Week 5: The Economics of Science

Carolyn Stein

Econ 220C: Topics in Industrial Organization

Science as a non-market incentive

A new economics of science

Scientific norms and their effects

- Mertonian origins
- Hill and Stein (2023)
- Hill and Stein (2024)

Basic vs. applied science

- **Basic science** seeks to expand human knowledge, but not to create or invent something. There is no obvious commercial value to the result of basic research
- **Applied science** seeks to solve practical problems and often yields something that is commercially valuable
- Basic research is important: "People cannot foresee the future well enough to predict what's going to develop from basic research. If we only did applied research, we would still be making better spears." – George Smoot
- But how do we incentivize people to produce it?

Basic vs. applied science: an example

Basic science

Applied science

Francisco Mojica studied bacteria in Spanish salt flats in the 80s and 90s. He noticed odd bits of repeated DNA in these bacteria...which paved the way for CRISPR

molecular microbiology

Long stretches of short tandem repeats are present in the largest replicons of the *Archaea Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning

F.J.M. Mojica, C. Ferrer, G. Juez, F. Rodríguez-Valera

First published: July 1995 | https://doi.org/10.1111/j.1365-2958.1995.mmi_17010085.x | Citations: 205

Vertex pharmaceuticals has developed a CRISPR gene editing treatment to cure patients with sickle-cell anemia, which is currently under FDA review

Academic freedom can be a strong motivator

- Stern (2004) "Do Scientists Pay to Be Scientists?" studies whether researchers take a pay cut to be given more scientific freedom
- Surveys biology post-docs with multiple job offers and collects characteristics of the jobs
  - Salary
  - Measures of scientific freedom (allowed to publish discoveries, allowed to continue postdoc projects, whether there are incentives to publish)
- Argues that all job offers should be roughly similarly attractive (formal offers only issued if candidate is serious)
- Can the run a hedonic regression with individual fixed effects

Researchers do value academic freedom

Results suggest postdocs accept a 20% pay cut in exchange for more academic freedom

Table 3 **Hedonic Wage Regression: Overall Sample Dependent Variable = LN(SALARY), # of Observations = 121**

	Permission to publish			Combination model	Science index model	
	(3–1)	(3–2)	(3–3)	(3–4)	(3–5)	(3–6)
	Baseline (NO FE)	Baseline (w/FE)	Full model (w/FE)	Full Model (w/FE)	Full Model (w/FE)	Full Model (w/FE)
PERMIT_PUB	0.027	**−0.266**	**−0.191**	−0.089		
	(0.186)	**(0.114)**	**(0.105)**	(0.103)		
CONTINUE RESEARCH				**−0.134**		
				(0.060)		
INCENT_PUB				−0.036		
				(0.028)		
SCIENCE INDEX					**−0.114**	−0.078
					(0.053)	(0.057)
EQUIPMENT				**0.053**	**0.057**	**0.053**
				(0.033)	**(0.030)**	**(0.031)**
CONTROLS						
PROMOTION			**0.041**	**0.046**	**0.042**	0.031
			(0.025)	**(0.021)**	**(0.021)**	(0.023)
STOCK_DUMMY			**0.196**	**0.234**	**0.260**	**0.190**
			(0.085)	**(0.074)**	**(0.067)**	**(0.077)**
ACCEPTED JOB			−0.013	0.002	−0.0001	−0.002
			(0.040)	(0.043)	(0.043)	(0.044)
JOBTYPE CONTROLS	no	no	yes (5; Sig.)	no	no	yes (5)
Individual fixed effects	no	yes (52; Sig.)	yes (52; Sig.)	yes (52; Sig.)	yes (52; Sig.)	yes (52; Sig.)
R-squared	0.001	0.915	0.955	0.958	0.954	0.958

Notes. Only persons with multiple job offers are included.

How else do we encourage basic research?

Many scientific norms can be viewed through the lens of providing incentives to engage in basic research:

- Grants
- Prizes
- Eponymy

All designed to compensate researchers. Maybe not with profits, but with credit, acclaim, etc.

"My love of natural science...has been much aided by the ambition to be esteemed by my fellow naturalists" – Charles Darwin

Science as a non-market incentive

A new economics of science

Scientific norms and their effects

- Mertonian origins
- Hill and Stein (2023)
- Hill and Stein (2024)

Dasgupta and David (1994)

- Economists have much to say about technological innovation but much less to say about basic science
- The "old" economics of science focused on three things:
  1. It is difficult to predict what basic science will be useful
  2. Property rights in basic science are difficult to enforce
  3. Thus, we expect market failures / underinvestment in basic science
- The "new" economics of science ought to go deeper:
  1. Understand scientific norms and their effects (an area richly studied by sociologists)
  2. Distinguish between codified vs. tacit knowledge, and think about incentives for disclosure and diffusion
  3. Think about the allocation across research areas and questions

