Academic Bubbles*

Richard Lowery† UT Austin
Michael Sockin‡ UT Austin
Matteo Tranchero§ UPenn

August 2025

Abstract

We develop a model of how a principal motivates innovation by contracting researchers to generate ideas that improve economic productivity. Because individual contributions are unobservable, the principal relies on career incentives tied to peer recognition, measured through citations. However, citations reward not only societal value but also reflect the number of researchers working on a given topic. As a result, researchers may inefficiently over-coordinate their efforts. This can give rise to "academic bubbles" in which researchers continue working in areas with little prospect of advancing knowledge. We empirically analyze the selection of topics in research on the genetic determinants of human disease. We document that crowding on a narrow set of topics inflates citation impact, with patterns that point to career concerns as a driver of misallocated scientific effort.

*We thank Cecil-Francis Brenninkmeijer and Charlie Guthmann for research assistance. All errors are our own.
†Email: richard.lowery@mccombs.utexas.edu.
‡Email: michael.sockin@mccombs.utexas.edu.
§Email: mtranc@wharton.upenn.edu.

1 Introduction

Scientific progress crucially depends on the development of new ideas. Researchers are motivated by the prospect of receiving credit for their contributions, which is largely awarded to those who are first to make a discovery public (Merton, 1957; Hill and Stein, 2025b). This priority-based system converts ideas into reputation, which in turn shapes access to funding and career advancement (Tuckman and Leahey, 1975; Diamond Jr, 1986; Dasgupta and David, 1994; Abbott et al., 2010; Ellison, 2013). In practice, however, credit is assigned through citations, which reflect the judgment of peers who decide which ideas are worth building on. This system offers a practical solution to a difficult problem: assessing research quality across fields with different methods, audiences, and standards. After all, what better way to quantify the significance of an idea than by measuring its influence on other experts in the field, as captured by objective citation counts?

Beyond shaping individual careers, citations play a deeper role in the organization of science. In the absence of centralized coordination, academia relies on the choices of independent investigators, each pursuing their own research interests (Polanyi, 1962). Yet these choices collectively determine which problems are studied and which are ignored, with lasting consequences for public health and economic growth. Citations help steer this process by concentrating recognition and resources on topics seen as important or timely. In principle, this rewards researchers for pursuing the most socially valuable questions (Hill and Stein, 2025a). This logic supports a longstanding faith in scientific institutions as vehicles for maximizing public benefit (Bush, 1945; Strevens, 2003). It also informs how recent declines in research productivity are interpreted. If incentives are working as intended, then the observed slowdown in discovery must be proof that the easy ideas have been explored and what remains is harder to find (Bloom et al., 2020; Park et al., 2023).

However, there is growing concern that citations might distort the direction of scientific progress. In systems where career advancement depends heavily on citation counts (Hager et al., 2024), researchers face strong pressures to avoid exploratory topics that attract little recognition (Wang et al., 2017) and instead gravitate toward popular areas where citation potential is greatest (Packalen and Bhattacharya, 2017). As Bhattacharya and Packalen (2020) argue, the widespread reliance on citations in evaluation skews the balance toward safer, incremental work in established fields. The result is a coordination trap: researchers cluster around familiar problems not because they are the most promising, but because they offer the clearest path to professional reward. This dynamic raises the possibility that the observed slowdown in breakthrough innovation may stem less from ideas becoming harder to find than from incentives that steer scientists toward increasingly crowded

domains with diminishing returns.

In this paper, we explore this possibility by framing academic research as a contracting problem between a principal, society, and many agents, researchers. The principal seeks to incentivize each agent to exert costly effort to conduct research, but cannot directly observe either that effort or the importance of the topic the researcher chooses to pursue, at least not until the topic matures. By contrast, publications and citations are readily observable and, as a result, contractible. While basing rewards on these imperfect signals may be necessary to sustain scientific effort, doing so can distort incentives. Researchers with career concerns may rationally gravitate toward less promising but heavily studied topics in order to secure citations. To examine this problem, we develop a formal model and test its empirical implications.

In the model, a population of researchers can exert costly effort over time to discover new topics or to produce findings within an existing one. Each active topic is either significant, meaning it increases economic output, or insignificant, meaning it reduces it. A topic's true value is revealed only upon maturation, which requires continued research activity. At inception, each new topic comes with a noisy binary public signal indicating whether it is promising, that is, likely to be significant, or speculative, that is, unlikely to be so. Speculative topics remain unresolved until they mature, though an active topic may also be downgraded if a negative signal arrives before maturation. Both discovery and maturation follow Poisson processes, with arrival rates increasing in the number of researchers working on the topic.

Researchers in our model are motivated by both impact and career advancement. We allow them to care about the importance of their work through a share in economic output and a preference for "prestige," which rises with success on significant topics and falls with failure. At the same time, they respond to citations, the dominant measure of academic recognition. This creates coordination incentives that can distort effort. Researchers are drawn to crowded topics where citations are easier to earn and where their work is more likely to mature. As a result, exploration is insufficient not only because private risk aversion exceeds social risk aversion, but also because of strategic interaction. This dynamic leads to the over-researching of speculative topics and the neglect of new ones. It represents a novel form of moral hazard in teams (Holmstrom, 1982) and parallels the distortions highlighted in the multi-tasking framework of Holmstrom and Milgrom (1991).

Our model is fully tractable and delivers several key insights. First, when career concerns are mild, researchers focus only on promising topics and abandon speculative ones. In contrast, when

career concerns are strong, researchers continue to pursue speculative topics through to maturity. Working on an active topic—however uncertain—generates citations, while searching for new topics does not. As a result, strong career incentives reduce the informational value of citations and lower overall research productivity. Promising topics that initially attract attention continue to be researched even after negative signals arrive, making it common knowledge that they are unlikely to be significant. We refer to this persistence as an "academic bubble". Yet in the absence of career incentives, effort may be too low to sustain productive research. In that case, academic bubbles may be a necessary cost of motivating socially beneficial effort.

To test the model, we identify a setting, the study of the genetic determinants of human disease, which closely parallels its structure. The space of inquiry includes a large but finite set of genes, each of which may or may not contribute to a human disease. Researchers face a choice between working on genetic targets that are already under investigation or exploring new genes, mirroring the model's distinction between entering crowded topics or seeking out new, potentially significant ones. We find that publications on well-studied genes receive more citations, that crowded genetic areas continue to attract researchers even after accounting for ex-ante scientific importance, and that researchers with plausibly weaker career concerns (because they are likely already tenured or employed directly by the National Institutes of Health rather than by universities) are more likely to pursue novel genetic targets rather than join heavily studied ones. These findings support the model's predictions and suggest that scientific effort may be significantly misallocated in a domain of high societal relevance.

These theoretical and empirical findings offer a new perspective on the widely cited decline in idea generation documented by Bloom et al. (2020) and Park et al. (2023), among others. The problem may not be that ideas are growing scarce, but rather that researchers are increasingly disincentivized from searching for them. Instead of exploring new ground, the current system of scientific rewards encourages concentrating on already crowded topics. This interpretation of scientific stagnation offers a more hopeful outlook. If the slowdown in research dynamism stems from incentives rather than from fundamental limits on human ingenuity, then changing how we evaluate and reward scientific work could help restore the pace of discovery.

Our paper contributes to several distinct but connected literatures. Most broadly, it advances our understanding of the direction of scientific progress and the forces that distort it (Nelson, 1962). The distortions we examine do not rely on information externalities (Hoelzemann et al., 2025), economic spillovers (Akcigit et al., 2021), congestion costs (Hopenhayn and Squintani,

2021), racing dynamics (Bryan and Lemus, 2017), the higher costs of novel projects (Carnehl and Schneider, 2025), or differences in researchers' preferences for autonomy (Aghion et al., 2008). Instead, we contribute to a growing body of work showing that the institutional structure of science and how it is evaluated can generate misallocation of efforts (Budish et al., 2015; Hill and Stein, 2025a; Wang et al., 2017). To our knowledge, we are among the first to offer a general equilibrium characterization of how research unfolds when strong coordination incentives arise from researchers' career concerns and risk aversion.

Our paper also relates to the literature on agency problems in innovation. Holmstrom (1989) examines the internal conflict within firms where innovation competes with other operational demands. Manso (2011) studies optimal contracts between a principal and an innovating agent, showing that insurance against failure is essential to sustaining innovative effort. Aghion et al. (2013) similarly argue that institutional ownership can promote innovation by reducing career risk for managers. Like these frameworks, we emphasize the role of risk aversion and the value of insurance in supporting innovation. However, our focus is on a different mechanism: the general equilibrium effects of the implicit insurance provided by career incentives based on accolades from peers. In our setting, distortions in the direction of scientific discovery do not arise from coordination failures, but from deliberate coordination rooted in shared career concerns. Researchers benefit from concentrating on the same topics, even when the likelihood of success is low. This dynamic creates contract externalities embedded in the informal relationship between individual researchers and the academic community.

Finally, our work also connects to the literature on resource misallocation (Hsieh and Klenow, 2009). For example, Eisfeldt and Rampini (2006) show that capital misallocation is countercyclical, while reallocation tends to be procyclical. Glode and Lowery (2016) highlight how resources can be wasted on zero-sum transfers in the financial sector. Other work focuses on the misallocation of attention due to the cost of acquiring information (e.g., Sims, 2003). While these studies emphasize the inefficient allocation of production inputs or attention, we examine a different margin: the consequences of how effort in science and innovation is allocated (Acemoglu, 2023; Myers, 2020).

2 Model

Researchers engage in research in continuous time, with $t \geq 0$. They work on topics that ultimately prove to be either significant or insignificant. Research accelerates the process through which topics mature, meaning they are either revealed to be unproductive or begin to raise economic output by

increasing total factor productivity (TFP). This process generates projects that contribute to the maturation of the topic. These projects, which can be understood as academic papers, also attract citations and generate citations for others. Researchers have a claim on economic productivity, seek prestige from working on topics that prove significant, and receive compensation based on citations to their work. These objectives can align, but they may also be in conflict.

2.1 Fields and Topics

There are L fields in which $N \geq 2$ researchers can potentially conduct research. Not all fields are active at a given time, and we assume $L \gg N$ so that the set of fields is large relative to the number of researchers. At some time t_j^l, a researcher publishes a seminal project j in field l, which establishes a topic indexed by its inception time. This publication can be interpreted as an academic journal article. The arrival of a new topic is governed by a compound Poisson process dJ_t^l with compensator $\lambda e_{lt} dt$, where $e_{lt} = \sum_{n=1}^{N} e_{lt}^n$ and $e_{lt}^n \in \{0, 1\}$ denotes the effort that researcher n devotes to field l. Each field can host only one active topic at a time. If a researcher does not wish to work on the current topic, they must redirect their effort to another field. Allowing multiple topics per field or modeling a finite but large set of potential topics would not meaningfully alter the model's predictions. In the context of biological sciences, for example, each human gene can be viewed as a field. Within a gene, researchers may study how it relates to a particular disease, which corresponds to a topic. Some genes may attract active investigation, while others may be unexplored.

Once a topic in a field is developed, other researchers can begin working on it. Let $r_{lt}^n \in \{0, 1\}$ indicate whether researcher n is working on the topic in field l at time t. A researcher can either work on one existing topic or search for a new one, but not both at the same time. For all $t \geq t_j^l$, the total number of researchers working on the topic is $r_{lt} = \sum_{n=1}^{N} r_{lt}^n$. Each of them produces a project within the topic, and each project cites others in the same line of work. The topic matures at a random stopping time $\tau_j^l > t_j^l$, governed by a second Poisson jump process dB_t^l with compensator $\lambda r_{lt} dt$. The more researchers contribute, the faster the topic matures. This rate may exceed the rate at which new topics are discovered. In academic research, topic maturity corresponds to a publication that resolves or decisively critiques the central questions posed by the literature.

When a topic matures, it either results in a significant discovery that advances the field or an insignificant one that does not. This outcome is indexed by $\pi_{lj} \in \{0, 1\}$, where $\pi_{lj} = 1$ indicates that the topic is significant. The true value of π_{lj} is hidden and remains unknown to researchers until the

topic matures. At the time of inception, all researchers observe a binary public signal, $S_{lj} \in \{0, 1\}$, which is drawn randomly and provides imperfect information about whether the topic is likely to be significant. The signal satisfies $\Pr(S_{lj} = 1 \mid \pi_{lj} = 1) = q_1$ and $\Pr(S_{lj} = 1 \mid \pi_{lj} = 0) = q_0$, where $q_0 < q_1$. The prior probability that a new topic is significant is given by p. Using Bayes' Rule, the posterior belief that a topic is significant after observing the signal S_{lj} is:

$$\Pr(\pi_{lj} = 1 \mid S_{lj} = 1) = \frac{\Pr(S_{lj} = 1 \mid \pi_{lj} = 1) \Pr(\pi_{lj} = 1)}{\Pr(S_{lj} = 1 \mid \pi_{lj} = 1) \Pr(\pi_{lj} = 1) + \Pr(S_{lj} = 1 \mid \pi_{lj} = 0) \Pr(\pi_{lj} = 0)},$$
$$\Pr(\pi_{lj} = 1 \mid S_{lj} = 0) = \frac{\Pr(S_{lj} = 0 \mid \pi_{lj} = 1) \Pr(\pi_{lj} = 1)}{\Pr(S_{lj} = 0 \mid \pi_{lj} = 1) \Pr(\pi_{lj} = 1) + \Pr(S_{lj} = 0 \mid \pi_{lj} = 0) \Pr(\pi_{lj} = 0)},$$

from which it follows that:

$$\Pr(\pi_{lj} = 1 \mid S_{lj} = 1) = \frac{pq_1}{pq_1 + (1-p)\, q_0} = p_1,$$
$$\Pr(\pi_{lj} = 1 \mid S_{lj} = 0) = \frac{p\,(1-q_1)}{1 - pq_1 - (1-p)\, q_0} = p_0.$$

The consequences of a topic being significant or insignificant are discussed in the sequel. Note that the unconditional probabilities of each signal realization are given by $\Pr(S_{lj} = 1) = pq_1 + (1-p)q_0$ and $\Pr(S_{lj} = 0) = 1 - pq_1 - (1-p)q_0$.[1]

After a topic is discovered, researchers may receive additional information beyond the initial signal S_{lj} that updates their beliefs about the topic's significance. We model this by introducing a second signal that may arrive before the topic matures. This second signal arrives according to a Poisson process dL_t^l with compensator ξdt, and it updates researchers' beliefs about the topic. For simplicity, we assume this second signal is always a negative update and applies only to topics that initially received a positive signal. When this negative update arrives, it reduces the perceived probability that the topic is significant to p_0, equivalent to the belief researchers would hold had they observed a low initial signal, $S_{lj} = 0$. This second signal can improve economic efficiency by helping researchers reassess whether active topics are worth pursuing. At the same time, this parsimonious structure helps to highlight a key inefficiency introduced by career concerns, which we examine in detail below.

We can then define the following pseudo-probability process $p_{lt} \in \{0, p_0, p_1\}$ with the law of

[1] In principle, one could allow for persistence in topic success by assuming that the likelihood a new topic is significant is positively correlated with the significance of the last topic that matured in that field.

motion

$$dp_{lt} = \left(\Pr\left(\pi_{lj} = 1 \mid S_{lj}\right) - p_{lt-}\right)\left(dJ_t^l - \lambda e_{lt} dt\right) + \left(p_0 - p_{lt-}\right)\left(dL_t^l - \xi dt\right) + \left(0 - p_{lt-}\right)\left(dB_t^l - \lambda r_{lt} dt\right), \tag{1}$$

where $dB_t^l - \lambda r_{lt} dt$, $dL_t^l - \xi dt$, and $dJ_t^l - \lambda e_{lt} dt$ are three martingale jump processes. The first updates p_{lt} to the conditional probability that the topic is significant, $\Pr(\pi_{lj} = 1 \mid S_{lj})$, when a new topic is discovered. The second updates beliefs when a topic with an initially high signal receives a negative signal, reducing the perceived likelihood that the topic is significant. Throughout the remainder of the paper, we refer to topics that receive a high initial signal and no subsequent negative update as *promising*. The third process resets p_{lt} to zero when the current topic in field l matures. Because signal arrival rates increase with the number of researchers working on a topic, the model exhibits social learning and experimentation behavior similar to social bandit strategies.

The cumulative citations received by a researcher who works on the current topic in field l from time s to t' are given by $\int_s^{t'} (t - s) r_{lt}\, dt$. Accordingly, the expected citations from the project initiated at time t_j^n, denoted by V_{j,t_j}, are

$$V_{j,t_j} = \mathbb{E}\left[\int_{t_j^l}^{\tau_j^l} (\tau_j^n - t_j^n) r_{lt}\, dt \,\middle|\, \mathcal{F}_{t_j}^c\right] = \mathbb{E}\left[\int_{t_j^n}^{\infty} e^{-\lambda t} (t - t_j^n) r_{lt}\, dt \,\middle|\, \mathcal{F}_{t_j}^c\right], \tag{2}$$

where $\mathcal{F}_t^c$ is the filtration capturing the common knowledge available at time t. The second equality follows from the memoryless property of Poisson counting processes. Since at most one topic can be active in a field at any given time, we omit the topic index j from the researcher count r_{lt}.

