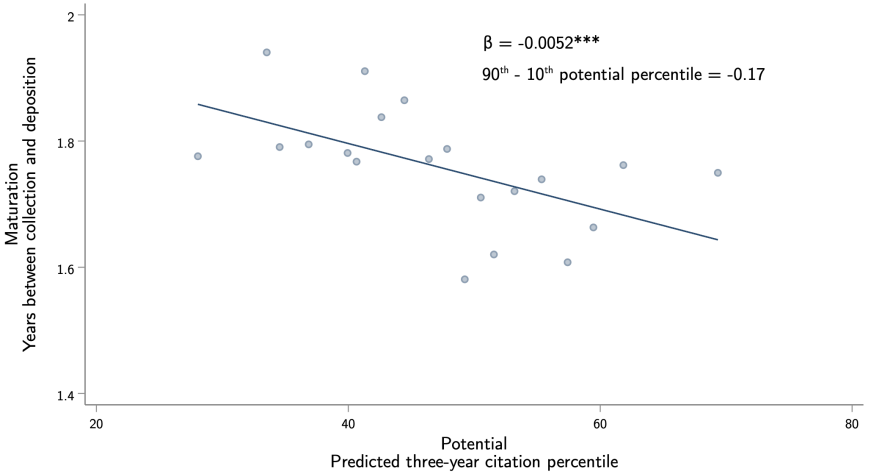
Proposition 1: High-Potential Projects are More Competitive



$$LogDepositsInCluster_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

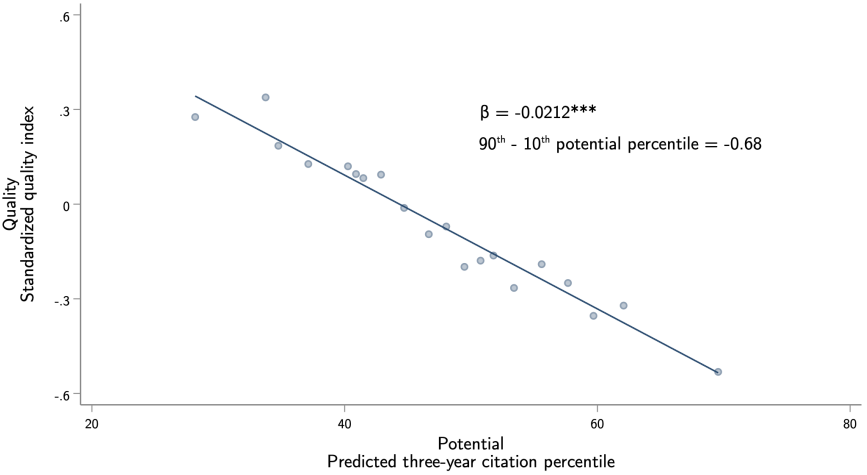
investment

Proposition 3: High-Potential Projects are Completed Faster...



$$Maturation_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

...So High-Potential Projects are Lower Quality



$$Quality_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

What About Project Complexity?

- If high P projects are also more complicated, this could drive our observed results
- Lower quality is driven by the difficulty / complexity of the project, not rushing

Strategy #1: Control for Complexity

- We are able to observe measures of molecule complexity in our data:
  - Molecular weight
  - Residue count
  - Atom site count

Strategy #1: Control for Complexity

- We are able to observe measures of molecule complexity in our data:
  - Molecular weight
  - Residue count
  - Atom site count
- Include these (and their squares), coefficient on potential remains stable:

Dependent variable:	Std. resolution	Std. R-free	Std. Rama. outliers	Std. quality index
Panel A. Without complexity controls				
Potential	-0.021***	-0.020***	-0.012***	-0.021***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.048	0.078	0.058	0.066
Panel B. With complexity controls				
Potential	-0.019***	-0.019***	-0.010***	-0.019***
	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.283	0.164	0.098	0.216
Observations	18,014	18,014	18,014	18,014