Science as a non-market incentive

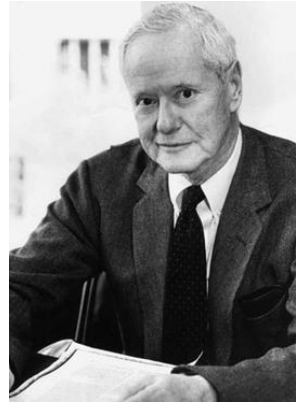
A new economics of science

Scientific norms and their effects

- Mertonian origins
- Hill and Stein (2023)
- Hill and Stein (2024)

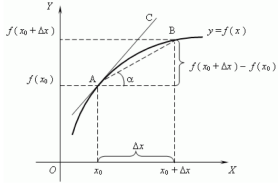
The importance of credit and recognition

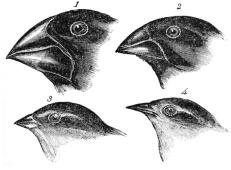
"In short, property rights in science become whittled down to just this one: the **recognition by others** of the scientist's distinctive part in having brought the result into being."
- Robert K. Merton (1957)

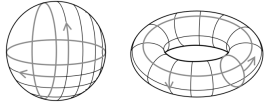


Priority in scientific discovery

- **Priority**: Credit given to the individual who *first* makes a scientific discovery.
- If being first yields more credit, not surprising that there are often fierce disputes over priority
- Notable scientific races and priority disputes:
  - Newton versus Leibniz - Calculus
  - Darwin versus Wallace - Natural Selection and Evolution
  - Perelman versus Yau, Zhu, and Cao - Proof of the Poincaré Conjecture
- Merton (1961) assembles 264 cases of "multiple discovery"







Science as a non-market incentive

A new economics of science

Scientific norms and their effects

- Mertonian origins
- Hill and Stein (2023)
- Hill and Stein (2024)

Scooped! Estimating rewards for priority in science

1. What is the causal effect of getting scooped?
   - Short-run effect on project: Publication, journal placement, and citations
   - Long-run effect on career: Future productivity of scientists

Scooped! Estimating rewards for priority in science

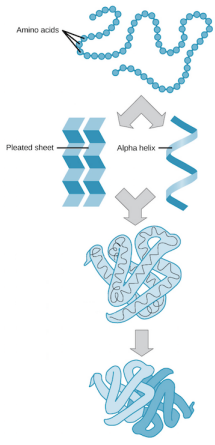
1. What is the causal effect of getting scooped?
   - Short-run effect on project: Publication, journal placement, and citations
   - Long-run effect on career: Future productivity of scientists
2. Does the priority reward system reinforce inequality in science? (Matthew Effect)
   - What drives citations: being first or being famous?

Key empirical challenges

1. Need a setting with well-defined problems and "one right answer."
2. Need an objective measure of scientific proximity.
3. Need a view of potential abandonments prior to publication.

What is structural biology?

- Structural biologists determine the molecular structure of proteins, DNA, and RNA.
- Proteins carry out most of the functions within cells, and often "form determines function."
- Structures are solved by X-ray crystallography. Successful experiments result in diffraction data and a model that describes the protein shape.

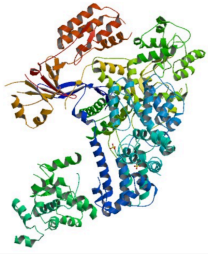


The Protein Data Bank

- The Protein Data Bank (PDB) contains structural data of 100,000+ proteins and meta-data about projects.
- Major scientific journals require scientists to submit their structure data to the PDB before publication.
- All structures are deposited confidentially a few months before article publication.
- Bioinformatics algorithm links projects with identical biological features.

PDB example: Cas-9

Biological Assembly 1



Macromolecule Content

- Total Structure Weight: 318476.84
- Atom Count: 18888
- Residue Count: 2744
- Unique protein chains: 1

Display Files | Download Files

4CMP unique structure ID

Crystal structure of S. pyogenes Cas9

DOI: 10.2210/pdb4CMP/pdb

Classification: HYDROLASE
Organism(s): Streptococcus pyogenes serotype M1
Expression System: Escherichia coli BL21(DE3)

Deposited: 2014-01-16 **Released:** 2014-02-12 key dates
Deposition Author(s): Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A.

Literature — Download Primary Citation

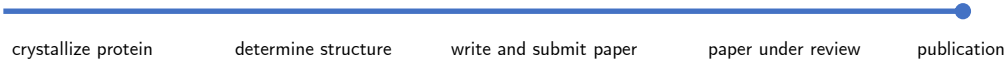
Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A.
(2014) Science **343**: 47997

PubMed: 24505130 Search on PubMed Search on PubMed Central
DOI: 10.1126/science.1247997
Primary Citation of Related Structures:
4OGE, 4OGC, 4CMQ

PubMed Abstract:
Type II CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated) systems use an RNA-guided DNA endonuclease, Cas9, to generate double-strand breaks in invasive DNA during an adaptive bacterial immune response. Cas9 h ...