2.2 Researchers

A researcher aims to complete projects over the course of her career and to earn citations. Each researcher is born at some time t with type $n \in \{1, \ldots, N\}$. Since only one researcher of each type exists at any time, we refer to her simply as researcher n. She retires at a random time T_i^n, determined by a Poisson shock process dQ_{nt} with compensator ηdt, and is immediately replaced by a new researcher of the same type.

At each instant, a researcher chooses which topic to work on, selecting only one topic at a time. Let $r_{lt}^n \in \{0, 1\}$ indicate whether the researcher is working in field l at time t. If she instead chooses to search for a new topic, such a topic may arrive according to the Poisson process described above. Let $e_{lt}^n \in \{0, 1\}$ indicate whether the researcher is exploring field l for a new topic at time t. A

researcher incurs a cost $\kappa > 0$ in certainty equivalent utility each instant she spends working on an active topic or exploring for a new one. The cost is the same in both cases, but a researcher searching for a new topic also bears the additional risk that the topic never arrives, and the payoff remains zero.

We assume that researcher n is risk-averse and has career concerns tied to his citation count, C_{nt}, and prestige or esteem, β_{nt}, which reflects promotions, awards, and broader recognition. Her compensation has two components: (1) a time-varying wage c_{nt}, which will later be linked to current output, and (2) a path-dependent high-powered incentive component, $\phi_X(\log \beta_{nt} + C_{nt})$, accumulated over his career. At each instant, the researcher also incurs an effort cost of $\kappa \sum_{l=1}^{L}(r_{lt}^n + e_{lt}^n)$ to engage in either active research or topic discovery. Let $\boldsymbol{r}_t^n = [r_{1t}^n, \ldots, r_{Lt}^n]$ be the vector of binary efforts the researcher devotes to active topics, and let $\boldsymbol{e}_t^n = [e_{1t}^n, \ldots, e_{Lt}^n]$ represent efforts toward discovering new topics, subject to the current availability of topics. Let ι_L denote the $L \times 1$ vector of ones. Following Holmstrom (1989), we assume the researcher has Constant Absolute Risk Aversion (CARA) flow preferences over compensation and effort:

$$u\left(c_{nt}, \beta_{nt}, C_{nt}, \boldsymbol{r}^n, \boldsymbol{e}^n\right) = -\beta_{nt}^{-\gamma\phi_X} e^{-\gamma(c_{nt}+\phi_X C_{nt})+\kappa(\boldsymbol{r}^n+\boldsymbol{e}^n)'\iota_L}, \tag{3}$$

where γ is the coefficient of absolute risk aversion over total compensation. The researcher has a subjective discount rate $\rho > 0$, and we assume $\rho + \eta > \lambda N$, meaning that discounting exceeds the maximum arrival rate of jump events.

If researcher n works on topic j in field l, then her citation count evolves according to the law of motion

$$dC_{nt} = \sum_l r_{lt} r_{lt}^n \, dt - C_{nt-} \, dQ_{nt}, \tag{4}$$

which reflects that researchers cite one another when working within the same topic. As a result, researchers prefer to work in fields that are already populated by others. If the topic matures and is significant (i.e., $\pi_{lj} = 1$), the researcher receives accolades such as promotions or awards. Her prestige β_{nt} evolves according to:

$$d\beta_{nt} = \beta_{nt-} \sum_l \pi_{lj} r_{lt-}^n \, dB_t^l + (1 - \beta_{nt-}) \, dQ_{nt}, \tag{5}$$

as an additional benefit. The citation count is initialized at zero and prestige begins at one when the researcher starts her career. We summarize the researcher's cumulative status as $X_{nt} =$

$\log \beta_{nt} + C_{nt}$, which evolves as:

$$dX_{nt} = \sum_l r_{lt} r^n_{lt}\, dt + \sum_{n'} \pi_{lj} r^n_{lt-}\, dB^l_t - X_{nt-}\, dQ_{nt}, \tag{6}$$

with initial condition $X_{n0} = 0$. Researcher n's preferences over compensation are then expressed as $-e^{-\gamma c_{nt} - \gamma \phi_X X_{nt}}$. Note that career rewards are not modeled as a fixed supply. For instance, if all researchers were to initiate new topics without contributing to active ones, the total pool of rewards would shrink, as citations would become scarce.

Finally, we assume that each researcher has an outside option upon entering the profession at time T^n_{i-1}. This alternative career requires no effort and, for simplicity and stationarity, provides a constant flow of consumption $c_{nT^n_{i-1}}$. The outside option delivers an expected value $U^0_{T^n_{i-1}}$ given by

$$U^0_{T^n_{i-1}} = -\mathbb{E}\left[\int_{T^n_{i-1}}^{T^n_i} e^{-\rho(t-T^n_{i-1})} e^{-\gamma c_{nT^n_{i-1}}}\, dt \,\middle|\, \mathcal{F}_{T^n_{i-1}}\right] = -\frac{e^{-\gamma c_{nT^n_{i-1}}}}{\rho+\eta}. \tag{7}$$

As a result, researcher n faces a participation constraint (PC):

$$U_{nT^n_{i-1}} \geq U^0_{T^n_{i-1}} \tag{8}$$

Since a researcher can be born at any instant, and the participation constraint is most binding when no active topic is available, we must verify that the constraint is satisfied in the absence of an active topic.

A researcher i of type n, born at time T^n_{i-1} (with $T^n_0 = 0$), solves the following optimization problem:

$$\begin{aligned} U_{nT^n_{i-1}} &= \sup_{\boldsymbol{r}^n, \boldsymbol{e}^n,\, (\boldsymbol{r}^n_t + \boldsymbol{e}^n_t)'\iota_L \leq 1\, \forall t} \mathbb{E}\left[\int_{T^n_{i-1}}^{T^n_i} e^{-\rho(t-T^n_{i-1})} \left(-e^{-\gamma c_{nt} - \gamma\phi_X X_{nt} + \kappa(\boldsymbol{r}^n_t + \boldsymbol{e}^n_t)'\iota_L}\right) dt \,\middle|\, \mathcal{F}_{T^n_{i-1}}\right] \\ &= \sup_{\boldsymbol{r}^n, \boldsymbol{e}^n,\, (\boldsymbol{r}^n_t + \boldsymbol{e}^n_t)'\iota_L \leq 1\, \forall t} \mathbb{E}\left[\int_{T^n_{i-1}}^{\infty} e^{-(\rho+\eta)(t-T^n_{i-1})} \left(-e^{-\gamma c_{nt} - \gamma\phi_X X_{nt} + \kappa(\boldsymbol{r}^n_t + \boldsymbol{e}^n_t)'\iota_L}\right) dt \,\middle|\, \mathcal{F}_{T^n_{i-1}}\right], \end{aligned} \tag{9}$$

subject to the participation constraint (8), using the memoryless property of Poisson processes. Each researcher fully internalizes how her own actions and those of others shape her career trajectory. Researchers also understand how firms generate output and, therefore, internalize how their work contributes to production. This may reflect private returns, such as royalties or institutional rewards, or may be interpreted as a direct preference for producing socially valuable research. The latter interpretation is particularly relevant when the value of research lies in public goods provision or improved policy. The special case where $\phi_X = 0$ corresponds to a researcher who cares only about

her impact on economic output.

2.3 Growth

While citations and prestige represent the private returns to research, such efforts also promote economic growth, which can benefit the broader economy. This connects our model to the endogenous growth literature. Alternatively, one may interpret output as the profits of the firms where researchers are employed. We capture this by letting aggregate productivity, A_t, evolve according to the law of motion:[2]

$$dA_t = \sum_{l=1}^{L} (2\pi_{lj} - 1)\, dB_t^l. \tag{10}$$

This expression implies that aggregate productivity rises when a significant topic matures and falls when an insignificant one does. The decline in productivity reflects that firms rely on research to guide product development and strategic decisions. When a topic proves unimportant, prior investments based on it are revealed to have been wasted. This can be interpreted as a correction following speculative misallocation. Although there is already an opportunity cost to pursuing failed research, allowing aggregate productivity to fall in such cases makes this cost explicit without complicating the model. Since the key mechanisms remain unchanged, modeling this drop adds realism without affecting the qualitative results.

The representative firm has an output given by: $Y_t = A_t K$, where capital $K > 0$ is fixed in the economy. As discussed in Section 2.2, each researcher receives a share of total output as part of their compensation, given by $c_{nt} = \phi_A Y_t$, where $\phi_A < \frac{1}{N}$. We assume that the initial output Y_0 is large enough that the principal can offer high-powered career incentives without violating feasibility. In particular, total compensation to researchers, given by $N\phi_A Y_t + \phi_X \sum_{n=1}^{N} X_{nt}$, must not exceed Y_t.

In what follows, we assume that the conditional probability that a topic is significant given a low signal, p_0, is sufficiently small. Specifically, we require $p_0 < \frac{1}{2}$. By l'Hôpital's rule, this implies $p_0 < \frac{1-e^{-\gamma\phi_A K}}{1-e^{-2\gamma\phi_A K}}$ for any $\phi_A > 0$, since the function $\frac{1-e^{-x}}{1-e^{-2x}}$ is increasing in x. In contrast, we assume that $p_1 \geq \frac{1-e^{-\frac{\gamma}{N}K}}{1-e^{-2\frac{\gamma}{N}K}}$.

[2]We could allow the productivity jump to depend on the number of researchers on a given topic, r_{lt}. However, this would bias researchers toward coordinating on the same topic and amplify the social cost of doing so when the topic is insignificant. To avoid this, we assume the jump in productivity is independent of how many researchers are working on the maturing topic.

2.4 Relationship Among Drivers of Compensation

Prestige and economic output play similar roles in motivating researchers, as both increase only when a significant topic matures. The key difference lies in how they relate to the value of a research project. Economic output is directly tied to the magnitude of a project's contribution, whereas prestige can be substantial even when the economic value of the research is modest. While this distinction is not central to the baseline analysis, the framework allows successful outcomes to influence payoffs flexibly.

In the main analysis, we focus on citations. Unlike prestige or economic output, citations are not necessarily associated with doing useful research, but they are observed continuously and serve as a primary performance measure. We place risk aversion on the citation-based component of pay to capture the asymmetry between a failed career and academic survival, as compared to the difference between survival and academic stardom. This reflects an aversion to downside risk in career outcomes.

2.5 Equilibrium Definition

We consider a Markov Perfect Equilibrium in which each researcher fully internalizes the impact of her actions on her own career, the careers of others, and the aggregate economy. A strategy for a researcher of type n is a sequence of project efforts $\{\vec{e}_t^n, \vec{r}_t^n\}_{t=T_{i-1}^n}^{T_i^n}$ that solves the dynamic program in Equation 9. An equilibrium is a fixed point in which researchers' conjectured project intensities $\{\hat{e}_{lt}, \hat{r}_{lt}\}_{l,t}$ match actual intensities $\{e_{lt}, r_{lt}\}_{l,t}$—that is, all researchers have rational expectations about the evolution of the game.

This setup defines a mean-field Nash Equilibrium, where the relevant "mean fields" are the number of researchers in each field, $\{e_{lt}, r_{lt}\}_{l,t}$. A researcher's problem is Markov in $L + N + 1$ state variables: the L probabilities p_{lt} that each topic is significant, which also indicate whether a field is currently active; the N researcher statuses X_{nt}; and aggregate productivity A_t. However, as we will show, the optimal closed-loop policies depend on a reduced state space. This allows strategies to be not only Markov, but also recursive. We initialize the economy at $t = 0$ with no active topics.

3 Equilibrium

In this section, we construct the equilibrium of the research game. Our central proposition concerns the behavior of the N researchers. We apply the dynamic programming principle to characterize each researcher's optimal policy. Given CARA preferences and the linear evolution of status, the

absolute level of a researcher's status does not influence her decision. As a result, all researchers follow identical policies in equilibrium. A researcher's choice depends only on whether there is an active topic and, if so, on the signal S_{lj} it received about its significance. We refer to an active topic with a high signal ($S_{lj} = 1$) as *promising*, given its high likelihood of being significant. An active topic with a low signal ($S_{lj} = 0$) is referred to as *speculative*, due to its lower likelihood of producing a valuable outcome. We then have the following key proposition:

Proposition 1. *The optimal research policy for the N researchers has the following properties:*

- *If there is a promising active topic, then all researchers work on that topic until either it matures or negative news arrives.*
- *If no promising topic is active and career concerns are sufficiently strong (i.e., $\phi_X \geq \phi_X^*$), then all researchers work on any available speculative topic until it matures.*
- *If no topic is active, then all researchers search for a new topic and are indifferent about which field to search in.*
- *This is an equilibrium if the effort cost κ is sufficiently small. Specifically, if speculative topics are researched, then $\kappa \leq \min\{\kappa_1^*, \kappa_1^{pc}\}$. If speculative topics are not researched, then $\kappa \leq \min\{\kappa_2^*, \kappa_2^{pc}\}$, where κ_1^*, κ_1^{pc}, κ_2^*, and κ_2^{pc} are defined in equations* (A.33)*,* (A.36)*,* (A.32)*, and* (A.38)*, respectively.*

In contrast, in the absence of career concerns, researchers abandon speculative topics and instead attempt to discover a new topic whenever no promising topic is active.

Proposition 1 shows that the equilibrium in our mean-field research game is surprisingly simple. All researchers concentrate on any active promising topic that has received a high signal ($S_{lj} = 1$), continuing until it matures. These topics are most likely to be significant, which increases both aggregate output and individual prestige. Coordinating on the same topic also maximizes citation accumulation. As a result, there is at most one active high-signal topic at any point in time.

The more interesting case arises when a topic has received a low signal ($S_{lj} = 0$), or when negative news arrives about a topic that was initially promising. Without career concerns, researchers would abandon such speculative topics and focus instead on discovering new, more promising ones. Since the conditional probability that a speculative topic is significant, p_0, is small, the socially optimal action is to stop working on it and continue searching. In the presence of career concerns, however, researchers may continue working on speculative topics through to maturity. This behavior is career-enhancing for two reasons. First, speculative topics generate citations, while searching for new topics does not. Second, speculative topics yield prestige if they succeed, but do not reduce

status if they fail. As a result, researchers may continue to invest in topics that are no longer believed to be valuable from a societal perspective. This leads to the emergence of academic bubbles—topics that persist because they serve individual careers even when they no longer serve scientific progress.

At the same time, rewarding citations and prestige helps relax the researcher's participation constraint (8), since it improves expected career payoffs. This is especially important when effort costs κ and risk aversion γ are high, or when the likelihood of discovering a significant topic (p_0 and p_1) is low. Risk aversion over citations and prestige dampens the incentive to explore, similar to the mechanism in Jones (1995), where private risk aversion exceeds social risk tolerance, discouraging innovation. The lower the likelihood that research yields meaningful results, the harder it becomes to sustain research effort. We summarize these insights in the following proposition:

Proposition 2. *An increase in the reward for career advancement, ϕ_X, or in the probability that topics are significant, p_0 and p_1, relaxes a researcher's participation constraint. In contrast, higher effort cost κ and higher risk aversion γ (for γ sufficiently large) tighten the constraint.*

Our analysis has several important implications. First, citation counts in academic literature may be difficult to interpret when they reflect career concerns. Researchers may prefer to work on topics that are unlikely to be significant, simply because such topics generate citations more reliably than riskier new ones. As a result, citations need not indicate social impact or, if economic growth is the relevant outcome, long-term value. Moreover, career incentives may shape the behavior of academic journals. Journals that internalize the citation motive may prefer to publish articles from speculative fields, knowing these will attract attention regardless of the underlying social value. This can shift editorial priorities away from research that is most promising from a societal perspective.

Second, when career concerns are strong, researchers may allocate excessive effort to speculative topics that eventually fade out, resulting in a clear form of misallocation. This behavior can distort the direction of scientific progress and offers an alternative explanation for the observed decline in research productivity (Bloom et al., 2020). However, our model also implies that disciplines with more informative signals about topic quality—those with high p_1 and low p_0, such as the natural sciences—are less vulnerable to this dynamic. In such fields, career incentives must be especially strong to induce inefficient crowding in speculative areas.[3]

[3]We show in the Online Appendix that this result holds in a dynamic setting where signals arrive gradually. In that case, there exists a threshold probability below which a topic is abandoned, and this threshold decreases as career concerns intensify.

Third, the model generates testable predictions about heterogeneity in researcher behavior. Researchers with weaker career concerns should crowd less into popular topics, since they face less pressure to accumulate citations.[4] For example, we expect to observe a drop in crowding after tenure, and among researchers who operate in environments where citations are not the primary measure of success. Moreover, if we can measure the potential significance of a topic, we should find that less career-driven researchers are overrepresented in high-potential but less popular areas. However, because most researchers face strong career incentives, the correlation between objective significance and publication output may be weak, or even undetectable. In the next section, we show that research on the genetic determinants of disease offers such a setting, with measurable topic significance and observable variation in career incentives across researchers.