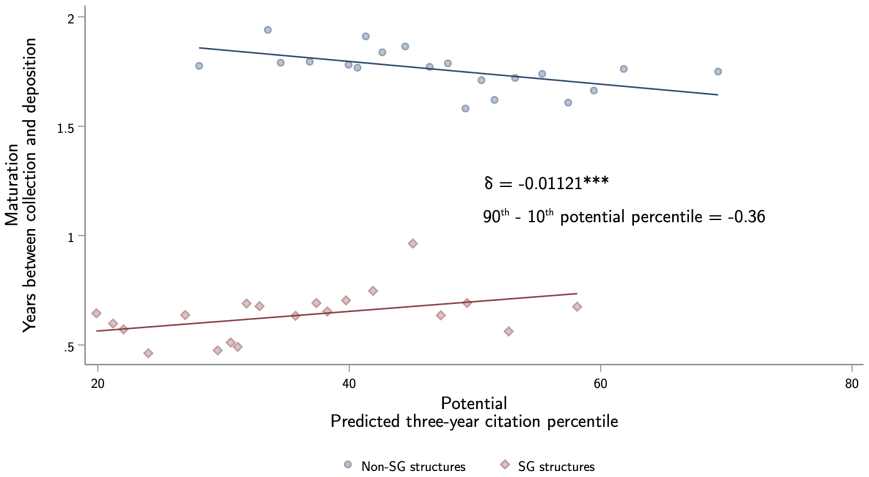
Strategy #2: Look at the Maturation Period

- If high-potential projects are low quality because they are complicated, we would expect they take *longer* to complete
- BUT we observe the opposite: high-potential projects are completed *faster*

Strategy #3: Structural Genomics Consortia

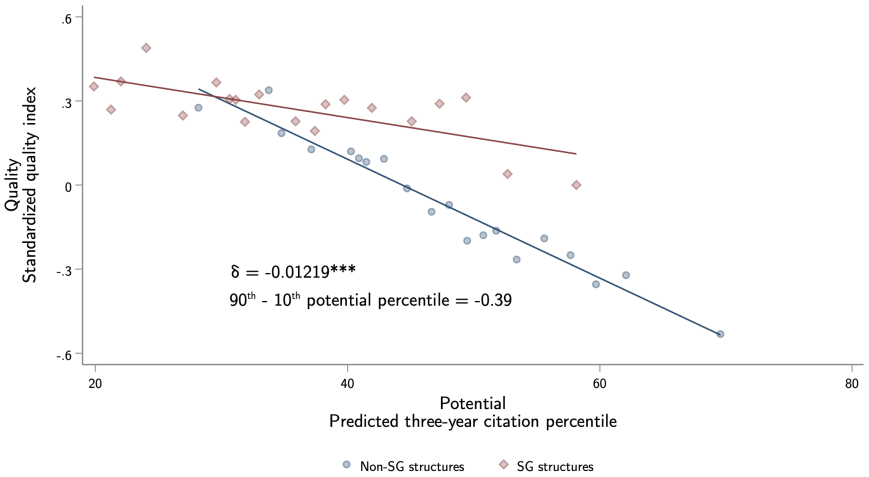
- Structural genomics consortia are publicly funded groups focused on achieving comprehensive coverage of the protein folding space
- Less focused on publishing and priority $\rightarrow$ competition is less important
- About 20% of structures in our sample were deposited by a structural genomics group

SG versus Non-SG Structures: Maturation



$$Maturation_{it} = \alpha + \beta PredictedCites_{it} + \gamma NonSG_{it} + \delta(PredictedCites_{it} * NonSG_{it}) + \tau_t + \epsilon_{it}$$

SG versus Non-SG Structures: Quality



$$Quality_{it} = \alpha + \beta PredictedCites_{it} + \gamma NonSG_{it} + \delta(PredictedCites_{it} * NonSG_{it}) + \tau_t + \epsilon_{it}$$