Project timeline



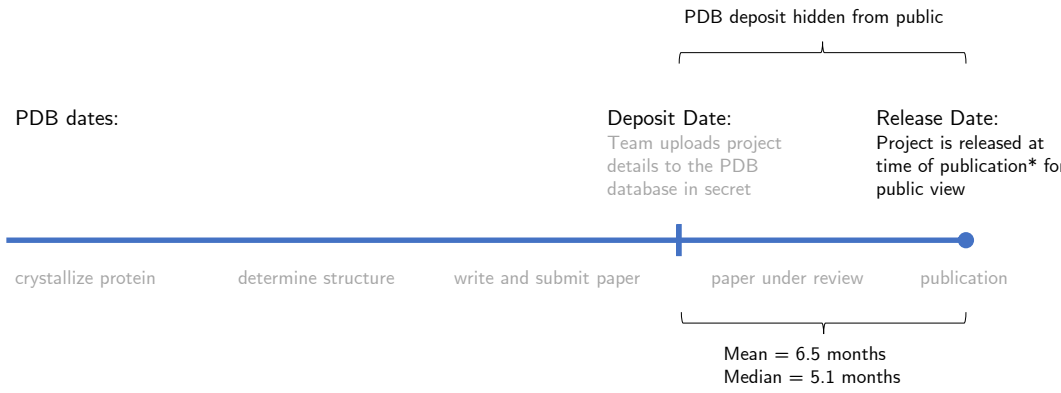
Project timeline

PDB dates:

Deposit Date:
Team uploads project details to the PDB database in secret

crystallize protein | determine structure | write and submit paper | paper under review | publication

Project timeline

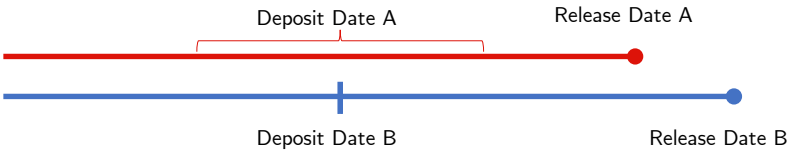


*If project goes unpublished, data is released publicly after one year

Scoop definition

Rules:
1. Take two projects that have identical sequence, different authors.
2. Assert that both projects are deposited before the first project is released.
3. Call the first to release the winner, call the second project "scooped."

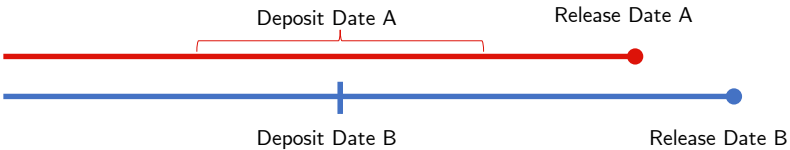
Scenario 1: Project A scoops Project B



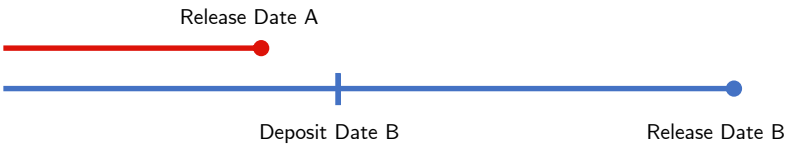
Scoop definition

Rules:
1. Take two projects that have identical sequence, different authors.
2. Assert that both projects are deposited before the first project is released.
3. Call the first to release the winner, call the second project "scooped."

Scenario 1: Project A scoops Project B

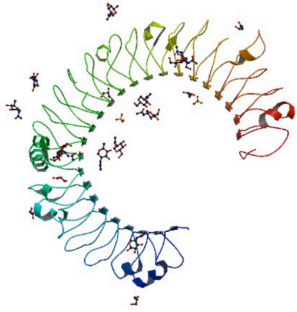


Scenario 2: Project A and Project B are excluded from racing sample



Example race: Toll-like receptor 3

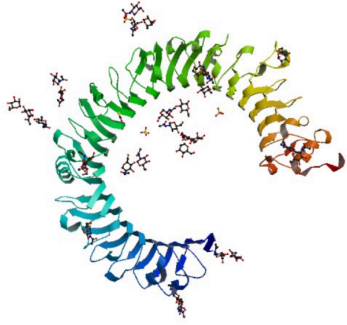
Winning Deposit: 1ZIW



Affiliation: Scripps Research Institute
Deposit Date: April 27, 2005
Release Date: June 28, 2005