Finally, our results raise a broader question about incentive design. If rewarding all contributions to a topic with citations encourages coordination and possible crowding, might selective rewards, such as prizes or awards for the most impactful work, foster more competition? If only the most influential paper receives recognition, risk-averse researchers may be more likely to explore less crowded topics or try to create new ones. This winner-take-all approach, as described by Dasgupta and David (1994), may be more aligned with social goals when diversification is desired. In contexts where society is risk-averse over innovation outcomes, prize-based incentives may be preferable. When financial or institutional resources are limited, selective recognition can provide an alternative to sustaining broad, citation-based incentives.

4 Career Concerns as a Contracting Outcome

In this section, we microfound the implicit academic contract that rewards citations and publications as the solution to an underlying agency problem. We extend our model by assuming that effort is not only costly for researchers, but also unobservable and non-contractible from the principal's (i.e., society's) perspective. This assumption is realistic: discovering new facts or ideas is inherently difficult, and much of the research process is not visible in the resulting publication. Moreover, the principal does not observe any signals about a topic until it matures. Because researchers contribute indirectly to shared knowledge and rely on one another to advance topics, the research environment resembles a moral hazard in teams problem from a social perspective (Holmstrom, 1982).

[4]The model can be extended to include researchers with different levels of career concern. The qualitative implications remain similar, so we focus on the homogeneous case for clarity.

We express the flow utility of researcher n as

$$u\left(w_{nt}, \vec{e}_t^{\,n}, \vec{r}_t^{\,n}\right) = -e^{-\gamma w_{nt} + \kappa(\vec{e}_t^{\,n} + \vec{r}_t^{\,n})' \iota_J},$$

where w_{nt} denotes the researcher's compensation paid by the social planner, the vectors $\vec{r}_t^{\,n}$ and $\vec{e}_t^{\,n}$ represent the effort that researcher n allocates to active topics and to the search for new topics, respectively, and the vector ι_L is an $L \times 1$ vector of ones. The researcher's effort is subject to the budget constraint $(\vec{e}_t^{\,n} + \vec{r}_t^{\,n})' \iota_L \leq 1$.

The first-best outcome would have the planner allocate all output across researchers at each instant. The maximum compensation the planner can provide to any researcher is $w_{nt} = \frac{1}{N} Y_{nt}$, and by ex-ante symmetry, it will choose to give each researcher exactly this amount. In addition, the planner would assign all m^* active researchers to work exclusively on promising active topics. The number of active researchers is limited by the flow certainty equivalent cost of effort, κ. This cost enters through the conditions for Pareto optimality: each active researcher must prefer following the planner's assignment over deviating. As a result, only $m^* - 1$ researchers may be incentivized to remain active. We summarize these implications in the following proposition:

Proposition 3. *In the first-best economy, which is Pareto efficient, it is socially optimal for there to be $m^* \leq N$ active researchers, and:*

- *if there is a promising active topic, all active researchers work on it until it matures;*
- *if no promising topic is active, all researchers search for a new topic and are indifferent between searching in the same or different fields;*
- *the number of active researchers m^* is the largest $m \leq N$ that satisfies condition* (A.54)*. If $\kappa \leq \kappa^*$, where κ^* is defined in equation* (A.55)*, then all N researchers are active;*
- *this is an equilibrium if $\kappa \leq \kappa^{pc}$, where κ^{pc} is given by equation* (A.57)*.*

Proposition 3 also highlights a key coordination failure that arises from a free-rider problem. A researcher's incentive to exert effort depends heavily on how many others are actively researching, and these incentives are weaker when the number of active researchers is small. Since effort is unobservable, researchers have an incentive to shirk, as the failure to discover a new topic or to mature an existing one may be attributed to bad luck rather than insufficient effort. Under asymmetric information, the first-best allocation may no longer be attainable, which motivates the need for incentive-compatible contracts.

In what follows, we assume $\kappa = \kappa^*$, as defined in Proposition 3, so that it is just optimal for all N researchers to be active in the first-best economy. We now illustrate the central free-rider problem that arises when effort is unobservable. If N is not too large, and all other researchers choose not to exert effort, then the remaining researcher also finds it optimal to deviate. This is because the benefits of research are shared, while the cost of effort is borne privately. Proposition 4 shows that, under these conditions, zero active researchers can constitute a Nash Equilibrium.

Proposition 4. *Suppose $\kappa = \kappa^*$. If $N < \tilde{N}$, where $\tilde{N}$ satisfies Equation A.59 with equality, then zero active researchers constitutes a Nash Equilibrium. This equilibrium is Pareto inferior to the first-best allocation characterized in Proposition 3.*

The possibility of coordination failure creates a role for the principal to offer an incentive-compatible contract that supports a second-best outcome. Since effort is unobservable, contracts must be based on observable measures such as output and citations. Because output is observable, the principal can also determine when a research topic proves successful. Therefore, contracts can either reward researchers continuously based on citation accumulation or only upon the successful maturation of a significant topic. The latter approach requires saving output to fund delayed rewards, which exposes researchers to considerable risk, or violating budget balance by paying more than current production allows. These constraints limit the feasibility of high-powered incentives and may induce distortions that prevent the first-best from being achieved.

This friction opens the possibility that a contract generating academic bubbles may be second-best, even if it diverts researchers from more socially valuable lines of inquiry. The standard trade-off in contract theory applies here: paying researchers only when output rises offers strong incentives to pursue promising topics but requires large transfers to risk-averse agents. When researchers are highly risk-averse and effort is costly, and when output changes are infrequent or difficult to verify, contracts that reward citations—and thus create career concerns—can be optimal despite their inefficiencies.

We conclude this section by characterizing the optimal contract, that is, the choice of ϕ_A and ϕ_X, that a principal (society) would offer. We assume the principal is risk-neutral and can commit at time 0 to a compensation scheme that maximizes the expected present value of output net of payments to researchers. The principal has a subjective discount rate ρ and chooses contract loadings subject to the researchers' participation and incentive compatibility constraints. Formally, the principal

solves:

$$U_P = \sup_{\phi_A, \phi_X} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(Y_t - \sum_{n=1}^{N} \phi_{nt}\right) dt\right], \tag{11}$$

subject to: researcher problem in Equation (9).

At time 0, we assume that no researcher has accumulated status and that no research topic is active.

Proposition 5 characterizes the key features of the optimal contract. In addition to the participation constraint binding, the principal chooses the contract loadings to balance the marginal cost of higher compensation with the marginal benefit of relaxing researchers' participation constraints.

Proposition 5. *The principal chooses the research contract loadings such that:*

- *If the researcher's effort cost κ is sufficiently small, the principal sets ϕ_A and ϕ_X to satisfy equations 8 and A.65, and researchers work only on promising topics.*
- *If κ is sufficiently large, the principal sets ϕ_A and ϕ_X to satisfy equations 8 and A.66, and researchers work on any active topic.*
- *If there is a risk of coordination failure, the principal must choose $\phi_X \geq \underline{\phi}_X$ to eliminate this equilibrium. If κ is high, then $\underline{\phi}_X$ must be set to ensure that researchers work on any active topic.*

Proposition 5 highlights that the optimal research contract depends on the cost of effort. When effort is relatively cheap, the principal can implement a socially efficient outcome by keeping ϕ_X low enough to discourage work on speculative topics. However, when effort is costly, the principal must reward all research to maintain participation. To prevent coordination failures, the planner may need to set the loading on status ϕ_X high enough to make even speculative research individually optimal.

5 Empirical Analysis

5.1 Research on the Genetic Roots of Human Diseases

In this section, we provide an empirical analysis to test the key implications of our model. This requires identifying a setting in which the set of potential research topics is observable ex ante, which makes it possible to assess whether researchers are crowding into a narrow subset of topics. Biomedical research on the relationship between genetic mutations and human diseases satisfies this condition. The space of possible topics, namely the set of human protein-coding genes, is large

but well defined, with over 19,000 candidate genes available for study. Moreover, the social value of successful research in this area is well established. Genes that harbor disease-causing mutations often serve as effective drug targets, increasing the likelihood of developing successful therapies (Nelson et al., 2015).

This setting aligns closely with our theoretical framework for several reasons. First, scientists face a clear choice between continuing to investigate well-studied genetic targets or exploring novel ones. Some regions of the genetic landscape are densely studied, raising the risk of inefficient crowding on a narrow set of topics (Edwards et al., 2011; Gates et al., 2021). Although incentives to establish priority in new areas exist (Hill and Stein, 2025a), researchers continue to focus on a relatively small subset of human genes. This is puzzling given widespread recognition that promising drug targets may lie among less-studied genes (Stoeger et al., 2018). Second, this context offers quantifiable ex ante measures of a gene's importance. Genes with mutations associated with disease are more likely to play a role in human pathology (Haynes et al., 2018; Richardson et al., 2024). This feature allows us to test whether researchers cluster in crowded areas because those genes are truly more ex ante promising, or whether they coordinate on less promising ones, reflecting misallocation and the potential emergence of academic bubbles.

Our model implies several empirical regularities about how researchers select topics in this setting. First, if crowding confers a citation advantage, we should observe that the number of citations a paper receives depends not only on the scientific importance of the gene it studies but also on how many other papers are being written about that gene. Second, we should see a causal relationship: exogenous increases in the number of papers on a gene should lead to more citations for work on that gene. More broadly, we expect a misalignment between the ex ante promise of a gene—such as its predicted relevance to disease—and the volume of research it attracts. Finally, the model predicts cross-sectional differences tied to career incentives. Researchers under stronger career pressures should be more likely to cluster in crowded areas, while those less constrained should explore less-studied genes. We measure variation in these incentives by comparing researchers at different career stages (tenured versus untenured), facing different career incentives (universities versus NIH's Intramural Research Program), or working in institutions where research output carries different weight in career advancement (more or less research-oriented universities).

5.2 Data

To test these ideas, we assembled a dataset of all papers focusing on human protein coding genes. Publications on protein-coding genes are identified using NIH's PubTator3 (Wei et al., 2024), which applies natural language processing to extract gene mentions from all articles indexed by PubMed. For ease of analysis and interpretation, we focus in our main analysis on publications studying only one gene (but results are all robust to examining multiple genes, see Appendix Table D2). This leaves us with 857,025 papers published between 1980 and 2018. Citation data come from NIH's iCite database,[5] which also tracks citations from clinical studies. The count of citations from USPTO patents is drawn from the data compiled by Marx and Fuegi (2020). Information on the principal investigator (PI), generally listed last in the authorship order, is obtained from Author-ity 2018 (Torvik and Smalheiser, 2021), which provides high-quality author disambiguation for all PubMed articles. Finally, institutional affiliations for each publication are drawn from OpenAlex (Priem et al., 2022).

To measure the ex ante scientific importance of a gene, we follow prior biomedical literature (Haynes et al., 2018) and use the number of distinct diseases with which it has been associated in genome-wide association studies (GWAS). GWAS are large-scale, atheoretical case–control studies that compare the DNA of individuals with and without a given condition to identify mutations statistically correlated with the disease (Tranchero, 2025). By design, GWAS do not restrict attention to a predefined set of genes, ensuring that results are not biased by the historical allocation of research effort (Visscher et al., 2012). Genes harboring mutations linked to many conditions are more likely to play central roles in human pathology and therefore represent more promising research targets (Haynes et al., 2018; Stoeger et al., 2018). We obtain these data from the University of New Mexico's *Target Illumination GWAS Analytics* (TIGA) platform (Yang et al., 2021), which aggregates evidence across studies to generate counts of gene–disease associations.[6] This measure serves as our baseline proxy for scientific importance and is independent of historical publication patterns. In Appendix C, we replicate all results using an alternative proxy based on molecular data capturing the probability that a gene is expressed in the presence of a disease.

Table 1 reports descriptive statistics. Papers in our sample receive, on average, 2.85 academic citations per year, though the distribution is highly skewed, with a right tail extending to more than 700 citations annually. The average publication studies a gene in the top eight percent of the most

[5] `https://icite.od.nih.gov/analysis`

[6] Data are publicly available at the following website: `https://unmtid-shinyapps.net/shiny/tiga/`.

studied ones. Roughly 44 percent of publications focus on understudied genes, defined as those below the median in cumulative prior publications. Genes are linked to an average of 40 disease traits in GWAS, but the distribution is wide, ranging from none to nearly one thousand. Most genes have a mouse homolog, allowing them to be studied in laboratory mice. We will use whether a gene has a mouse homolog as an exogenous source of variation in crowding by researchers into genes in our instrumental variable approach. Senior PIs account for 90% of our sample, while NIH intramural scientists authored only 1.5% of the publications. Finally, papers receive, on average, 0.09 clinical citations and 0.01 patent citations per year, highlighting the relative selectivity of translational uptake compared to academic impact.

6 Results

6.1 Crowdedness and Citations in Genetic Research

We begin by documenting that crowdedness is a salient feature of research in the genetic space, as captured in our data. We rank genes by the total number of publications over the sample period and plot the number of publications by rank in Panel (a) of Figure 1. The distribution is strikingly skewed, with a handful of genes attracting the vast majority of research while most receive little or none. Is this concentration efficient, with effort directed toward the most scientifically promising genes? To assess this, we examine how many conditions have been associated with each gene in GWAS studies. Panel (b) plots the same ranking against the logarithm of the number of diseases related to each gene. Although publication counts and GWAS associations are positively correlated (Appendix Figure D1), many promising genes are entirely neglected, while others with limited therapeutic potential continue to draw substantial attention. This misalignment echoes prior work showing that numerous scientifically important genes remain overlooked (Haynes et al., 2018; Stoeger et al., 2018; Richardson et al., 2024).

Overall, we find a substantial discrepancy between the volume of publications on a gene and its scientific importance. This suggests a potential misalignment between the citations a paper receives and the underlying scientific value of its contribution. To examine this possibility, we estimate whether papers on under-explored genes receive fewer citations even controlling for the number of diseases related to them. Our main specification is:

$$y_{i,t} = \alpha + \beta \cdot \text{crowdedness}_i + \theta \cdot \text{importance}_i + \mu_{j,t} + d_c + a_p + \varepsilon_{i,t}, \tag{12}$$

where the dependent variable $y_{i,t}$ is the number of citations received per year by paper i. The key

independent variable is crowdedness$_i$, defined as the percentile of the cumulative publication count for the focal gene, capturing how extensively it has been studied. importance$_i$ is our time-invariant proxy for the scientific importance of gene i. The specification includes journal-by-year fixed effects $(\mu_{j,t})$, MeSH disease-class fixed effects (d_c), and PI fixed effects (a_p). In the most stringent version, papers are compared within the same author, disease class, and journal-by-year cell.

Table 2 presents the results. We invert the gene rank so that higher values correspond to less studied genes and find a statistically significant decline in citation impact for publications on these genes. In the most stringent specification with PI fixed effects, studying the least studied gene (top percentile) rather than the most studied gene (bottom percentile) is associated with 0.9 fewer citations per year, relative to a mean of 2.85. Strikingly, this effect persists even after controlling for scientific importance, consistent with our key theoretical prediction. Although scientific importance is directionally associated with higher citations, the coefficient is imprecisely estimated. In Panel B, we re-estimate the specification using a dummy indicating whether the gene had received a below-median number of publications at the time of publication and find similar results: papers on below-median genes receive 0.154 fewer citations per year. Figure 2 visualizes these patterns by plotting yearly citations against the number of disease associations for each gene. While the relationship between scientific importance and citations is weakly positive, there is a sharp discontinuity favoring crowded genes at every level of importance, revealing a consistent citation penalty for studying understudied genes.

Although the baseline specification includes tight controls, we also address the risk of unobservable confounders with an instrumental variable strategy. The approach exploits variation in genetic similarity between humans and mice. Studying a human gene is easier and less costly when scientists can investigate its counterpart in laboratory mice. This is possible when the two species share a homolog gene, defined as a gene inherited from a common ancestor that performs equivalent functions in both species. About 10 percent of human protein-coding genes lack such a counterpart for evolutionary reasons. These genes have historically been neglected for reasons of convenience rather than importance (Stoeger et al., 2018; Richardson et al., 2024), making the absence of a mouse homolog a plausibly exogenous source of variation in crowdedness. Table 3 reports the results. As expected, genes with a mouse homolog are studied more frequently, with a strong first-stage F-statistic. Even in our most stringent specification with PI fixed effects, we continue to find evidence of a significant citation penalty. The instrumented estimates confirm a robust citation penalty, showing that the effect is driven by crowdedness itself rather than by unobserved

characteristics of the gene.[7]

6.2 Heterogeneity in Researchers' Career Incentives

While the aggregate results suggest that scientists may concentrate on crowded genetic topics to avoid a citation penalty, this alone does not establish that the penalty drives their choices. Scientists might instead be intrinsically motivated to study important genes regardless of their impact on their citation counts. To disentangle these explanations, we conduct a series of descriptive heterogeneity analyses that test whether researchers facing stronger career pressures, or greater incentives to maximize citation impact, are less likely to work on uncrowded genes. These analyses provide a suggestive way to explore whether the citation penalty is the mechanism underlying our results.