Proposition 2: Competitive Projects are Rushed and Lower Quality

- Could estimate:

$$Quality_{it} = \alpha + \beta LogDepositsInCluster_{it} + \tau_t + \epsilon_{it}$$

but this is a noisy measure of competition
$\rightarrow \beta$ will be attenuated

Dependent variable	Maturation	Std. quality index
Panel A. Ordinary least squares		
Competition	-0.144***	-0.044***
	(0.032)	(0.016)
Mean of dependent variable	1.75	-0.07
Observations	16,278	18,014

Proposition 2: Competitive Projects are Rushed and Lower Quality

- Could estimate:

$$Quality_{it} = \alpha + \beta LogDepositsInCluster_{it} + \tau_t + \epsilon_{it}$$

but this is a noisy measure of competition $\rightarrow \beta$ will be attenuated

- Alternatively: could instrument for competition using potential
  - Proposition 1 is the first stage
  - Proposition 3 is the reduced form

Dependent variable	Maturation	Std. quality index
Panel A. Ordinary least squares		
Competition	-0.144***	-0.044***
	(0.032)	(0.016)
Panel B. Two-stage least squares		
Competition	-0.673***	-2.311***
	(0.171)	(0.134)
First-stage *F* statistic	484.4	550.8
Mean of dependent variable	1.75	-0.07
Observations	16,278	18,014

Will Follow-on Work Fix the Problem?

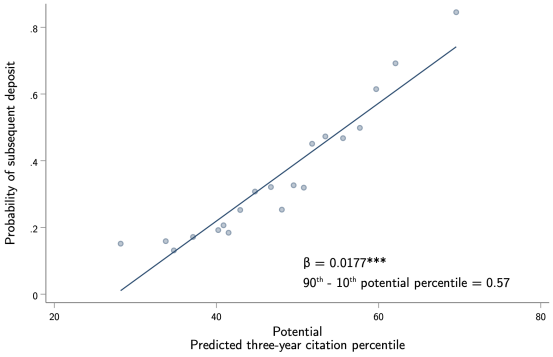
- In a standard quality ladder model, researchers could costlessly build on rushed, lower quality structures

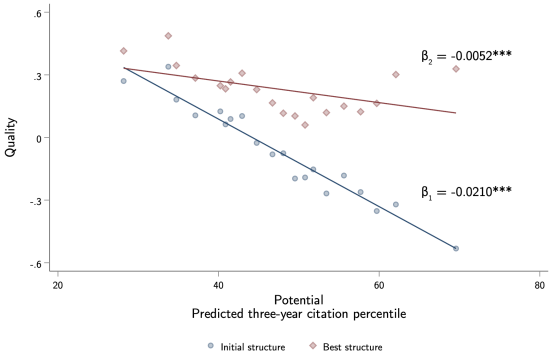
Will Follow-on Work Fix the Problem?

- In a standard quality ladder model, researchers could costlessly build on rushed, lower quality structures
- In our setting, making a marginal quality improvement requires re-sinking all the same costs (typically over a year of time and $100K)
  - Only worth fixing particularly bad / important structures
  - More efficient to do it well the first time

Follow-on Work Mitigates The Negative Relationship

- High potential low quality structures very likely to be re-deposited
- Enough to diminish the negative relationship between potential and quality



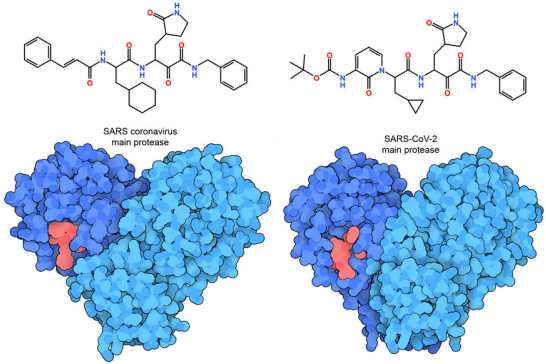


Does Quality Matter for Structure's Usefulness?