Journal: *Science*
Journal Impact Factor: 30.9
5-year Citations: 196

Scooped Deposit: 2A0Z



Affiliation: National Institutes of Health
Deposit Date: June 27, 2005
Release Date: August 2, 2005

Journal: *PNAS*
Journal Impact Factor: 10.2
5-year Citations: 129

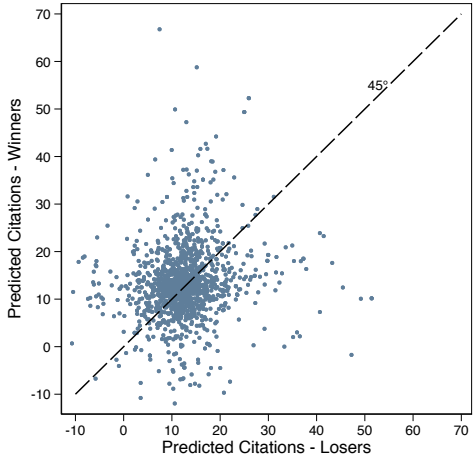
Predicted citation balance

Race winners are not randomly assigned, but seem highly unpredictable.

Lasso model of predicted citations:

- Team size and age
- Past deposits and publications
- University rank and location

Difference in predicted citations: 0.66 (p-value = 0.076)



Estimating the scoop penalty

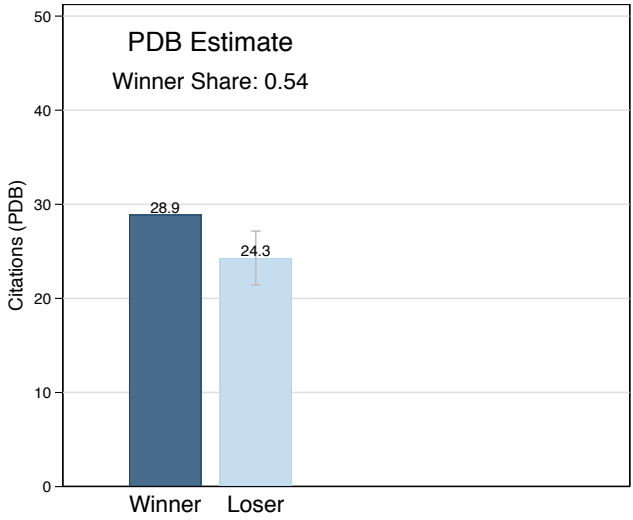
- Basic specification: For deposit i of protein (race) p:

$$Y_{ip} = \alpha + \beta Scooped_i + X_i'\delta + \gamma_p + \epsilon_{ip}$$

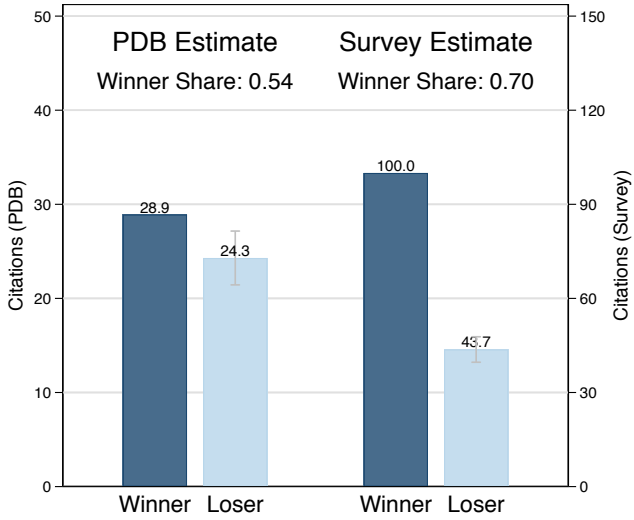
where

- $Scooped_i$ is a dummy for losing priority race.
- γ_p is the coefficient on a protein (i.e. race) fixed effect.
- X_i is a vector of individual and lab controls selected by PDS-Lasso method (Belloni et al. 2014).

Citation penalty



Citation penalty



Scoop penalty: alternative outcomes

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls					
Scooped	-0.025	-0.192***	-0.066***	-0.245***	-0.037***
	(0.015)	(0.044)	(0.020)	(0.071)	(0.014)
Panel B. Base controls					
Scooped	-0.026**	-0.183***	-0.064***	-0.216***	-0.028**
	(0.013)	(0.044)	(0.021)	(0.063)	(0.014)
Panel C. PDS-Lasso selected controls					
Scooped	-0.026***	-0.186***	-0.063***	-0.208***	-0.036***
	(0.010)	(0.032)	(0.015)	(0.045)	(0.010)
Winner Y mean	0.879	-0.027	0.320	28.830	0.149
Observations	3,279	3,279	3,279	2,514	2,514

Notes: This table presents regression estimates of the scoop penalty, following equation 2 in the text. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. Column 4 regression uses asinh(five-year citations) as the dependent variable, but Winner Y Mean is reported in levels for ease of interpretation.