A first source of heterogeneity is whether the principal investigator (PI) is likely tenured. Tenure reduces career risk by providing long-term job security, which in turn lowers researchers' sensitivity to short-term citation incentives and allows them to pursue less common directions (Manso, 2011; Tripodi et al., 2025). We proxy for tenure by coding PIs as senior if they have been publishing for at least seven years, but results are similar using different cut-offs (Appendix Table D1). Consistent with this mechanism, Panel A of Table 4 shows that senior researchers are systematically more likely to publish on less studied genes, both by moving up the inverse percentile ranking of crowdedness (columns 1 and 2) and by focusing more often on genes with a below-median number of publications (columns 3 and 4). This relationship is visualized in the first panel of Figure 3, which plots the likelihood that a paper is authored by a senior PI against the crowdedness of the focal gene. Taken together, these results suggest that by alleviating career pressures, tenure reduces sensitivity to citation penalties and encourages researchers to explore less studied genes.

The second source of heterogeneity we examine is whether the paper's PI is affiliated with the NIH Intramural Research Program. Unlike academic researchers, intramural scientists are employed and funded directly by the NIH (Azoulay et al., 2013). In terms of our model, they are directly monitored by the NIH and do not face the up-or-out pressures of academia, which often push researchers toward short-term publication impact. Consistent with this prediction, we find that NIH-affiliated PIs are significantly more likely to publish on less-studied genes (Panel B of Table 4). The second panel of Figure 3 shows the positive relation between the likelihood of a paper

[7]This result is consistent with earlier evidence by Pfeiffer and Hoffmann (2007), who showed that genes studied more frequently are more likely to appear in higher–impact journals. Our analysis extends this result by demonstrating that the effect persists even after accounting for journal and author fixed effects, and further by using an instrumental variable strategy that isolates exogenous variation in how much a gene has been studied. Together, these results strengthen the case of a citation advantage derived from working on crowded topics.

having an NIH-affiliated author and the featuring of less studied genes. These findings complement prior evidence that intramural researchers generate more novel scientific output (Xu et al., 2025) and suggest that scientists not facing the incentives of academic careers are more willing to explore less studied parts of the genetic landscape.

Finally, we examine the role of university prestige. Researchers at top institutions are often embedded in larger laboratories that depend on a steady flow of publications and grants to sustain personnel (Zhang et al., 2022). Such labs tend to favor established lines of work over disruptive exploration, since their scale creates stronger incentives to avoid failure (Wu et al., 2019). In addition, novel contributions are more likely to appear in lower-impact journals, making them less attractive under tenure systems in elite departments (Wang et al., 2017). To test whether these dynamics shape topic selection, we restrict our sample to papers with U.S.-based PIs and compare those affiliated with top 10 universities to others. The results, shown in Panel C of Table 4, provide some evidence that researchers outside the most highly ranked departments are more likely to study less common genes. Although the within-author variation across rankings is limited, the binscatter in Figure 3 shows a suggestive association.

Taken together, the heterogeneity analyses suggest that career pressures play a central role in driving researchers toward crowded genes. PIs who are tenured or employed within the NIH intramural program, and thus less exposed to citation-based incentives, are significantly more likely to publish on less-studied genes. By contrast, researchers at elite universities, where career and resource pressures are most intense, show a tendency to remain in crowded areas. The consistent pattern across these comparisons attunes with the model's predictions that citation concerns discourage exploration, and help explain why attention clusters on genes that are already heavily studied rather than those that are more likely to be scientifically important.

6.3 Impacts on Downstream Innovation

We next examine how the allocation of scientific attention shapes downstream developments in medicine. We focus on two paper-level outcomes that capture potential real-world impact rather than academic recognition. The first is the number of citations a paper receives from clinical studies, a crucial step in translating discoveries into therapeutic applications. Clinical research might be expected to draw most heavily on papers studying the most scientifically significant genes. In practice, because it builds on the broader scientific literature, it may inherit its biases, particularly the tendency to concentrate on already crowded genes. To test which force dominates,

we re-estimate our main specification using clinical citations as the dependent variable. The results, reported in the first panel of Table 5, mirror those for academic citations: papers on less-studied genes consistently receive fewer clinical citations, even after controlling for the number of diseases associated with the gene. This indicates that the misallocation of attention in basic science carries through to the translational stage, with direct consequences for medical progress. Appendix Figure D2 illustrates this result, again showing that clinical research disproportionately rewards work on already crowded genes.

The second outcome we consider is the number of citations a paper receives from firm patents. Unlike academic citations, patent citations may be less sensitive to how concentrated prior research is, since corporate inventors are motivated by the most commercially promising applications for drug development. Accordingly, we estimate our main specification using citations from USPTO patents as the dependent variable. The results, reported in the bottom panel of Table 5, show that patent citations are indeed less influenced by whether a gene is heavily studied. Using the percentile rank, papers on less-studied genes attract more patent citations, while the below-median indicator yields less precise estimates. Figure D2 visualizes these findings, again showing no systematic bias of patent citations toward crowded genetic areas. Taken together, the evidence suggests that firms searching for commercially valuable innovations draw more heavily on research in less-studied genes, highlighting how academic incentives can steer attention away from precisely those targets with the greatest potential for industrial application.

6.4 Robustness Checks

A. Alternative Proxy of Genes' Scientific Importance: In Appendix C, we replicate all main analyses using an alternative measure of scientific importance based on the probability that a gene is differentially expressed in human disease. This proxy captures the likelihood that a gene is "switched on" in disease contexts, providing molecular evidence complementary to our baseline GWAS-based measure. The data are obtained from Northwestern University's *Find My Understudied Genes* database (Richardson et al., 2024). Results are robust and papers focusing on less-studied genes continue to experience a significant citation penalty, even after controlling for differential expression.

B. Including Publications Studying Multiple Genes: Appendix Table D2 extends the analysis to include publications that study multiple genes simultaneously. For these papers, we follow the standard assumption that researchers allocate their attention equally across all genes mentioned and

compute the average measure of scientific importance and crowdedness across them. The main findings are unaffected by this adjustment, confirming that the citation penalty for less-studied genes is not driven by our focus on monogenic publications.

7 Conclusion

In this paper, we model the process of academic inquiry and show that rewarding publications and citations can lead to excessive coordination among researchers and, in some cases, the formation of academic bubbles. Empirical evidence from genetics research supports both that concentrating on well-studied areas increases citations and that stronger career incentives predict greater crowding into these areas. Both patterns are consistent with the mechanisms identified in our model. Our normative analysis suggests that such incentives may still be necessary to motivate researchers, particularly when they are risk-averse, effort is costly, and the likelihood of discovering a promising topic is low.

More broadly, our findings underscore the importance of considering researcher incentives when evaluating the quality and impact of scientific work. From a principal-agent perspective, the decline in breakthrough ideas documented by Bloom et al. (2020) and Park et al. (2023) may reflect distorted incentives rather than diminishing scientific returns. If research stagnation stems from the structure of incentives rather than the limits of human creativity, then rethinking how we evaluate and reward researchers may help revitalize scientific progress.

References

Abbott, A., D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, and R. Van Noorden (2010): "Metrics: Do metrics matter?" *Nature*, 465, 860–862.

Acemoglu, D. (2023): "Distorted innovation: does the market get the direction of technology right?" *AEA Papers and Proceedings*, 113, 1–28.

Aghion, P., M. Dewatripont, and J. C. Stein (2008): "Academic freedom, private-sector focus, and the process of innovation," *The RAND Journal of Economics*, 39, 617–635.

Aghion, P., J. Van Reenen, and L. Zingales (2013): "Innovation and institutional ownership," *American Economic Review*, 103, 277–304.

Akcigit, U., D. Hanley, and N. Serrano-Velarde (2021): "Back to basics: Basic research spillovers, innovation policy, and growth," *The Review of Economic Studies*, 88, 1–43.

Azoulay, P., J. S. Graff Zivin, and G. Manso (2013): "National Institutes of Health peer review: Challenges and avenues for reform," *Innovation Policy and the Economy*, 13, 1–22.

Bhattacharya, J. and M. Packalen (2020): "Stagnation and scientific incentives," Tech. rep., National Bureau of Economic Research.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020): "Are ideas getting harder to find?" *American Economic Review*, 110, 1104–1144.

Bryan, K. A. and J. Lemus (2017): "The direction of innovation," *Journal of Economic Theory*, 172, 247–272.

Budish, E., B. N. Roin, and H. Williams (2015): "Do firms underinvest in long-term research? Evidence from cancer clinical trials," *American Economic Review*, 105, 2044–2085.

Bush, V. (1945): "Science, the Endless Frontier," *A Report to the President by the Director of the Office of Scientific Research and Development*.

Carnehl, C. and J. Schneider (2025): "A quest for knowledge," *Econometrica*, 93, 623–659.

Dasgupta, P. and P. A. David (1994): "Toward a new economics of science," *Research Policy*, 23, 487–521.

Diamond Jr, A. M. (1986): "What is a citation worth?" *Journal of Human Resources*, 200–215.

Edwards, A. M., R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and F. H. Yu (2011): "Too many roads not taken," *Nature*, 470, 163–165.

Eisfeldt, A. L. and A. A. Rampini (2006): "Capital reallocation and liquidity," *Journal of Monetary Economics*, 53, 369–399.

Ellison, G. (2013): "How does the market use citation data? The Hirsch index in economics," *American Economic Journal: Applied Economics*, 5, 63–90.

Gates, A. J., D. M. Gysi, M. Kellis, and A.-L. Barabási (2021): "A wealth of discovery built on the Human Genome Project—by the numbers," *Nature*, 590, 212–215.

Glode, V. and R. Lowery (2016): "Compensating financial experts," *The Journal of Finance*, 71, 2781–2808.

Hager, S., C. Schwarz, and F. Waldinger (2024): "Measuring science: Performance metrics and the allocation of talent," *American Economic Review*, 114, 4052–4090.

Haynes, W. A., A. Tomczak, and P. Khatri (2018): "Gene annotation bias impedes biomedical research," *Scientific Reports*, 8, 1362.

Hill, R. and C. Stein (2025a): "Race to the bottom: Competition and quality in science," *The Quarterly Journal of Economics*, 140, 1111–1185.

——— (2025b): "Scooped! Estimating rewards for priority in science," *Journal of Political Economy*, 133, 793–845.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2025): "The streetlight effect in data-driven exploration," Tech. rep., National Bureau of Economic Research.

Holmstrom, B. (1982): "Moral hazard in teams," *The Bell Journal of Economics*, 324–340.

——— (1989): "Agency costs and innovation," *Journal of Economic Behavior & Organization*, 12, 305–327.

Holmstrom, B. and P. Milgrom (1991): "Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design," *The Journal of Law, Economics, and Organization*, 7, 24–52.

Hopenhayn, H. and F. Squintani (2021): "On the direction of innovation," *Journal of Political Economy*, 129, 1991–2022.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124, 1403–1448.

Jones, C. I. (1995): "R & D-based models of economic growth," *Journal of Political Economy*, 103, 759–784.

Manso, G. (2011): "Motivating innovation," *The Journal of Finance*, 66, 1823–1860.

Marx, M. and A. Fuegi (2020): "Reliance on science: Worldwide front-page patent citations to scientific articles," *Strategic Management Journal*, 41, 1572–1594.

Merton, R. K. (1957): "Priorities in scientific discovery: A chapter in the sociology of science," *American Sociological Review*, 22, 635–659.

Myers, K. (2020): "The elasticity of science," *American Economic Journal: Applied Economics*, 12, 103–134.

Nelson, M. R., H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, et al. (2015): "The support of human genetic evidence for approved drug indications," *Nature Genetics*, 47, 856–860.

Nelson, R. R. (1962): *The rate and direction of inventive activity: Economic and social factors*, Princeton University Press.

Packalen, M. and J. Bhattacharya (2017): "Neophilia ranking of scientific journals," *Scientometrics*, 110, 43–64.

Park, M., E. Leahey, and R. J. Funk (2023): "Papers and patents are becoming less disruptive over time," *Nature*, 613, 138–144.

Pfeiffer, T. and R. Hoffmann (2007): "Temporal patterns of genes in scientific publications," *Proceedings of the National Academy of Sciences*, 104, 12052–12056.

Polanyi, M. (1962): "The Republic of Science: Its Political and Economic Theory," *Minerva*, 1, 54–73.

Priem, J., H. Piwowar, and R. Orr (2022): "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv preprint arXiv:2205.01833*.

Richardson, R., H. T. Navarro, L. A. N. Amaral, and T. Stoeger (2024): "Meta-research: understudied genes are lost in a leaky pipeline between genome-wide assays and reporting of results," *eLife*, 12, RP93429.

Rodriguez-Esteban, R. and X. Jiang (2017): "Differential gene expression in disease: a comparison between high-throughput studies and the literature," *BMC Medical Genomics*, 10, 59.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Stoeger, T., M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral (2018): "Large-scale investigation of the reasons why potentially important genes are ignored," *PLoS Biology*, 16, e2006643.

Strevens, M. (2003): "The role of the priority rule in science," *The Journal of Philosophy*, 100, 55–79.

Torvik, V. and N. Smalheiser (2021): "Author-ity 2018 PubMed Author Name Disambiguated Dataset," *University of Illinois at Urbana-Champaign*.

Tranchero, M. (2025): "Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies," *The Wharton School*.

Tripodi, G., X. Zheng, Y. Qian, D. Murray, B. F. Jones, C. Ni, and D. Wang (2025): "Tenure and research trajectories," *Proceedings of the National Academy of Sciences*, 122, e2500322122.

Tuckman, H. P. and J. Leahey (1975): "What is an article worth?" *Journal of Political Economy*, 83, 951–967.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang (2012): "Five years of GWAS discovery," *The American Journal of Human Genetics*, 90, 7–24.

Wang, J., R. Veugelers, and P. Stephan (2017): "Bias against novelty in science: A cautionary tale for users of bibliometric indicators," *Research Policy*, 46, 1416–1436.

Wei, C.-H., A. Allot, P.-T. Lai, R. Leaman, S. Tian, et al. (2024): "PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge," *Nucleic Acids Research*, 52, W540–W546.

Wu, L., D. Wang, and J. A. Evans (2019): "Large teams develop and small teams disrupt science and technology," *Nature*, 566, 378–382.

Xu, L., B. Li, S. Chen, and M. Li (2025): "Research productivity and novelty under different funding models: evidence from NIH-funded research projects," *Scientometrics*, 1–29.

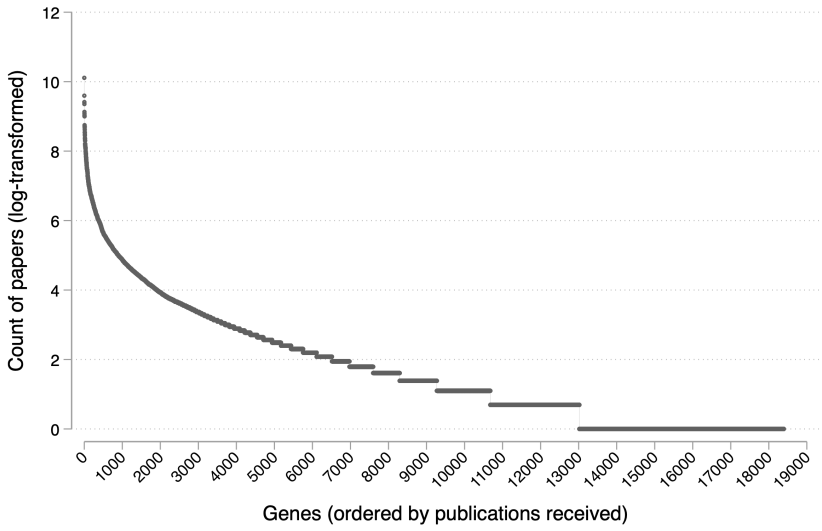
Yang, J. J., D. Grissa, C. G. Lambert, C. G. Bologa, et al. (2021): "TIGA: target illumination GWAS analytics," *Bioinformatics*, 37, 3865–3873.

Zhang, S., K. H. Wapman, D. B. Larremore, and A. Clauset (2022): "Labor advantages drive the greater productivity of faculty at elite universities," *Science Advances*, 8, eabq7056.

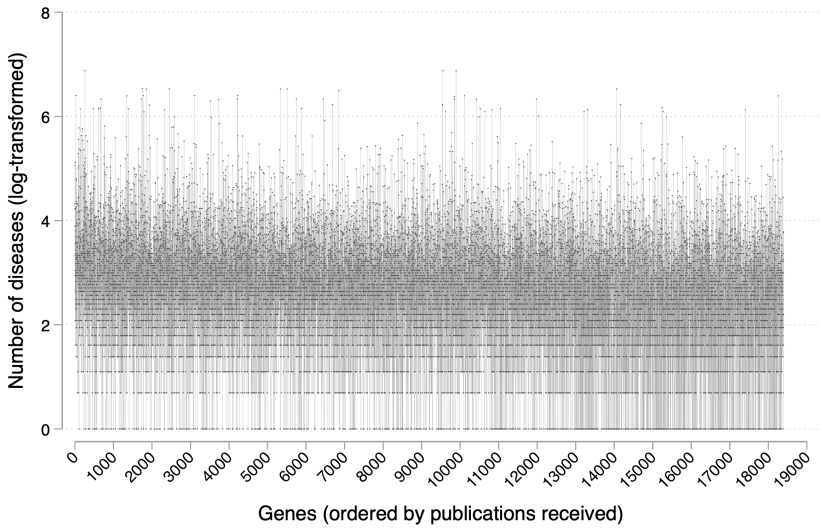
8 Figures and Tables

Figure 1: Research is Concentrated on Few Genes, While Many Disease-Relevant Genes Remain Neglected.