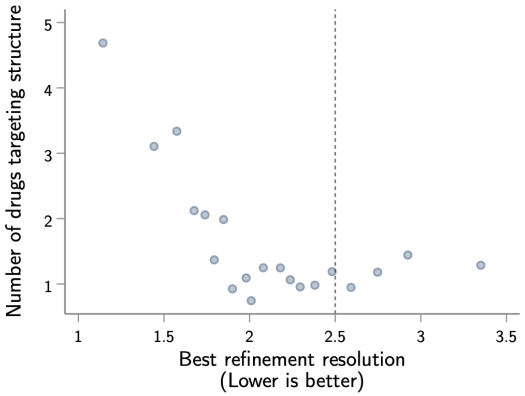
- Short answer: depends on the structure's use case
- For structure-based drug design, quality is important (Anderson 2003):
  - Resolution should be 2.5 Å or better (35% of non-SG structures don't meet this cutoff)
  - R-free should be 0.25 or better (45% of non-SG structures don't meet this cutoff)
- We will demonstrate that these thresholds appear to matter

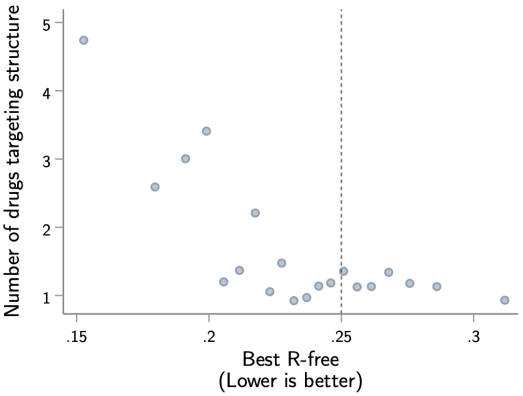
Linking Target Protein Structures and Drugs

- A drug target is the protein that the drug binds to, in order to have its effect
- Use data from DrugBank to link drugs to their targets, and targets to their PDB ID(s)



More Drug Development when Structures Exceed Quality Thresholds





Agenda

1. Introduction
2. A Model of Competition and Quality in Science
3. Structural Biology and the PDB
4. Empirical Results
5. Welfare Considerations

Welfare Losses Relative to First Best

- Scientist's objective:

$$\max_{I_i, m_i} \underbrace{g(I_i)}_{\Pr(i \text{ enters})} \times \underbrace{e^{-rm_i} PQ(m_i) \times \left[\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right]}_{\text{expected (private) return if } i \text{ enters}} - \underbrace{I_i}_{\text{cost to } i}$$

Welfare Losses Relative to First Best

- Scientist's objective:

$$\max_{I_i, m_i} \underbrace{g(I_i)}_{\Pr(i \text{ enters})} \times \underbrace{e^{-rm_i} PQ(m_i) \times \left[\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right]}_{\text{expected (private) return if } i \text{ enters}} - \underbrace{I_i}_{\text{cost to } i}$$

- Social planner's objective:

$$\max_{I,m} \underbrace{\left[1 - (1 - g(I))^2\right]}_{\Pr(\text{either } i \text{ or } j \text{ enters})} \times \underbrace{e^{-rm} kPQ(m)}_{\text{social return if either } i \text{ or } j \text{ enters}} - \underbrace{2I}_{\text{cost to } i \text{ and } j}$$

Welfare Losses Relative to First Best

- Scientist's objective:

$$\max_{I_i, m_i} \underbrace{g(I_i)}_{\Pr(i \text{ enters})} \times \underbrace{e^{-rm_i} PQ(m_i) \times \left[\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})\right]}_{\text{expected (private) return if } i \text{ enters}} - \underbrace{I_i}_{\text{cost to } i}$$

- Social planner's objective:

$$\max_{I,m} \underbrace{\left[1 - (1 - g(I))^2\right]}_{\Pr(\text{either } i \text{ or } j \text{ enters})} \times \underbrace{e^{-rm} kPQ(m)}_{\text{social return if either } i \text{ or } j \text{ enters}} - \underbrace{2I}_{\text{cost to } i \text{ and } j}$$