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

The long-run consequences of being scooped

- Long run outcomes (excluding winning/scooped paper):
  - Active in PDB five years later
  - Total publications - five years
  - Total citations - five years
- Estimate for scientist s, deposit i, for protein (race) p:

$$Y_{isp} = \alpha + \beta Scooped_{is} + X'_{is}\delta + \gamma_p + \epsilon_{isp}$$

- Estimate separately for novices ($<$1 year of PDB experience) and veterans.

Long-run results

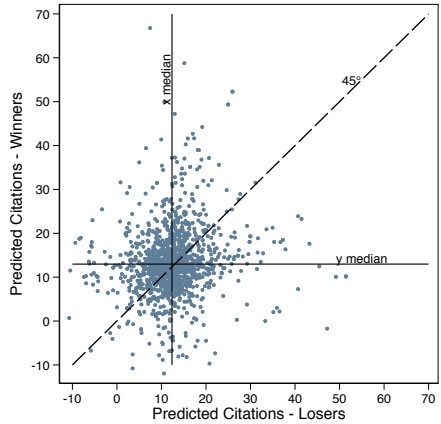
Dependent variable	Any PubMed within five years (1)	Any PDB within five years (2)	Total count within five years after race: PubMed publications (3)	PDB publications (4)	Top-ten publications (5)	Citation-weighted publications (6)	Top-10% cited publications (7)
Panel A. All scientists							
Scooped	-0.018***	-0.042***	-1.165	-0.085	-0.114	-0.172***	-0.414**
	(0.006)	(0.010)	(1.051)	(0.220)	(0.100)	(0.044)	(0.180)
Winner Y mean	0.841	0.702	45.869	7.154	3.610	497.203	7.741
Observations	8,624	8,624	8,624	8,624	8,624	6,484	6,484
Panel B. Novices							
Scooped	-0.057***	-0.040**	-0.021	0.003	0.104	-0.321***	-0.102
	(0.018)	(0.019)	(0.276)	(0.168)	(0.068)	(0.103)	(0.109)
Winner Y mean	0.469	0.356	4.243	1.890	0.616	75.691	1.165
Observations	2,033	2,033	2,033	2,033	2,033	1,529	1,529
Panel C. Veterans							
Scooped	-0.006*	-0.040***	-1.200	-0.176	-0.197	-0.130***	-0.568**
	(0.003)	(0.012)	(1.556)	(0.308)	(0.144)	(0.043)	(0.252)
Winner Y mean	0.990	0.839	61.681	9.261	4.787	667.421	10.388
Observations	5,821	5,821	5,821	5,821	5,821	4,378	4,378

Priority and inequality

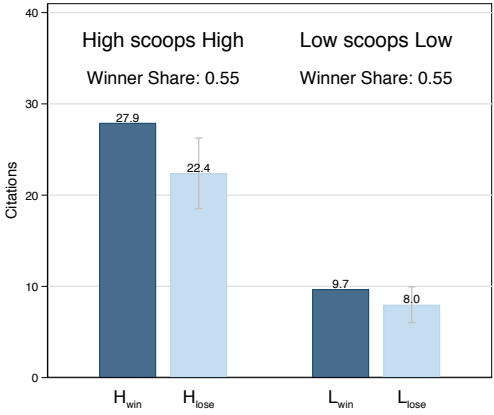
- Merton proposes two key drivers of academic attention:
  - Priority
  - Matthew Effect
- We test which of these effects dominates by comparing citations in races between high- and low-reputation teams.
- See the statistical discrimination model in the paper.

Defining reputation

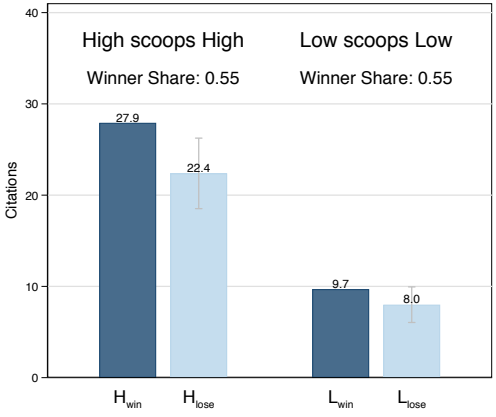
- Define pre-existing reputation using LASSO-generated predicted citations.
- Define H teams as those with above median predicted citations and L teams as those with below median.