(a) *Count of Scientific Publications on Each Gene*



(b) *Count of Diseases Associated with Each Gene*



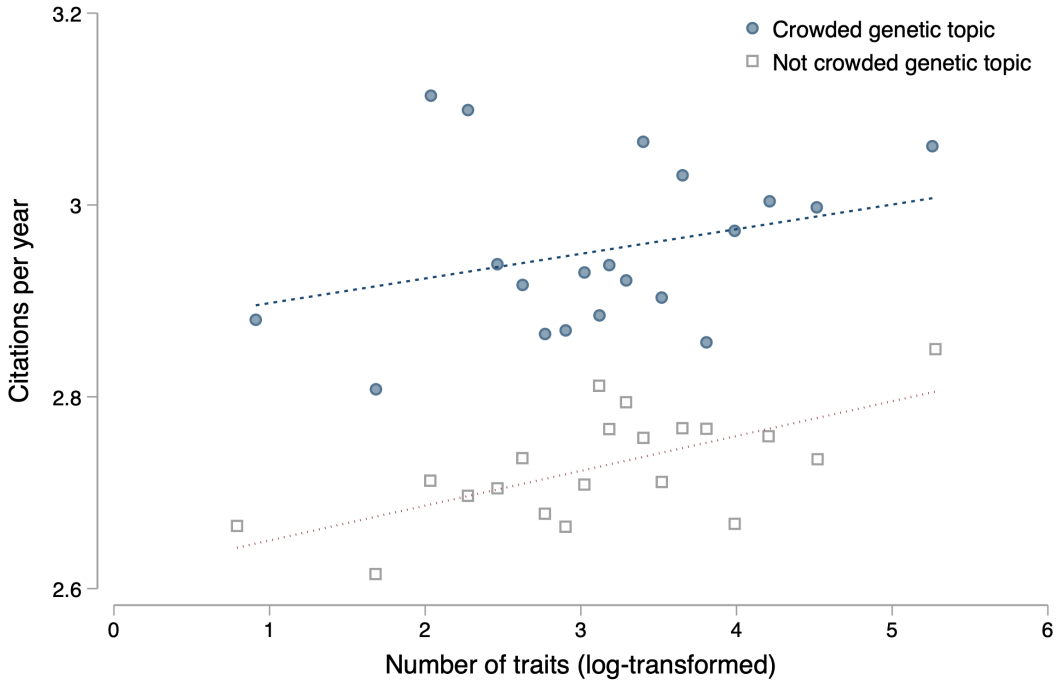
Note: The figure compares the distribution of scientific effort across human genes with the number of diseases associated with mutations in those same genes. Panel (a) shows the count of publications targeting each gene on the Y axis, with genes on the X axis sorted from the most to the least studied. Panel (b) shows the count of human diseases associated with each gene on the Y axis, with genes on the X axis sorted from the most to the least studied. Both panels present the genes sorted in the same way on the X axis to allow a comparison. The sample is the full analysis sample as defined in the text. See text for details.

Table 1: Descriptive Statistics.

	mean	median	st d	min	max	N
Citations per Year	2.853	1.364	6.939	0.000	750.400	857,025
Inverse Study Rank	7.454	2.000	14.413	0.000	99.000	857,025
Understudied (0/1)	0.439	0.000	0.496	0.000	1.000	857,025
Gene-Trait Associations	40.333	23.000	67.141	0.000	971.000	857,025
Has Mouse Ortholog (0/1)	0.917	1.000	0.276	0.000	1.000	857,025
Senior Author (0/1)	0.899	1.000	0.302	0.000	1.000	851,796
NIH Author (0/1)	0.015	0.000	0.121	0.000	1.000	857,025
Outside Top 10 University (0/1)	0.732	1.000	0.443	0.000	1.000	136,529
Clinical Citations per Year	0.091	0.000	0.431	0.000	107.800	857,023
Patent Citations per Year	0.001	0.000	0.016	0.000	3.500	857,025

Note: This table reports summary statistics at the paper level for 857,025 publications focusing on one human protein-coding genes between 1980 and 2018. *Citations per Year*: annual forward citations received from other academic articles (source: NIH iCite). *Inverse Study Rank*: percentile rank (inverted) of cumulative publication counts on the focal gene, where higher values indicate less studied genes. *Understudied (0/1)*: indicator equal to one if the focal gene had below-median cumulative publications at the time of publication. *Gene–Trait Associations*: number of diseases associated with the focal gene in data-driven genome-wide association studies (GWAS), drawn from TIGA (Yang et al., 2021). *Has Mouse Ortholog (0/1)*: indicator equal to one if the human gene has a homolog in laboratory mice, used in the instrumental variable analysis. *Senior Author (0/1)*: indicator equal to one if the last author (PI) has at least seven years of publishing history. *NIH Author (0/1)*: indicator equal to one if the PI is affiliated with the NIH Intramural Research Program. *Outside Top 10 University (0/1)*: indicator equal to one if the PI is not affiliated with a top 10 U.S. biomedical department, based on the 2006 CWTS Leiden Rankings. *Clinical Citations per Year*: annual forward citations received from clinical trial publications (source: NIH iCite). *Patent Citations per Year*: annual forward citations received from USPTO patents (source: Marx and Fuegi (2020)). See text for details.

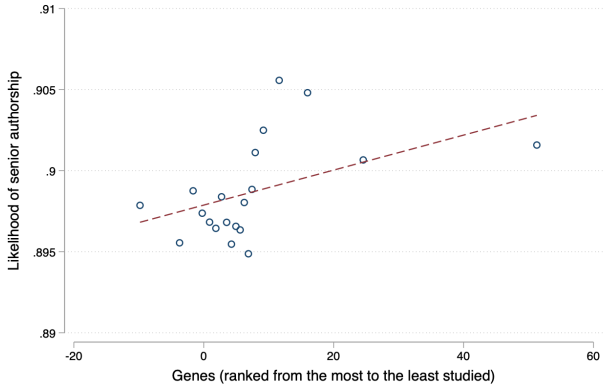
Figure 2: Publications in Crowded Genetic Fields Receive More Citations, Irrespective of the Disease Importance of the Gene.



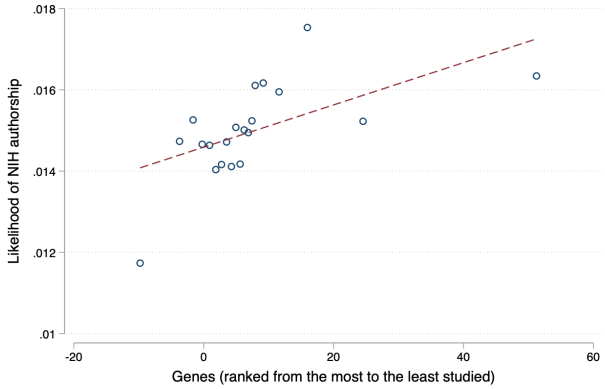
Note: This figure plots the relationship between yearly citations received by a publication and the biological importance of the gene it studies. The importance of a gene is proxied by the number of diseases associated with it in unbiased GWAS studies. The plot is presented as a binned scatterplot, separately for genes below and above the median of publications received. For both groups of genes, we residualize yearly citations and biological importance with respect to an indicator for each journal-year bin. We divide the sample into 20 equal-sized groups based on the ventiles of the biological importance measure and plot the mean of yearly cites against the mean of importance in each bin. The sample is the full analysis sample as defined in the text. See text for details.

Figure 3: Researchers With Plausibly Fewer Career Pressures Explore Less Studied Genes.

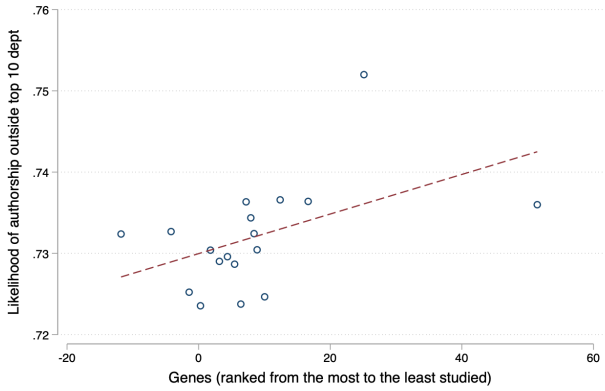
(a) *Senior Authors and Exploration of Less Studied Genes*



(b) *NIH Intramural Authors and Exploration of Less Studied Genes*



(c) *Authors in Less Prestigious Universities and Exploration of Less Studied Genes*



Note: This figure shows the likelihood that a publication features a senior author (panel (a)), an NIH intramural researcher (panel (b)), or a researcher affiliated with a less prestigious university (panel (c)), as a function of how much prior research has been conducted on the gene studied. The x-axis divides genes into 20 equal-sized bins based on their rank in the distribution of prior publications, from the most to the least studied. The plotted values are residualized with respect to journal-year fixed effects. The sample is the full analysis sample as defined in the text. See text for details.

Table 2: Citation Penalty to Papers Focusing on Less Studied Genes.

Panel A: Percentile Rank

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Inverse Study Rank	-0.0127***	-0.0122***	-0.00927***	-0.00924***
	(0.000503)	(0.000537)	(0.000867)	(0.000866)
Scientific Importance				0.0000847
				(0.000175)
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	No	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	832,994	790,650	604,906	604,906

Panel B: Dummy Version

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Understudied (0/1)	-0.297***	-0.260***	-0.156***	-0.154***
	(0.0152)	(0.0161)	(0.0257)	(0.0257)
Scientific Importance				0.000132
				(0.000175)
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	No	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	832,994	790,650	604,906	604,906

Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Citations*: average yearly scientific citations received by the publication; *Inverse Study Rank*: percentile rank of the gene studied by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied*: 0/1 = 1 for protein-coding human genes with a below-median number of publications; *Scientific Importance*: count of diseases linked to mutations in the gene, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. See text for details.

Table 3: Citation Penalty to Papers Focusing on Less Studied Genes Using an Instrumental-Variable Approach.

Panel A: Percentile Rank

	(1) Inverse Study Rank	(2) Citations	(3) Inverse Study Rank	(4) Citations
Mouse Homolog (0/1)	-0.356***		-0.361***	
	(0.0634)		(0.0925)	
Inverse Study Rank		-0.454***		-0.254*
		(0.0994)		(0.121)
F-Statistic (First Stage)	31.558		15.241	
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	790,650	790,650	604,906	604,906

Panel B: Dummy Version

	(1) Understudied (0/1)	(2) Citations	(3) Understudied (0/1)	(4) Citations
Mouse Homolog (0/1)	-0.0185***		-0.0150***	
	(0.00214)		(0.00301)	
Understudied (0/1)		-8.73***		-6.11*
		(1.50)		(2.76)
F-Statistic (First Stage)	75.202		24.972	
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	790,650	790,650	604,906	604,906

Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Citations*: average yearly scientific citations received by the publication; *Inverse Study Rank*: percentile rank of the gene studied by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied*: 0/1 = 1 for protein-coding human genes with a below-median number of publications; *Mouse Homolog (0/1)*: 0/1 = 1 for protein-coding genes with a homolog gene in the mouse, which allows them to be studied using the laboratory mouse; *Scientific Importance*: count of diseases linked to mutations in the gene, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. Both panels report the Kleibergen-Paap F statistic for the first-stage regression. See text for details.

Table 4: Researcher Characteristics and Exploration of Less Studied Genes.

Panel A: Senior Researchers

	(1) Inverse Study Rank	(2) Inverse Study Rank	(3) Understudied (0/1)	(4) Understudied (0/1)
Senior Author (0/1)	0.267***	0.331*	0.0128***	0.00877†
	(0.0536)	(0.166)	(0.00194)	(0.00520)
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	785,891	604,906	785,891	604,906

Panel B: NIH Intramural Researchers

	(1) Inverse Study Rank	(2) Inverse Study Rank	(3) Understudied (0/1)	(4) Understudied (0/1)
NIH Intramural Author (0/1)	0.582***	0.500*	0.0345***	0.0214**
	(0.164)	(0.252)	(0.00455)	(0.00663)
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	790,650	604,906	790,650	604,906

Panel C: Researchers in Lower-Ranked Institutions

	(1) Inverse Study Rank	(2) Inverse Study Rank	(3) Understudied (0/1)	(4) Understudied (0/1)
Author Outside Top 10 Dept (0/1)	0.251*	-0.0276	0.00757*	0.00773
	(0.117)	(0.375)	(0.00355)	(0.0107)
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	111,566	80,520	111,566	80,520

Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Inverse Study Rank*: percentile rank of the gene studied by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied*: 0/1 = 1 for protein-coding human genes with a below-median number of publications; *Senior Author*: 0/1 = 1 if the last author of the publication has been active in publishing for at least 7 years (and thus is likely tenured); *NIH Intramural Author*: 0/1 = 1 if the last author of the publication is affiliated exclusively with the NIH; *Outside Top 10 Dept*: 0/1 = 1 if the last author of the publication is affiliated with a U.S. university whose biomedical departments are outside the top 10 in the 2006 CWTS Leiden rankings; *Scientific Importance*: count of diseases linked to mutations in the gene, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. Panel A has a smaller sample because it is restricted to papers whose PI's seniority could be established. Panel C has a smaller sample because it is restricted to papers whose PI is affiliated with a U.S. university. See text for details.

Table 5: Clinical and Patent Citations to Papers Focusing on Less Studied Genes.

Panel A: Percentile Rank

	(1) Clinical Cites	(2) Clinical Cites	(3) Patent Cites	(4) Patent Cites
Inverse Study Rank	-0.00111***	-0.000713***	0.000007***	0.000006**
	(0.0000255)	(0.0000344)	(0.0000017)	(0.0000021)
Scientific Importance Control	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	790,650	604,906	790,650	604,906

Panel B: Dummy Version

	(1) Clinical Cites	(2) Clinical Cites	(3) Patent Cites	(4) Patent Cites
Understudied (0/1)	-0.0394***	-0.0249***	0.000124**	0.0000915
	(0.00100)	(0.00145)	(0.0000481)	(0.0000603)
Scientific Importance Control	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	790,650	604,906	790,650	604,906

Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Clinical Cites*: average yearly citations from clinical studies received by the publication; *Patent Cites*: average yearly citations from USPTO patents received by the publication; *Inverse Study Rank*: percentile rank of the gene studied by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied*: 0/1 = 1 for protein-coding human genes with a below-median number of publications; *Scientific Importance*: count of diseases linked to mutations in the gene, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. See text for details.

Academic Bubbles

Online Appendix

Richard Lowery, Michael Sockin, and Matteo Tranchero

A Appendix A: Proof of Propositions

Proof of Proposition 1:

In what follows, let $\mathcal{A}_t$ be the set of active fields at time t, and $\mathcal{E}_t$ its complement (i.e., the set of inactive fields), $R_{-n,lt}$ the number of researchers working in active topic l other than researcher n, and $\boldsymbol{X}_{-nt}$ the vector of statuses of other researchers. In addition, let $\boldsymbol{u_i}$ be $N \times 1$ Euclidean basis vector that is all zeros except for the i^{th} entry that is 1.

For simplicity, we first begin with the special case in which $\xi = 0$ and $\kappa = 0$, and no negative news ever arrives after a topic has an initial high signal, $S_{lj} = 1$. We then consider the more general case in the sequel.

Step 1: Researcher n's Hamilton-Jacobi-Bellman Equation

Recall that the state variable X_{nt} to be the status of researcher n with law of motion by Ito's Lemma

$$dX_{nt} = \sum_l r_{lt} r_{lt}^n dt + \log\left(1 + \sum_l \pi_{lj} r_{lt-}^n\right) dB_t^l - X_{nt-} dQ_{nt}. \tag{A.1}$$

A new researcher begins with a X_{nt} of $X_0 = 0$ because they have an initial citation of 0 and prestige of 1. In addition, we recognize her consumption is per capita output, i.e., $c_{nt} = \phi_A A_t K$.

To condense notation, let $\boldsymbol{p}_t = [p_{1t}, ..., p_{Lt}]'$ to be the vector of pseudo-probabilities of significance for each of the N fields (which also indexes which fields are active), and $\boldsymbol{X}_{-nt}$ to be the vector of other researchers' statuses $X_{n't}$. Let $\partial_y g$ denote the first partial derivative of the function g with respect to y, respectively, and $\nabla_{\boldsymbol{z}} g$ the gradient of g with respect to the vector $\boldsymbol{z}$. Further, let $\boldsymbol{u_i}$ be $N \times 1$ Euclidean basis vector that is all zeros except for the i^{th} entry that is 1.