- Consequences:
  - Scientists rush: $m^{C^*} < m^{SP^*}$
  - If k is large, investment is too low: $I^{C^*} < I^{SP^*}$
  - Note: if priority rewards are equal ($\overline{\theta} = \underline{\theta}$), then we get $m^{C^*} = m^{SP^*}$

Optimal Policy

- Reasonable policy lever: rewards granted to the winner ($\overline{\theta}$) and loser ($\underline{\theta}$)
  - Several journals have discussed adding a "scoop protection policy" in the past few years

Optimal Policy

- Reasonable policy lever: rewards granted to the winner ($\overline{\theta}$) and loser ($\underline{\theta}$)
  - Several journals have discussed adding a "scoop protection policy" in the past few years
- Making priority rewards more equal ($\overline{\theta} \downarrow$ and $\underline{\theta} \uparrow$) leads to a potential tradeoff:
  - Increases maturation times, getting closer to the social optimum
  - May decrease investment, distorting us away from the social optimum example

Optimal Policy

- Reasonable policy lever: rewards granted to the winner ($\overline{\theta}$) and loser ($\underline{\theta}$)
  - Several journals have discussed adding a "scoop protection policy" in the past few years
- Making priority rewards more equal ($\overline{\theta} \downarrow$ and $\underline{\theta} \uparrow$) leads to a potential tradeoff:
  - Increases maturation times, getting closer to the social optimum
  - May decrease investment, distorting us away from the social optimum example
- Tradeoff implies that optimal priority rewards may be lopsided ($\overline{\theta}^* > \underline{\theta}^*$)
  - Suggests that the negative relationship between potential and quality is not inconsistent with a constrained second-best solution
  - Hill and Stein (2020) find that winner's credit share $\overline{\theta}/(\overline{\theta} + \underline{\theta}) = 0.55$ in structural biology
  - However, surveyed scientists believe the winner's credit share is about 0.7

Alternative Policy: Ending Races Early

- If races ended when the first team successfully entered the project, there would be no maturation distortion (no competition $\rightarrow$ no need to rush)

Alternative Policy: Ending Races Early

- If races ended when the first team successfully entered the project, there would be no maturation distortion (no competition $\rightarrow$ no need to rush)
- In fact, in the 1970s researchers used to publish their protein crystals, which signaled that other teams should "back off"
  - "There was a tradition that if someone had produced crystals of something, they were usually left alone to solve the problem" (Ramakrishnan, 2018)

Alternative Policy: Ending Races Early

- If races ended when the first team successfully entered the project, there would be no maturation distortion (no competition $\rightarrow$ no need to rush)
- In fact, in the 1970s researchers used to publish their protein crystals, which signaled that other teams should "back off"
  - "There was a tradition that if someone had produced crystals of something, they were usually left alone to solve the problem" (Ramakrishnan, 2018)
- This norm collapsed once the field became too large, but still interesting to note that the field "organically" solved this problem at one point

Conclusions and Future Work

- Calibration of the optimal priority rewards is beyond the scope of this project
- Competition likely affects science in ways we have not considered here:
  - May reduce collaboration and free sharing of ideas
  - Impacts who enters certain fields and who is deterred
- Brings up questions of alternative models of science:
  - More collaborative models: Protein Structure Initiative, Human Genome Project

Choosing Maturation

After entering the project, researcher i chooses maturation:

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

where

- r is the discount rate
- $\pi(m_i, m_j)$ is probability i publishes first
- $\overline{\theta}$, $\underline{\theta}$ are first, second place credit shares

Choosing Maturation

After entering the project, researcher i chooses maturation:

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

where

- r is the discount rate
- $\pi(m_i, m_j)$ is probability i publishes first
- $\overline{\theta}$, $\underline{\theta}$ are first, second place credit shares

First-order condition:

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

back

Choosing Investment

When deciding how much to invest in entry, researcher i solves:

$$\max_{I_i} \quad \underbrace{g(I_i)}_{\Pr(\text{enter})} \underbrace{e^{-rm_i} PQ(m_i^*)}_{\text{PDV of project}} \left[\underbrace{\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})}_{\text{expected credit share}} \right] - \underbrace{I_i}_{\text{cost}}$$

where

- r is the discount rate
- $\pi(m_i, m_j)$ is probability i publishes first
- $\overline{\theta}$, $\underline{\theta}$ are first, second place credit shares

Choosing Investment

When deciding how much to invest in entry, researcher i solves:

$$\max_{I_i} \underbrace{g(I_i)}_{\Pr(\text{enter})} \underbrace{e^{-rm_i} PQ(m_i^*)}_{\text{PDV of project}} \left[\underbrace{\overline{\theta} - \frac{1}{2} g(I_j)(\overline{\theta} - \underline{\theta})}_{\text{expected credit share}} \right] - \underbrace{I_i}_{\text{cost}}$$

where

- r is the discount rate
- $\pi(m_i, m_j)$ is probability i publishes first
- $\overline{\theta}$, $\underline{\theta}$ are first, second place credit shares

First-order condition:

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

back

Sample Construction

We start with the universe of PDB x-ray structures from 1971 to 2018 (128,876 structures, 71,685 papers)

- Restrict to single structure-paper pairs (35,625 obs)
- Restrict to new structure discoveries (26,620 obs)
- Restrict to non-missing outcomes (22,308 obs)

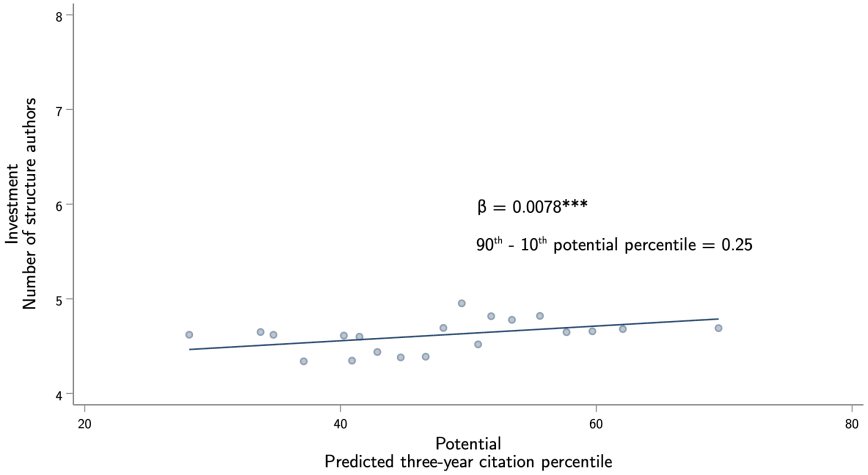
back

LASSO Details

- LASSO predictors include:
  - Macromolecule type (protein, DNA, RNA)
  - Classification (membrane protein, oxygen transport)
  - Taxonomy (homo sapiens, e. coli, influenza virus)
  - Gene linkage (gag-pol gene, CA2 gene)
  - Prior citations to protein (papers prior to structure discovery, from UniProt)
  - Publication year