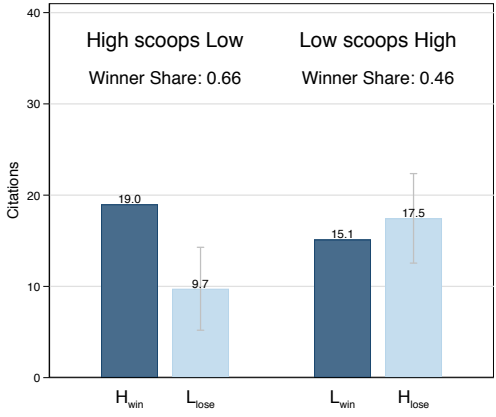


Evenly-matched and mismatched races



Evenly-matched and mismatched races





Conclusion

Getting scooped lowers citations, but rewards are more evenly distributed than previously thought.

Normative implications: Is the premium for priority too large or too small?

- Priority may incentivize effort and timely disclosure.
- Racing may incentivize speed at the expense of quality and transparency.

Science as a non-market incentive

A new economics of science

Scientific norms and their effects

- Mertonian origins
- Hill and Stein (2023)
- Hill and Stein (2024)

Competition and quality in science

- 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

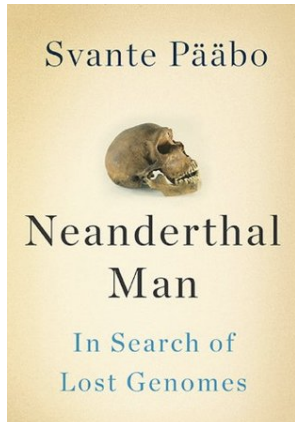
Competition and quality in science

- 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
  - **Focus of this project**

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*



Summary of the model

- Projects vary in their ex-ante potential

Summary of the model

- Projects vary in their ex-ante potential
- Scientists decide how long to work on a project (maturation), trading off improving the quality of their work against the threat of being scooped (Bobtcheff et. al 2017)

Summary of the model

- Projects vary in their ex-ante potential
- Scientists decide how long to work on a project (maturation), trading off improving the quality of their work against the threat of being scooped (Bobtcheff et. al 2017)
- **Key ingredient:** entry into projects is endogenous $\rightarrow$ more likely to be competition in high potential projects

Summary of the model

- Projects vary in their ex-ante potential
- Scientists decide how long to work on a project (maturation), trading off improving the quality of their work against the threat of being scooped (Bobtcheff et. al 2017)
- **Key ingredient:** entry into projects is endogenous $\rightarrow$ more likely to be competition in high potential projects
- **Key result:** high potential projects will be executed with lower quality

Key propositions

- **Proposition 1:**
  High potential projects are more attractive to enter $\rightarrow$ are more competitive

Key propositions

- **Proposition 1:**
  High potential projects are more attractive to enter $\rightarrow$ are more competitive
- **Proposition 2:**
  Competitive projects completed faster $\rightarrow$ are lower quality

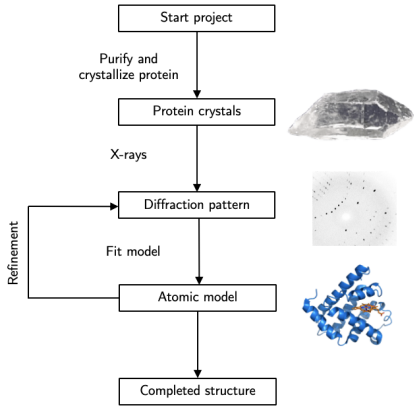
Key propositions

- **Proposition 1:**
  High potential projects are more attractive to enter $\rightarrow$ are more competitive
- **Proposition 2:**
  Competitive projects completed faster $\rightarrow$ are lower quality
- **Proposition 3 (key model prediction):**
  High potential projects completed faster $\rightarrow$ are lower quality

How do scientists solve protein structures?

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

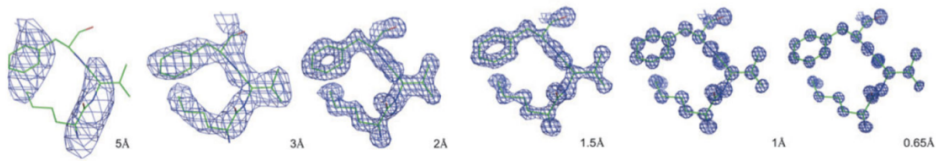
1. First, proteins are purified and crystallized
2. Next, the crystals are placed in an x-ray beam, which produces a diffraction pattern
3. Finally, the diffraction data is used to infer the structure. Biologists will "refine" their structure by comparing their model to the diffraction data, trying to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