The value function of the researcher $V_{nT_{i-1}^n}(t, X_{nt}, \boldsymbol{p}_t, \boldsymbol{X}_{-nt}, A_t)$ for a researcher n born at time T_{i-1}^n then satisfies the Hamilton-Jacobi-Bellman (HJB) Equation according to the Dynamic Pro-

gramming Principle

$$0 \geq \sup_{\boldsymbol{r}^n, \boldsymbol{e}^n, (\boldsymbol{r}_t^n + \boldsymbol{e}_t^n)' \iota_L \leq 1 \ \forall \ t} e^{-(\rho+\eta)t} \left(-e^{-\gamma \phi_A A_t K - \gamma \phi_X X_{nt} + \kappa (\boldsymbol{r}_t^n + \boldsymbol{e}_t^n)' \iota_L} \right) dt + \mathbb{E} \left[dV_{nT_{i-1}^n} \left(t, X_{nt}, \boldsymbol{p}_t, \boldsymbol{X}_{-nt}, A_t \right) \right]. \tag{A.2}$$

The HJB equation remains valid for random optional stopping times, and consequently is equipped to handle researcher retirement.

Let us conjecture that researcher n's value function takes the form

$$V_{nT_{i-1}^n} \left(t, X_{nt}, \boldsymbol{p}_t, \boldsymbol{X}_{-nt}, A_t \right) = -e^{-(\rho+\eta)\left(t - T_{i-1}^n\right) - v_a A_t - v_n X_{nt}} f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right).$$

We can then expand this HJB Equation by Ito's Lemma and factor out the $e^{-(\rho+\eta)t - v_a A_t - v_n X_{nt}}$ terms to arrive at

$$\begin{aligned} 0 \geq \sup_{\boldsymbol{r}^n, \boldsymbol{e}^n, (\boldsymbol{r}_t^n + \boldsymbol{e}_t^n)' \iota_L \leq 1 \ \forall \ t} \Bigg\{ & - e^{v_a A_t + v_n X_{nt} - \gamma \phi_A A_t K - \gamma \phi_X X_{nt} + \kappa (\boldsymbol{r}_t^n + \boldsymbol{e}_t^n)' \iota_L} + v_n f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right) \sum_l \left(R_{-n,lt} + r_{lt}^n \right) r_{lt}^n \\ & - \lambda \sum_{l \in \mathcal{A}_t} \left(R_{-n,lt} + r_{lt}^n \right) \left(\left(\frac{p_{lt} e^{-v_a}}{\left(1 + r_{lt}^n\right)^{v_n}} + \left(1 - p_{lt}\right) e^{v_a} \right) f_n \left(\boldsymbol{p}_t - p_{lt} \boldsymbol{u_l}, \boldsymbol{X}_{-nt-} + \boldsymbol{\Delta X}_{-nt} \right) - f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right) \right) \\ & - \sum_{l \in \mathcal{E}_t} \lambda \left(E_{-n,lt} + e_{lt}^n \right) \left(\mathbb{E} \left[f_n \left(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}, \boldsymbol{X}_{-nt} \right) \right] - f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right) \right) + \left(\rho + \eta \right) f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right) \\ & - \sum_{i,l} \left[\nabla_{\boldsymbol{X}_{-n}} f_n \right]_i \left(R_{-n,lt} + r_{lt}^n \right) r_{it}^l - \sum_{n' \neq n} \eta \left(f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} - X_{-n't} \boldsymbol{u_{n'}} \right) - f_n \left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt} \right) \right) \Bigg\}, \end{aligned} \tag{A.3}$$

where $R_{-n,lt}$ and $E_{-n,lt}$ are the number of researchers researching and investigating in each topic without researcher n, respectively. The second term reflects the marginal impact of an increase in citations on the researcher's status. The third is the jump in the researcher's value function that accumulates when a topic matures. The fourth is the researcher's value function when a new topic is discovered. The fifth is the flow discounting in value from subjective discounting and the instantaneous probability of the researcher's retirement. The sixth term reflects the growth in the status of other researchers and the last term reflects the stochastic their retirement. The inequality becomes an equality under the optimal policy.

Matching coefficients on the A_t and X_{nt} terms, we find that

$$v_a = \gamma \phi_A K,$$

and

$$v_n = \gamma \phi_X,$$

which confirms the conjecture.

Step 2: Optimal Research Policies

Suppose researcher n chooses in which topic to research only among active fields. We take the convention that $r_{lt}+1$ represents the number of researchers in topic l with researcher n, and r_{lt} the number without researcher n, and similarly with e_{lt}. Their optimal choice is then

$$\begin{aligned} v_{nt}^R = \sup_{l \in \mathcal{A}_t} \Big\{ & \gamma \phi_X f_n\left(r_{lt}+1\right) - \lambda p_{lt} e^{-\gamma \phi_A K} \left(\left(r_{lt}+1\right) 2^{-\gamma \phi_X} - r_{lt} \right) f_n\left(\boldsymbol{p}_t - p_{lt}\boldsymbol{u_l}, \boldsymbol{X}_{-nt-} + \boldsymbol{\Delta X}_{-nt}\right) \\ & - \lambda \left(\left(1-p_{lt}\right) e^{\gamma \phi_A K} f_n\left(\boldsymbol{p}_t - p_{lt}\boldsymbol{u_l}, \boldsymbol{X}_{-nt-} + \boldsymbol{\Delta X}_{-nt}\right) - f_n\left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt}\right) \right) - \sum_i \left[\nabla_{\boldsymbol{X}_{-n}} f_n\right]_i r_{lt}^i \Big\}, \end{aligned} \tag{A.4}$$

for fields in which $p_{lt} \in (0,1)$. The $\sum_{i,n'} \left[\nabla_{\boldsymbol{X}_{-n}} f_n\right]_i r_{it}^{n'}$ reflects that researcher n internalizes that they can raise the status of other researchers by investigating in field l because $r_{lt} = \sum_{i=1}^N r_{it}^l$. The last two positive terms reflect that the researcher internalizes that their choice of field impacts the number of researchers working in that field, and consequently the citations of other researchers and variance of the signals about the quality of each topic. It is immediate that the benefit of researching a given active topic is increasing in the number of other researchers in the topic (i.e., r_{lt}).

If the researcher instead tries to investigate their own topic, then the payoff from this is risky endeavor is

$$v_{nt}^E = \sup_{l \in \mathcal{E}_t} -\lambda \left(\mathbb{E}\left[f_n\left(\boldsymbol{p}_t + p_{lj}\boldsymbol{u_l}, \boldsymbol{X}_{-nt}\right)\right] - f_n\left(\boldsymbol{p}_t, \boldsymbol{X}_{-nt}\right)\right). \tag{A.5}$$

The optimal choice for researcher n is to research an active topic if $v_{nt}^R > v_{nt}^E$ and to research a new topic if $v_{nt}^R \leq v_{nt}^E$.

Notice now that v_{nt}^E and v_{nt}^R are independent of X_{nt} for researcher n. By symmetry, then, $v_{n't}^E$ and $v_{n't}^R$ are independent of $X_{n't}$ for researcher n'. If no researcher's optimal research policy depends on his status, then by rational expectations, researcher n recognizes that researcher n''s research policy does not depend on his own status. Consequently, researcher n does not need to keep track of researcher n''s status, or that of any other researcher. Consequently, by this argument, it must

be the case that $f_n(\boldsymbol{p}_t, \boldsymbol{X}_{-nt}) = f_n(\boldsymbol{p}_t)$, which is the same for all researchers, and equations (A.4) and (A.5) simplify to

$$v_{nt}^R = \sup_{l \in \mathcal{A}_t} \Big\{ -\lambda \left(p_{lt} e^{-\gamma \phi_A K} \left((R_{-n,lt} + 1) 2^{-\gamma \phi_X} - R_{-n,lt} \right) + (1 - p_{lt}) e^{\gamma \phi_A K} \right) f_n(\boldsymbol{p}_t - p_{lt} \boldsymbol{u_l}) + (\lambda + \gamma \phi_X (R_{-n,lt} + 1)) f_n(\boldsymbol{p}_t) \Big\}, \tag{A.6}$$

and

$$v_{nt}^E = \sup_{l \in \mathcal{E}_t} -\lambda \left(\mathbb{E}\left[f_n(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}) \right] - f_n(\boldsymbol{p}_t) \right). \tag{A.7}$$

where

$$\mathbb{E}\left[f_n(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}) \right] = \Pr(S_{lj} = 1) f_n(\boldsymbol{p}_t + p_1 \boldsymbol{u_l}) + \Pr(S_{lj} = 0) f_n(\boldsymbol{p}_t + p_0 \boldsymbol{u_l}), \tag{A.8}$$

$\Pr(S_{lj} = 1) = pq_1 + (1 - p) q_0$, and $\Pr(S_{lj} = 0) = 1 - pq_1 - (1 - p) q_0$. Because the value of investigating a new topic is always non-negative, a researcher will always prefer to investigate than to not research at all. Consequently, the optimal choice of topic is always a non-empty set surely for all time.

Step 3: Researcher n's Maximized Hamilton-Jacobi-Bellman Equation

Let l_{nt}^* be the optimal topic for researcher n to research at time t. From the researcher's Hamilton-Jacobi-Bellman Equation (A.3), the researcher's maximized Hamilton-Jacobi-Bellman equation can be expressed as

$$0 = -1 + \left(\gamma \phi_X r_{l_{nt}^* t} r_{l_{nt}^* t}^n + \lambda \sum_l (r_{lt} + e_{lt}) + \rho + \eta \right) f_n(\boldsymbol{p}_t) - \sum_{l \in \mathcal{E}_t} \lambda e_{lt} \mathbb{E}\left[f_n(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}) \right] - \lambda \sum_{l \in \mathcal{A}_t} r_{lt} \left(p_{lt} (1 + r_{lt}^n)^{-\gamma \phi_X} e^{-\gamma \phi_A K} + (1 - p_{lt}) e^{\gamma \phi_A K} \right) f_n(\boldsymbol{p}_t - p_{lt} \boldsymbol{u_l}), \tag{A.9}$$

Because researchers will always either research an existing or a new topic from Step 2, it follows that $\sum_l (r_{lt} + e_{lt}) = N$. Substituting this and the expectation (A.8) into equation (A.9), we arrive

at

$$
\begin{aligned}
0 = &-1 + \left(\rho + \eta + \gamma\phi_X r_{l^*_{nt}t} r^n_{l^*_{nt}t} + \lambda N\right) f_n\left(\boldsymbol{p}_t\right) \\
&- \lambda \sum_{l \in \mathcal{E}_t} e_{lt} \left(\left(pq_1 + (1-p)\, q_0\right) f\left(\boldsymbol{p}_t + p_1 \boldsymbol{u_l}\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f_n\left(\boldsymbol{p}_t + p_0 \boldsymbol{u_l}\right)\right) \\
&- \lambda \sum_{l \in \mathcal{A}_t} r_{lt} \left(p_{lt} \left(1 + r^n_{lt}\right)^{-\gamma\phi_X} e^{-\gamma\phi_A K} + \left(1 - p_{lt}\right) e^{\gamma\phi_A K}\right) f_n\left(\boldsymbol{p}_t - p_{lt} \boldsymbol{u_l}\right). \qquad \text{(A.10)}
\end{aligned}
$$

Step 4: Incentives to Coordinate on Research Activities

We first consider the incentives to investigate different new topics. Notice that all inactive research fields are ex ante identical, and therefore deliver equivalent value to all researchers. As such, $f_n\left(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}\right) = f_n\left(\boldsymbol{p}_t + p_{l'j} \boldsymbol{u_{l'}}\right)$ if $l, l' \in \mathcal{E}_t$. Consequently, we have that for all $n \in 1, ..., N$

$$
v^E_{nt} = \mathbb{E}\left[f_n\left(\boldsymbol{p}_t + p_{lj} \boldsymbol{u_l}\right)\right] - f_n\left(\boldsymbol{p}_t\right) = \mathbb{E}\left[f_n\left(\boldsymbol{p}_t + p_{l'j} \boldsymbol{u_{l'}}\right)\right] - f_n\left(\boldsymbol{p}_t\right) \;\forall\, l, l' \in \mathcal{E}_t.
$$

Further, because the total value in the Hamilton-Jacobi-Bellman Equation (A.10) of investigating new research topics for researcher n is

$$
\sum_{l \in \mathcal{E}_t} \lambda e_{lt} \left(\left(pq_1 + (1-p)\, q_0\right) f\left(\boldsymbol{p}_t + p_1 \boldsymbol{u_{l'}}\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f\left(\boldsymbol{p}_t + p_0 \boldsymbol{u_{l'}}\right) - f\left(\boldsymbol{p}_t\right)\right),
$$

it follows that any combination of investigative efforts $e^{nN}_{ltn=1, l \in \mathcal{E}_t}$ will deliver the same expected value. To see this, suppose there are 2 researchers and there are two fields, 1 and 2, that are both inactive. Then, if both researchers only investigate field 1,each earns an expected value from investigating of

$$
2\lambda \left(\left(pq_1 + (1-p)\, q_0\right) f_n\left(p_1, 0\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f_n\left(p_0, 0\right) - f_n\left(0, 0\right)\right),
$$

while if one researcher investigates field 1 and one field 2, each earns an expected value from investigating of

$$
\begin{aligned}
&\lambda \left(\left(pq_1 + (1-p)\, q_0\right) f_n\left(p_1, 0\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f_n\left(p_0, 0\right) - f_n\left(0, 0\right)\right) \\
+&\lambda \left(\left(pq_1 + (1-p)\, q_0\right) f_n\left(0, p_1\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f_n\left(0, p_0\right) - f_n\left(0, 0\right)\right) \\
=&2\lambda \left(\left(pq_1 + (1-p)\, q_0\right) f_n\left(p_1, 0\right) + \left(1 - pq_1 - (1-p)\, q_0\right) f_n\left(p_0, 0\right) - f_n\left(0, 0\right)\right),
\end{aligned}
$$

because both fields deliver the same conditional continuation value regardless of in which a new topic is discovered. Consequently, any combination of efforts to investigate new topics delivers the

same value, and only the total effort is identified. Further, conditional continuation values are the same for all inactive topics.

We now consider the incentives to investigate different active topics. In contrast to the case of inactive topics, researchers are not indifferent between researching one active topic l versus another l' because each may garner different citation counts. Notice that in the absence of career concerns (i.e., $\gamma\phi_X = 0$), active researchers are indifferent to which active topic they research with the same probability of being significant p_{lj} because any combination of research efforts $r^{nN}_{ltn=1,l\in\mathcal{A}_t}$ will deliver the same expected value. To see this, suppose there are again 2 researchers and there are two fields, 1 and 2, that are both active and likely significant (i.e., $p_{1j} = p_{2j} = p_1$). Then, if both researchers only research field 1,each earns an expected value from researching of

$$2\lambda\left(\left(p_1 e^{-\gamma\phi_A K} + (1-p_1)\, e^{\gamma\phi_A K}\right) f_n\left(0, p_1\right) - f_n\left(p_1, p_1\right)\right),$$

while if one researcher works on field 1 and one on field 2, each earns an expected value from researching of

$$\begin{aligned}
&\lambda\left(\left(p_1 e^{-\gamma\phi_A K} + (1-p_1)\, e^{\gamma\phi_A K}\right) f_n\left(0, p_1\right) - f_n\left(p_1, p_1\right)\right)\\
+&\lambda\left(\left(p_1 e^{-\gamma\phi_A K} + (1-p_1)\, e^{\gamma\phi_A K}\right) f_n\left(p_1, 0\right) - f_n\left(p_1, p_1\right)\right)\\
=&2\lambda\left(\left(p_1 e^{-\gamma\phi_A K} + (1-p_1)\, e^{\gamma\phi_A K}\right) f_n\left(0, p_1\right) - f_n\left(p_1, p_1\right)\right),
\end{aligned}$$

because both fields deliver the same conditional continuation value regardless of in which one matures. Consequently, any combination of efforts to investigate active topics delivers the same value, and only the total effort is identified. Further, conditional continuation values are the same for all active topics.

However, in the presence of career concerns (i.e., $\gamma\phi_X > 0$), researcher n is no longer indifferent toward which topic he researches because of the flow benefit $\gamma\phi_X r_{lt} f_n\left(\boldsymbol{p}_t\right)$ from the r_{lt} additional citations and the $\left(1 + r^n_{lt}\right)^{-\gamma\phi_X}$ term in the $p_{lt}\left(1 + r^n_{lt}\right)^{-\gamma\phi_X} e^{-\gamma\phi_A K} f_n\left(\boldsymbol{p}_t - p_{lt}\boldsymbol{u_l}\right) - f_n\left(\boldsymbol{p}_t\right)$ benefit from the field maturing. Both of these additional forces favor coordinating on the same topic as other researchers.

Consequently, in the presence of career concerns, all researchers who research an active field coordinate on the same field with the highest likelihood of being significant, i.e., a field for which $p_{lj} = p_1$ before a field for which $p_{lj} = p_0$, although on which specific field they coordinate is indeterminate.