back

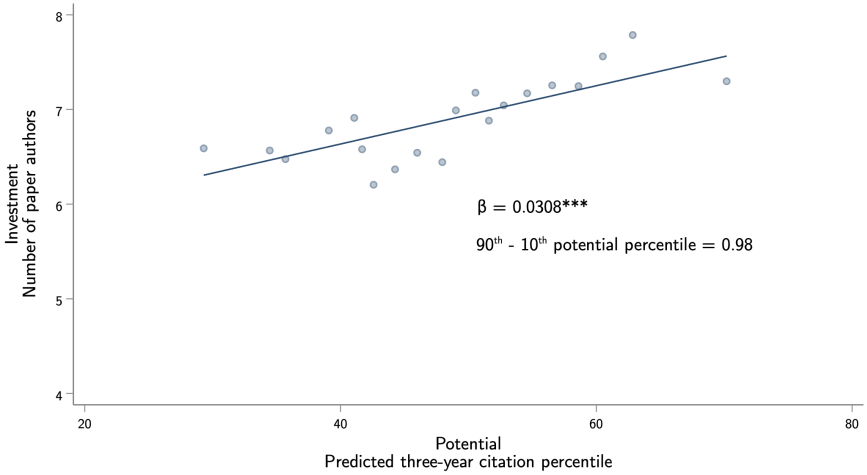
Proposition 1: High-Potential Projects Generate More Investment



$$NumberAuthors_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

back

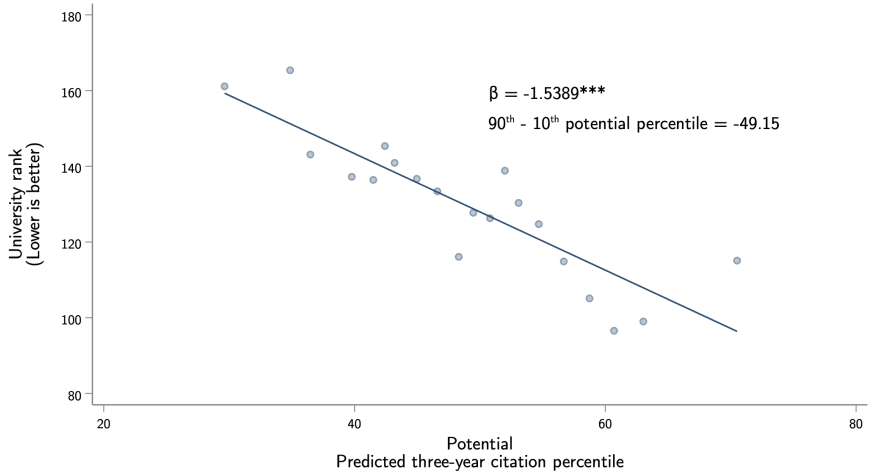
Proposition 1: High-Potential Projects Generate More Investment