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

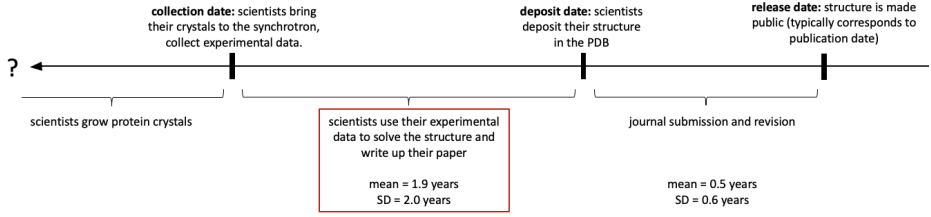


2. R-free: model fit, estimated on a holdout sample of the experimental data
3. Outlier share: errors in the model based on chemical properties

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):

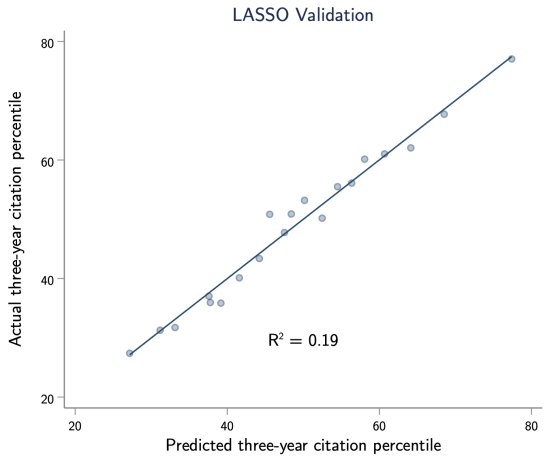


Mapping to the model: competition

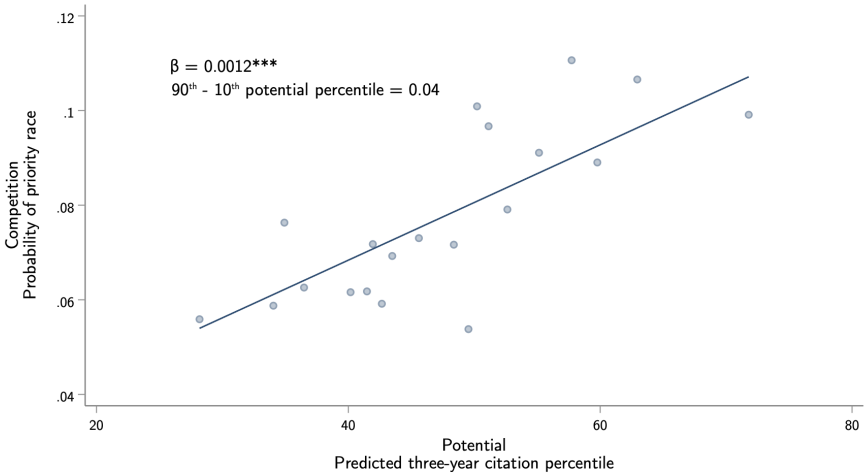
- Want to measure if multiple teams are working on the same project contemporaneously
- Use the priority race measure developed for the previous paper