Because all researchers behave in a symmetric manner and do not distinguish between active topics that received a high signal $S_{lj} = 1$ or between active topics that received a low signal $S_{lj} = 0$, we can shrink the state space into the number of active fields that received a high signal $S_{lj} = 1$, L_1, and the number of active fields that received a low signal $S_{lj} = 0$,L_0. The number of inactive fields is then $L - L_1 - L_0$. It follows $f_n(\boldsymbol{p}_t) = f(L_1, L_0)$. Further, because all researchers prioritize researching an active topic with a high signal over that with a low signal or investigating, we need only consider the case where $L_1 \in 0, 1$.

Step 5: Solving for the Maximized Hamilton-Jacobi-Bellman Equation

Because all researchers face the same incentives to research, it follows that all researchers will either research an active topic or all will investigate a new topic. Suppose they all investigate a new topic, in which case there is no active topic with a high signal $S_{lj} = 1$, and consequently $L_1 = 0$ (otherwise they would research it instead). Then, because all new topics deliver $f(1, L_0)$ with probability $pq_1 + (1-p)q_0$ and $f(0, L_0 + 1)$ with probability $1 - (pq_1 + (1-p)q_0)$, it follows that researcher n's value Hamilton-Jacobi-Bellman Equation in this case is

$$f(0, L_0) = \frac{1 + \lambda N (pq_1 + (1-p)q_0) f(1, L_0)}{\rho + \eta + \lambda N} + \frac{\lambda N (1 - pq_1 - (1-p)q_0) f(0, L_0 + 1)}{\rho + \eta + \lambda N}. \tag{A.11}$$

Similarly, if all researchers research an active topic that received a high signal $S_{lj} = 1$ and $L_1 = 1$, then researcher n's value Hamilton-Jacobi-Bellman Equation in this case is instead

$$f(1, L_0) = \frac{1 + \lambda N \left(p_1 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_1) e^{\gamma\phi_A K}\right) f(0, L_0)}{\rho + \eta + \gamma\phi_X N + \lambda N}. \tag{A.12}$$

Finally, if all research an active topic that received a low signal $S_{lj} = 0$, which will only occur if $L_1 = 0$ and there is no active topic with a high signal $S_{lj} = 1$, then researcher n's value Hamilton-Jacobi-Bellman Equation in this case is

$$f(0, L_0) = \frac{1 + \lambda N \left(p_0 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_0) e^{\gamma\phi_A K}\right) f(0, L_0 - 1)}{\rho + \eta + \gamma\phi_X N + \lambda N}. \tag{A.13}$$

Because researchers will always research an active topic that that received a high signal $S_{lj} = 1$,

we can substitute equation (A.12) (shifting the L_1 index forward by 1) into (A.11) to find

$$f(0,L_0) = \frac{1+\lambda N\left(pq_1+(1-p)q_0\right)\frac{1+\lambda N\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_1)e^{\gamma\phi_A K}\right)f(0,L_0)}{\rho+\eta+\gamma\phi_X N+\lambda N}}{\rho+\eta+\lambda N} + \frac{\lambda N\left(1-pq_1-(1-p)q_0\right)f(0,L_0+1)}{\rho+\eta+\lambda N}. \tag{A.14}$$

We now have two cases depending on whether it is preferable to research an active topic that received a low signal $S_{lj}=0$ or to continue investigating until researchers discover a topic that received a high signal $S_{lj}=1$.

Case 1: It is preferable to research topics with low signals to investigating new topics. In this case, we also have that $L_0 \in 0,1$, and we can recover $f(0,1)$ from equation (A.13) and substitute it (shifting the L_0 index forward by 1) into equation (A.14) to find

$$f(0,0) = \frac{1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}}{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\delta\right)}, \tag{A.15}$$

where

$$\begin{aligned}\delta &= \left(1-pq_1-(1-p)q_0\right)\left(p_0 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_0)e^{\gamma\phi_A K}\right)\\ &+\left(pq_1+(1-p)q_0\right)\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_1)e^{\gamma\phi_A K}\right)\\ &>0.\end{aligned}$$

For the solution to be well-defined, we require that

$$\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\delta\right)>0.$$

Consequently, the optimal research policy for all researchers is to coordinate to research any active topic, then investigate a new topic once it has matured.

Case 2: It is preferable to investigate new topics than to research topics with low signals. In this case, we never use equation (A.13). Instead, we recognize that $f(0,L_0+1)=f(0,L_0)=f(0)$ (we can ignore the count of low signal topics), i.e., the new topic with a low signal is abandoned, and equation (A.14) reduces to

$$f(0) = \frac{1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\left(pq_1+(1-p)q_0\right)}{\rho+\eta+\lambda N(1-\chi)}, \tag{A.16}$$

where

$$\chi = 1 - (pq_1 + (1-p)\,q_0)\left(1 - \frac{\lambda N\left(p_1 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_1)\,e^{\gamma\phi_A K}\right)}{\rho + \eta + \gamma\phi_X N + \lambda N}\right) > 0,$$

which is decreasing in $\gamma\phi_X$, and it follows that

$$f(1) = \frac{1 + \lambda N\left(p_1 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_1)\,e^{\gamma\phi_A K}\right) f(0)}{\rho + \eta + \gamma\phi_X N + \lambda N}. \tag{A.17}$$

For the solution to be well-defined, we require that

$$\rho + \eta + \lambda N(1-\chi) > 0.$$

Consequently, the optimal research policy for all researchers is to coordinate to research only active topics with high signals $S_{lj} = 1$, and to investigate a new topic otherwise.[8]

Step 6: Pecking Order for Research Activities

We have already identified that all researchers will prioritize researching topics with high signals $S_{lj} = 1$. We now need to examine when an individual researcher has incentive to research an active topic with a low signal. The marginal value of researching the active topic, v_{nt}^R, from equation (A.6) in this case simplifies to

$$v_{nt|S_{lj}=0}^R = (\lambda + \gamma\phi_X r_{lt})\,f(0,1) - \lambda\left(p_0 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_0)\,e^{\gamma\phi_A K}\right) f(0,0), \tag{A.18}$$

while the marginal value of investigating from equation (A.5) is

$$v_{nt}^E = -\lambda\left((pq_1 + (1-p)\,q_0)\,f(1,1) + (1 - pq_1 - (1-p)\,q_0)\,f(0,2) - f(0,1)\right). \tag{A.19}$$

For it to be preferable to investigate new topics and ignore active topics with low signals, we require that $v_{nt}^E \geq v_{nt|S_{lj}=0}^R$, i.e.,

$$\begin{aligned}\gamma\phi_X r_{lt} f(0,1) - \lambda\left(p_0 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + (1-p_0)\,e^{\gamma\phi_A K}\right) f(0,0) \\ \leq -\lambda\left((pq_1 + (1-p)\,q_0)\,f(1,1) + (1 - pq_1 - (1-p)\,q_0)\,f(0,2)\right).\end{aligned} \tag{A.20}$$

[8]This characterization fails in the extreme case that all fields have active topics with low signals, in which case there is a bottleneck and researchers are forced to complete one of these fields to then investigate to try to discover a high signal field. We consequently consider L to be sufficiently large that this happens with an arbitrarily low probability.

Considering we are in Case 2 from Step 5, we recognize $f(0,2) = f(0,1) = f(0,0) = f(0)$, and substituting with equations (A.17) and (A.16), condition (A.20) when one researcher deviates to research the active topic (i.e., $r_{lt} = 1$) (and assuming $\rho + \eta + \lambda N(1-\chi) > 0$) reduces to

$$\begin{aligned} 0 \geq \gamma\phi_X + \lambda\left(1 - p_0 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} - (1-p_0)\, e^{\gamma\phi_A K}\right) \\ - (p_1 - p_0)\left(e^{\gamma\phi_A K} - 2^{-\gamma\phi_X} e^{-\gamma\phi_A K}\right) \lambda N \frac{\lambda\left(pq_1 + (1-p)\, q_0\right)}{\rho + \eta + \gamma\phi_X N + \lambda N}. \end{aligned} \tag{A.21}$$

This is a conservative incentive compatibility condition because if one researcher has incentive to deviate, then $m > 1$ researchers have stronger incentive to deviate through the higher citation count to deviating. This ensures there is no coordination failure in which all researchers may suddenly revert to not researching active speculative topics.

Suppose there are no career incentives (i.e., $\phi_X = 0$), then the condition (A.21) reduces to

$$1 - p_0 e^{-\gamma\phi_A K} - (1-p_0)\, e^{\gamma\phi_A K} - (p_1 - p_0)\left(e^{\gamma\phi_A K} - e^{-\gamma\phi_A K}\right) \frac{\lambda N\left(pq_1 + (1-p)\, q_0\right)}{\rho + \eta + \lambda N} \leq 0. \tag{A.22}$$

Given that $p_1 > p_0$ by construction, and $e^{\gamma\phi_A K} > e^{-\gamma\phi_A K}$,it follows that the final term on the left-hand side of condition (A.22) is negative. Further, since $1 - p_0 e^{-\gamma\phi_A K} - (1-p_0)\, e^{\gamma\phi_A K}$ is increasing in p_0 and negative when $p_0 < \frac{1-e^{-\gamma\phi_A K}}{1-e^{-2\gamma\phi_A K}}$, we have that for $p_0 < \frac{1-e^{-\gamma\phi_A K}}{1-e^{-2\gamma\phi_A K}}$, condition (A.22) is satisfied and researchers will never research an active topic with a low signal $S_{jl} = 0$.

In contrast, suppose researchers have career concerns (i.e., $\phi_X > 0$). Let LHS represent the left-hand side of condition (A.21). Then, it is immediate that

$$\begin{aligned} \frac{dLHS}{d\phi_X} = 1 + \lambda\left(p_0 - (p_1 - p_0)\frac{\lambda N\left(pq_1 + (1-p)\, q_0\right)}{\rho + \eta + \gamma\phi_X N + \lambda N}\right) 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} \log 2 \\ + (p_1 - p_0)\left(e^{\gamma\phi_A K} - 2^{-\gamma\phi_X} e^{-\gamma\phi_A K}\right)\left(pq_1 + (1-p)\, q_0\right)\left(\frac{\lambda N}{\rho + \eta + \gamma\phi_X N + \lambda N}\right)^2 \end{aligned}$$

which is increasing in ϕ_X for ϕ_X sufficiently large. Consequently, there is a critical ϕ_X, ϕ_X^*, such that for $\phi_X \geq \phi_X^*$, condition (A.21) fails and researchers will instead research active topics that receive low signals. Further, if $p_0 \geq (p_1 - p_0)\frac{\lambda N(pq_1 + (1-p)q_0)}{\rho + \eta + \gamma\phi_X N + \lambda N}$, then $\frac{dLHS}{d\phi_X} \geq 0$, and for $\phi_X < \phi_X^*$, condition (A.22) is satisfied.

Step 7: Transversality and Sufficiency

It is a necessary condition that the value function satisfy the transversality condition

$$\lim_{T\to\infty} \mathbb{E}\left[V_{nT_{i-1}^n}\left(T, X_{nT}, \boldsymbol{p}_T, \boldsymbol{X}_{-nT}, A_T\right)\right] = \lim_{T\to\infty} \mathbb{E}\left[-e^{-(\rho+\eta)\left(T-T_{i-1}^n\right)-\gamma\phi_A K A_T-\gamma\phi_X X_{nT}} f\left(L_0, L_1\right)\right] = 0. \tag{A.23}$$

It is immediate from Step 5 that $f\left(L_0, L_1\right) \geq 0$ is bounded. Because status X_{nt} is non-negative, $-e^{-\gamma\phi_X X_{nt}}$ is bounded from above by 0 and below by -1. In addition, $-e^{-(\rho+\eta)\left(T-T_{i-1}^n\right)-\gamma\phi_A K A_T}$ is bounded from above by 0 and, given the jumps in A_t are finite and (at maximum) arrive slower than the rate of discounting $\rho+\eta$, it follows that $-e^{-(\rho+\eta)\left(T-T_{i-1}^n\right)-\gamma\phi_A K A_T}$ is also bounded in expectation from below. Consequently, the transversality condition (A.23) is satisfied.

Standard arguments establish that it is sufficient that the value function $V_{nt}\left(X_{nT}, \boldsymbol{p}_T, \boldsymbol{X}_{-nT}, A_T\right)$ satisfies the Hamilton-Jacobi-Bellman Equation (A.10) and the transversality condition (A.23) for it to solved researcher n's problem and delivers value $U_{nT_{i-1}^n}$.

Consequently, we have solved the researcher n's problem and, as a consequence, the full equilibrium in the special case in which $\xi = 0$.

Step 8: The Case in which $\xi > 0$ and $\kappa > 0$

We can draw insights from the special case in which $\xi = 0$ to discern the equilibrium in which $\xi > 0$, and negative news about promising topics can arrive. This is true regardless of effort cost κ.

It is immediate that the best situation for researchers still is that there is a promising topic that received an initial signal of $S_{lj} = 1$. Amplified by career concerns, all researchers will research any active topic with a high signal. However, there is now the possibility that topic will receive a second, negative signal before it matures. Because researching an active promising topic is preferable to researching a speculative topic or investigating a new topic, researchers will research the topic until either it matures or the negative signal arrives.

If a negative signal arrives, then by the memory-less property of Poisson processes, this is equivalent to the arrival of new speculative topic. As such, our analysis for whether the researchers research the speculative topic or abandon it remains valid.

The normalized value of an active promising topic consequently now satisfies

$$f\left(1, L_0\right) = \frac{e^{\kappa} + \lambda N\left(p_1 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} + \left(1-p_1\right) e^{\gamma\phi_A K}\right) f\left(0, L_0\right) + \xi f\left(0, L_0+1\right)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}, \tag{A.24}$$

We again have two cases depending on whether it is preferable to research an active topic that received a low signal $S_{lj} = 0$ or to continue investigating until researchers discover a topic that received a high signal $S_{lj} = 1$.

Case 1: It is preferable to research topics with low signals to investigating new topics. We can repeat our procedure in Step 5 in this case to find

$$f(0,0) = \frac{1 + \frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}}{\rho + \eta + \lambda N\left(1 - \frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)} e^{\kappa}, \tag{A.25}$$

where

$$\begin{aligned}\theta = &-(p_1 - p_0)\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\left(pq_1 + (1-p)q_0\right)\left(e^{\gamma\phi_A K} - 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}\right) \\ &+ p_0 2^{-\gamma\phi_X}e^{-\gamma\phi_A K} + (1-p_0)e^{\gamma\phi_A K}.\end{aligned} \tag{A.26}$$

Notice because $\frac{\partial\theta}{\partial\phi_X} < 0$, we have that $\frac{\partial f(0,0)}{\partial\phi_X} < 0$.

For the solution to be well-defined, we require that

$$\rho + \eta + \lambda N\left(1 - \frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right) > 0.$$

Case 2: It is preferable to investigate new topics than to research topics with low signals. We can repeat our procedure in Step 5 in this case to find

$$f(0) = \frac{1 + \frac{\lambda N(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}}{\rho + \eta + \lambda N(1-\omega)} e^{\kappa}, \tag{A.27}$$

where

$$\omega = 1 - \left(pq_1 + (1-p)q_0\right)\left(1 - \frac{\lambda N\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K} + (1-p_1)e^{\gamma\phi_A K}\right) + \xi}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\right), \tag{A.28}$$

which is again decreasing in ϕ_X, and that

$$f(1) = \frac{e^{\kappa} + \lambda N\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K} + (1-p_1)e^{\gamma\phi_A K}\right)f(0) + \xi f(0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}. \tag{A.29}$$

Notice again that because $\frac{\partial\omega}{\partial\phi_X} < 0$, we have that $\frac{\partial f(0)}{\partial\phi_X} < 0$.

We can further derive the necessary and sufficient condition for a researcher not to deviate to

research an active speculative topic (i.e., $r_{lt} = 1$) instead of investigating a new one. Based on our derivation in Step 6, we can substitute equations (A.27) and (A.29) into condition (A.20), recognizing $f(0,2) = f(0,1) = f(0,0) = f(0)$ and assuming $\rho + \eta + \lambda N(1-\omega) > 0$, to arrive at

$$\begin{aligned} 0 \geq\ & \gamma\phi_X + \lambda\left(1 - p_0 2^{-\gamma\phi_X} e^{-\gamma\phi_A K} - (1-p_0)\, e^{\gamma\phi_A K}\right) \\ & - (p_1 - p_0)\left(e^{\gamma\phi_A K} - 2^{-\gamma\phi_X} e^{-\gamma\phi_A K}\right) \lambda N \frac{\lambda\left(pq_1 + (1-p)\, q_0\right)}{\rho + \eta + \xi + \gamma\phi_X N + \lambda N}. \end{aligned} \tag{A.30}$$

Analogous arguments to those in Step 6 establish that for $p_0 < \frac{1-e^{-\gamma\phi_A K}}{1-e^{-2\gamma\phi_A K}}$, a researcher would never deviate to research a speculative topic when there are no career incentives, i.e., $\gamma\phi_X = 0$. Similarly, there is a critical ϕ_X, ϕ_X^*, such that for $\phi_X \geq \phi_X^*$, condition (A.30) fails and researchers will instead research active topics that receive low signals. Noticeably, these results do not on the effort cost, κ.