$$NumberAuthors_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

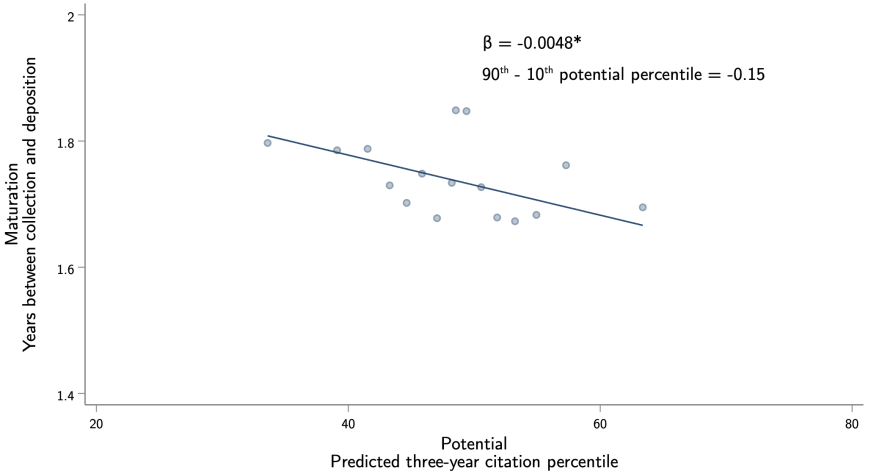
back

Lab Sorting



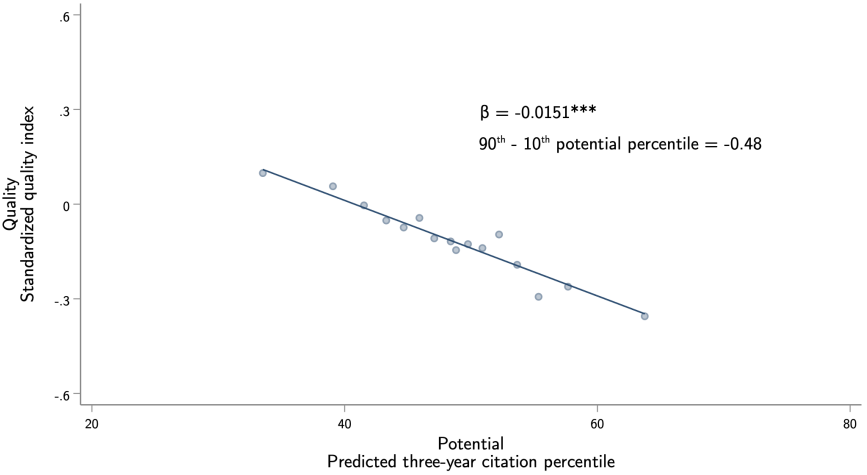
$$UniversityRank_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \gamma_l + \epsilon_{it}$$

Maturation vs. Potential Within Labs



$$Maturation_{ijtl} = \alpha + \beta PredictedCites_{ijtl} + \tau_t + \gamma_l + \epsilon_{ijtl}$$

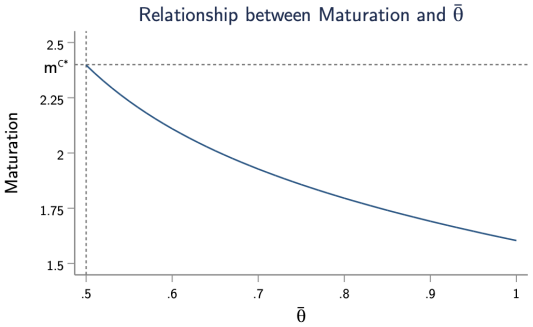
Quality versus Potential Within Labs

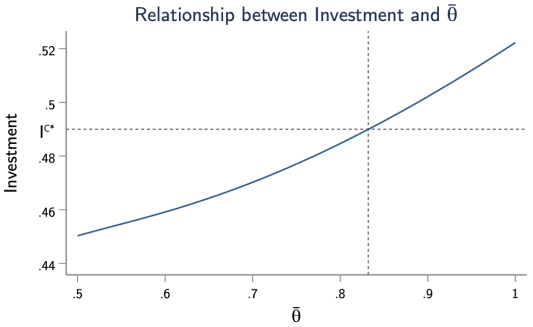


$$Quality_{itl} = \alpha + \beta PredictedCites_{itl} + \tau_t + \gamma_l + \epsilon_{itl}$$

Optimal Policy: Example

Let $\overline{\theta} + \underline{\theta} = 1$. If SP makes $\overline{\theta}$ and $\underline{\theta}$ more equal ($\overline{\theta} \to 0.5$), it creates tradeoff between longer maturation (good) and lower investment (bad):

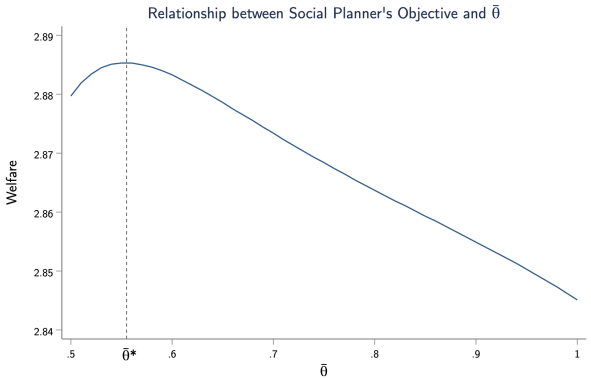




back

Optimal Policy: Example

Implies that the optimal $\overline{\theta}^*$ can be greater than $\frac{1}{2}$:



This will lead to racing and negative relationship between potential and quality, even though we are at the second best