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

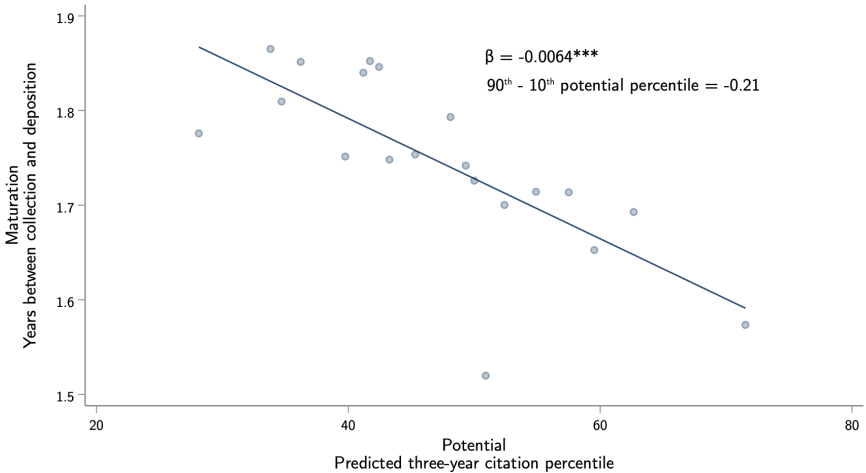


Proposition 1: high-potential projects are more competitive



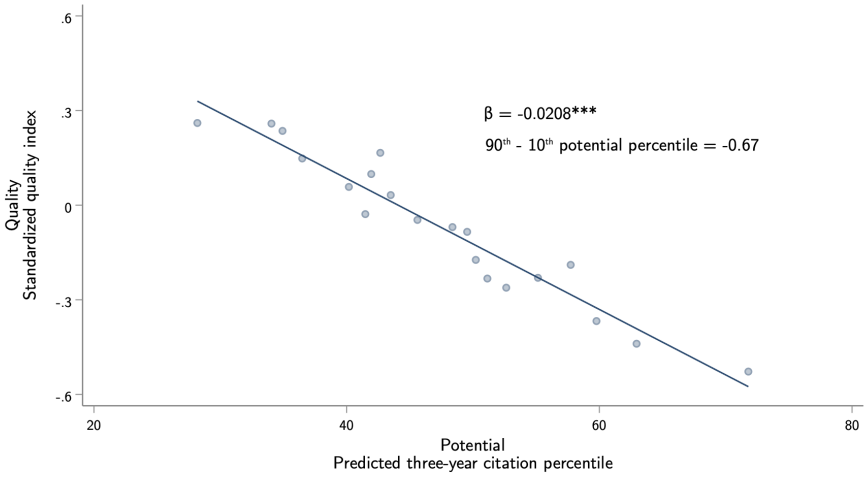
$$LogDepositsInCluster_{it} = \alpha + \beta PredictedCites_{it} + \tau_t + \epsilon_{it}$$

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



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

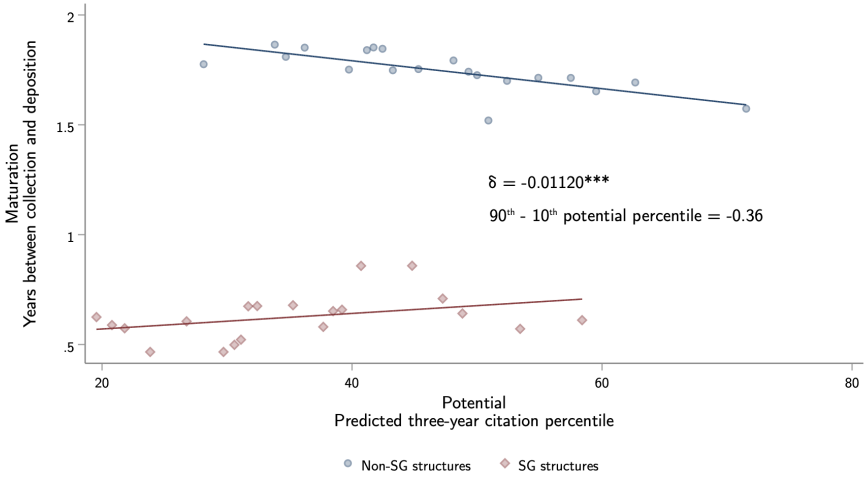
What about project complexity?

- **In general:** omitted variables bias might be a concern
- **In particular:** if high potential projects are also more complicated, this could drive our results. Lower quality is caused by the difficulty / complexity of the project, not rushing
- In the paper, we perform several analyses to rule out this story; today focus on just one

Structural genomics consortia

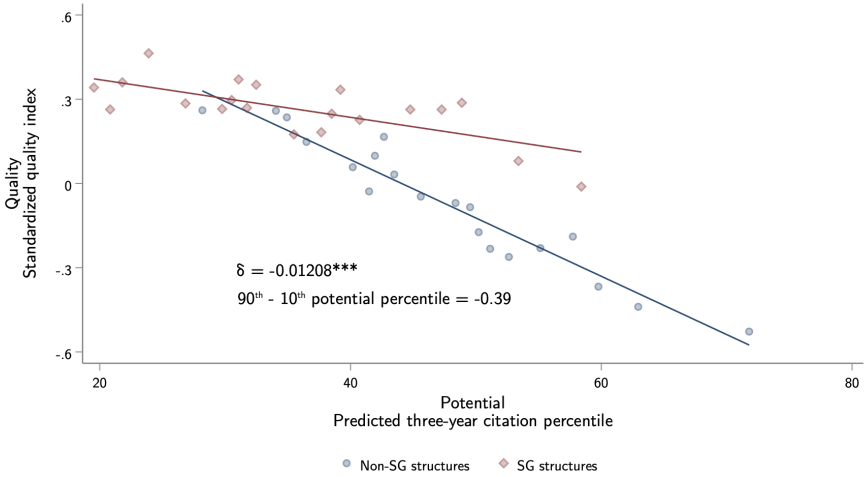
- **Key idea:** scientists affiliated with different types of institutions face different incentives
- Structural genomics consortia are publicly funded groups focused on achieving comprehensive coverage of the protein folding space
- Less focused on publishing and priority $\rightarrow$ competition 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



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

Conclusions and future work

- **Positive conclusion:** competition in science leads researchers to work faster and produce lower quality work
- **Normative analysis:** back-of-the envelope analysis suggests we have spent $2-5 billion on "cleaning up" these low-quality deposits. A social planner would prefer that we do them well the first time!
- However, taking a stand on optimal competition is hard. 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