The equilibrium strategies we characterized in the special case when $\xi = 0$ and $\kappa = 0$ therefore remain valid when $\xi > 0$ and $\kappa > 0$.

What remains to be shown is that a researcher has the benefits of research exceed the effort costs. Notice that researching an active topic must yield (weakly) conditionally more value to a researcher than investigating to find a new topic. Consequently, we need only show that a researcher wishes to investigate a new topic despite the effort cost.

Substituting our verified solution, $V_{nt}(X_{nt}, \boldsymbol{p}_t, \boldsymbol{X}_{-nt}) = -e^{-(\rho+\eta)t - v_a A_t - v_n X_{nt}} f_n(L_1, L_0)$, the Hamilton-Jacobi-Bellman equation for a investigating a new topic is given by

$$\begin{aligned} 0 \geq\ & \min 1, e^{\kappa} + \lambda N\left(\left(pq_1 + (1-p)\, q_0\right) f(1,0) + \left(1 - pq_1 - (1-p)\, q_0\right) f(0,1) - f(0,0)\right) \\ & - (\rho + \eta)\, f(0,0). \end{aligned}$$

Consequently, we need only verify the one-shot deviation principle that

$$1 \geq e^{\kappa} + \lambda N\left(\left(pq_1 + (1-p)\, q_0\right) f(1,0) + \left(1 - pq_1 - (1-p)\, q_0\right) f(0,1) - f(0,0)\right), \tag{A.31}$$

for it to be optimal for a researcher always to engage in research.

In Case 2, when active speculative topics are abandoned, $f(0,1) = f(0)$ and $f(1,0) = f(1)$, and

condition (A.31) from equations (A.27) and (A.29) reduces to

$$\kappa \leq \kappa_1^* = -\log\left(\left(1+\frac{\lambda N\left(pq_1+(1-p)\,q_0\right)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\left(1-\frac{\rho+\eta+\xi+\gamma\phi_X N+\lambda N+\lambda N\left(pq_1+(1-p)\,q_0\right)}{\frac{(\rho+\eta)(\rho+\eta+\xi+\gamma\phi_X N+\lambda N)}{\rho+\eta+\gamma\phi_X N+\lambda N\left(1-p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}-(1-p_1)e^{\gamma\phi_A K}\right)}+\lambda N\left(pq_1+(1-p)\,q_0\right)}\right)\right)\right). \tag{A.32}$$

Similarly, for Case 1, when active speculative topics are researched to maturity, condition (A.31) from equation (A.25) reduces to

$$\begin{aligned}\kappa \leq \kappa_2^* = -\log\Bigg(1+\left(\frac{\left(pq_1+(1-p)\,q_0\right)\xi}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}+1\right)\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N} \\ +\frac{\lambda N\left(1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\right)\chi}{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)}\Bigg).\end{aligned} \tag{A.33}$$

where

$$\begin{aligned}\chi = \left(pq_1+(1-p)\,q_0\right)\frac{\lambda N\left(p_1-p_0\right)\left(2^{-\gamma\phi_X}e^{-\gamma\phi_A K}-e^{\gamma\phi_A K}\right)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N} \\ +\left(\frac{\left(pq_1+(1-p)\,q_0\right)\xi}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}+1\right)\frac{\lambda N\left(p_0 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_0)\,e^{\gamma\phi_A K}\right)}{\rho+\eta+\gamma\phi_X N+\lambda N}-1.\end{aligned} \tag{A.34}$$

Step 9: The Participation Constraint

Notice that the participation constraint (8) is tightest when there is no active topic. Consequently, if it is satisfied in this case, then it will also be satisfied when there is an active topic. We again have two cases depending on whether it is preferable to research an active topic that received a low signal $S_{lj}=0$ or to continue investigating until researchers discover a topic that received a high signal $S_{lj}=1$.

Case 1: It is preferable to research topics with low signals to investigating new topics. At inception when there is no active topic, a researcher's value function because $X_{nT_{i-1}^n}=0$ and $c_{nT_{i-1}^n}=\phi_A A_{T_{i-1}^n}K$

$$V_{nT_{i-1}^n}\left(T_{i-1}^n, X_{nT_{i-1}^n}, A_{T_{i-1}^n}\right) = -e^{-\gamma c_{nT_{i-1}^n}} f\left(0,0\right) \geq -\frac{e^{-\gamma c_{nT_{i-1}^n}}}{\rho+\eta},$$

from which it follows that we require from equation (A.25)

$$\frac{1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}}{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)}e^{\kappa} \leq \frac{1}{\rho+\eta}, \tag{A.35}$$

We can rewrite the participation constraint as

$$\kappa \leq \kappa_1^{pc} = \log\left(\frac{1}{\rho+\eta}\frac{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)}{1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}}\right). \tag{A.36}$$

Case 2: It is preferable to investigate new topics than to research topics with low signals. At inception when there is no active topic, a researcher's value function because $X_{nT_{i-1}^n} = 0$ and $c_{nT_{i-1}^n} = \phi_A A_{T_{i-1}^n} K$

$$V_{nT_{i-1}^n}\left(T_{i-1}^n, X_{nT_{i-1}^n}, A_{T_{i-1}^n}\right) = -e^{-\gamma c_{nT_{i-1}^n}} f(0) \geq -\frac{e^{-\gamma c_{nT_{i-1}^n}}}{\rho+\eta},$$

from which it follows that we require from equation (A.25)

$$\frac{1+\frac{\lambda N(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}}{\rho+\eta+\lambda N(1-\omega)}e^{\kappa} \leq \frac{1}{\rho+\eta}, \tag{A.37}$$

and the left-hand-side is again decreasing in ϕ_X. We can rewrite the participation constraint as

$$\kappa \leq \kappa_1^{pc} = \log\left(\frac{1}{\rho+\eta}\frac{\rho+\eta+\lambda N(1-\omega)}{1+\frac{\lambda N(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}}\right). \tag{A.38}$$

Consequently, the participation constraint is relaxed the higher is the reward for citations, ϕ_X.

Proof of Proposition 2:

We consider both cases depending on whether it is preferable to research an active topic that received a low signal $S_{lj} = 0$ or to continue investigating until researchers discover a topic that received a high signal $S_{lj} = 1$.

Case 1: From Step 8 of the proof of Proposition **1,** $f(0,0)$ is decreasing in ϕ_X. It is also immediate that $f(0,0)$ is increasing in κ. In addition, we recognize from rewriting θ from equation (A.26) as

$$\begin{aligned}\theta = {} & (1-pq_1+(1-p)q_0)\left(p_0 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_0)e^{\gamma\phi_A K}\right)\\ & +(pq_1+(1-p)q_0)\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_1)e^{\gamma\phi_A K}\right)\\ & +\frac{\xi(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\left(p_0 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+(1-p_0)e^{\gamma\phi_A K}\right),\end{aligned} \tag{A.39}$$

that θ is increasing in γ for γ sufficiently large, and therefore so is $f(0,0)$. It is also immediate that θ is decreasing in p_0 and p_1, and therefore so is $f(0,0)$.

Consequently, the left-hand-side of the participation constraint (A.35) is decreasing in ϕ_X, p_0, and p_1, relaxing the constraint, and increasing in κ and γ for γ sufficiently large, tightening the constraint.

Case 2: From Step 8 of the proof of Proposition **1,** $f(0)$ is decreasing in ϕ_X. It is further immediate that $f(0)$ is increasing in κ. In addition, we recognize that ω from equation (A.28) is increasing in γ for γ sufficiently large, and therefore so is $f(0)$, while it is decreasing in p_1, and therefore so is $f(0)$.

Consequently, the left-hand-side of the participation constraint (A.37) is decreasing in ϕ_X and p_1, relaxing the constraint, and increasing in κ and γ for γ sufficiently large, tightening the constraint.

Proof of Proposition 3:

Let the value function for researcher n in the first-best economy with costly effort be $e^{-(\rho+\eta)t}v_n(A_t, L_1, L_0)$, where L_0 is again the number of active speculative and L_1 is the number of active promising research topics.

Researching Active topic**s:** Because effort is contractible, for it to be optimal to have researcher n research an active speculative topic if $m-1$ other researchers are actively researching it at time t, it must be the case by the Hamilton-Jacobi-Bellman Equation for researcher n that

$$\begin{aligned}-e^{-\frac{\gamma}{N}A_tK+\kappa}+\lambda m\left(p_0v_n\left(A_t+1,L_1,L_0-1\right)+\left(1-p_0\right)v_n\left(A_t-1,L_1,L_0-1\right)-v_n\left(A_t,L_1,L_0\right)\right)\\ -\rho v_n\left(A_t,L_1,L_0\right)\geq\\ -e^{-\frac{\gamma}{N}A_tK}-\lambda\left(m-1\right)\left(p_0v_n\left(A_t+1,L_1,L_0-1\right)+\left(1-p_0\right)v_n\left(A_t-1,L_1,L_0-1\right)-v_n\left(A_t,L_1,L_0\right)\right)\\ -\rho v_n\left(A_t,L_1,L_0\right).\end{aligned} \tag{A.40}$$

We focus on the expected benefit for the M^{th} researcher to ensure that having M researchers actively researching represents a Pareto improvement. The remaining $N-M$ researchers who are not actively researching are clearly better off if those who are required to provide effort are better off.

Conjecturing that $v_n(A_t, L_1, L_0) = -e^{-\frac{\gamma}{N}A_tK}h_n(L_1, L_0)$,condition A.40 reduces to

$$e^{\kappa}+\lambda\left(\left(p_0e^{-\frac{\gamma}{N}K}+\left(1-p_0\right)e^{\frac{\gamma}{N}K}\right)h_n\left(L_1,L_0-1\right)-h_n\left(L_1,L_0\right)\right)\leq 1. \tag{A.41}$$

If having a speculative topic is socially valuable, then it must be the case that $h_n(L_1, L_0) <$

$h_n(L_1, L_0 - 1)$. In addition, because $p_0 < \frac{1}{2}$, we have that$p_0 < \frac{1-e^{-\frac{\gamma}{N}K}}{1-e^{-2\frac{\gamma}{N}K}}$ and consequently $p_0 e^{-\frac{\gamma}{N}K} + (1-p_0)\, e^{\frac{\gamma}{N}K} > 1$. These two observations imply that

$$\left(p_0 e^{-\frac{\gamma}{N}K} + (1-p_0)\, e^{\frac{\gamma}{N}K}\right) h_n(L_1, L_0 - 1) - h_n(L_1, L_0) > h_n(L_1, L_0 - 1) - h_n(L_1, L_0) > 0,$$

and because $\kappa > 0$ that condition A.41 fails. Consequently, it is not socially optimal (for any m) to research a speculative topic.

In contrast, for it to be optimal to have researcher n research an active promising topic if $m-1$ other researchers are actively researching it at time t, it must be the case by the Hamilton-Jacobi-Bellman Equation for researcher n that

$$\begin{aligned}
-e^{-\frac{\gamma}{N}A_t K + \kappa} + \lambda m \left(p_1 v_n(A_t + 1, L_1 - 1, L_0) + (1-p_1)\, v_n(A_t - 1, L_1 - 1, L_0) - v_n(A_t, L_1, L_0)\right) \\
-\rho v_n(A_t, L_1, L_0) + \xi\left(v_n(A_t, L_1 - 1, L_0 + 1) - v_n(A_t, L_1, L_0)\right) \geq \\
-e^{-\frac{\gamma}{N}A_t K} - \lambda(m-1)\left(p_1 v_n(A_t + 1, L_1 - 1, L_0) + (1-p_1)\, v_n(A_t - 1, L_1 - 1, L_0) - v_n(A_t, L_1, L_0)\right) \\
-\rho v_n(A_t, L_1, L_0) + \xi\left(v_n(A_t, L_1 - 1, L_0 + 1) - v_n(A_t, L_1, L_0)\right),
\end{aligned} \tag{A.42}$$

where the last term in each expression reflects the possibility that bad news arrives about the topic. Because this arrival is independent of the number of active researchers in the topic, it does not meaningfully impact the optimality condition.

Conjecturing that $v_n(A_t, L_1, L_0) = -e^{-\frac{\gamma}{N}A_t K} h_n(L_1)$ (where we drop the L_0 argument because it is never optimal to research a speculative topic), condition A.42 reduces to

$$e^{\kappa} + \lambda\left(\left(p_1 e^{-\frac{\gamma}{N}K} + (1-p_1)\, e^{\frac{\gamma}{N}K}\right) h_n(L_1 - 1) - h_n(L_1)\right) \leq 1. \tag{A.43}$$

This is the key condition we will evaluate after solving for researcher's value function.

Investigating a New topic**:** We assume for now that it is optimal for m researchers to to research an active promising topic. In this case, in equilibrium, there will be either zero or one active promising topics. In contrast, we have established that the principal would abandon any speculative active topic.

For it to be optimal for $M \leq N$ researchers to investigate a new topic, there cannot be an active promising topic it must be the case by the Hamilton-Jacobi-Bellman Equation for researcher n's

value function in the first-best economy with costly effort, $e^{-(\rho+\eta)t}v_n\left(A_t, L_1\right)$,

$$\begin{aligned} -e^{-\frac{\gamma}{N}A_tK+\kappa} - \lambda M\left(pq_1+\left(1-p\right)q_0\right)\left(v_n\left(A_t,1\right)-v_n\left(A_t,0\right)\right) \\ -\left(\rho+\eta\right)v_n\left(A_t,0\right) \geq \\ -e^{-\frac{\gamma}{N}A_tK} + \lambda\left(M-1\right)\left(pq_1+\left(1-p\right)q_0\right)\left(v_n\left(A_t,1\right)-v_n\left(A_t,0\right)\right) \\ -\left(\rho+\eta\right)v_n\left(A_t,0\right), \end{aligned} \tag{A.44}$$

where $pq_1+\left(1-p\right)q_0$ is the probability a new topic is promising, and

$$\left(\rho+\eta+\lambda M\right)v_n\left(A_t,0\right) = -e^{-\frac{\gamma}{N}A_tK+\kappa} + \lambda M\left(pq_1+\left(1-p\right)q_0\right)\left(v_n\left(A_t,1\right)-v_n\left(A_t,0\right)\right), \tag{A.45}$$

and

$$\begin{aligned} \left(\rho+\eta\right)v_n\left(A_t,1\right) = -e^{-\frac{\gamma}{N}A_tK+\kappa} + \lambda m\left(p_1v_n\left(A_t+1,0\right)+\left(1-p_1\right)v_n\left(A_t-1,0\right)-v_n\left(A_t,1\right)\right), \\ +\xi\left(v_n\left(A_t,0\right)-v_n\left(A_t,1\right)\right) \end{aligned} \tag{A.46}$$

are the value functions conditional on having researcher n investigating and researching an active promising topic, respectively. The last term in equation A.46 reflects the arrival of negative news.

Let us conjecture that $v_n\left(A_t,L_1\right) = -e^{-\frac{\gamma}{N}A_tK}h_n\left(L_1\right)$,from which follows that condition (A.44) reduces to

$$1-e^{\kappa}+\lambda\left(pq_1+\left(1-p\right)q_0\right)\left(h_n\left(0\right)-h_n\left(1\right)\right)\geq 0, \tag{A.47}$$

equation (A.45) reduces to

$$h_n\left(0\right) = \frac{e^{\kappa}+\lambda M\left(pq_1+\left(1-p\right)q_0\right)h_n\left(1\right)}{\rho+\eta+\lambda M\left(pq_1+\left(1-p\right)q_0\right)}, \tag{A.48}$$

and equation (A.46) reduces to

$$h_n\left(1\right) = \frac{e^{\kappa}+\lambda m\left(p_1e^{-\frac{\gamma}{N}K}+\left(1-p_1\right)e^{\frac{\gamma}{N}K}\right)h_n\left(0\right)+\xi h_n\left(0\right)}{\rho+\eta+\xi+\lambda m}, \tag{A.49}$$

which confirms the conjecture. Notice that $p_1 \geq \frac{1-e^{-\frac{\gamma}{N}K}}{1-e^{-2\frac{\gamma}{N}K}}$ implies that $p_1e^{-\frac{\gamma}{N}K}+\left(1-p_1\right)e^{\frac{\gamma}{N}K} \leq 1$. Combining equations $(A.48)$ and (A.49), we find

$$h_n\left(0\right) = \frac{e^{\kappa}}{\rho+\eta+\lambda m\left(1-p_1e^{-\frac{\gamma}{N}K}-\left(1-p_1\right)e^{\frac{\gamma}{N}K}\right)\frac{\lambda M\left(pq_1+\left(1-p\right)q_0\right)}{\rho+\eta+\xi+\lambda m+\lambda M\left(pq_1+\left(1-p\right)q_0\right)}}, \tag{A.50}$$

and

$$h_n(0) - h_n(1) = \frac{e^{\kappa}}{\lambda M\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda m+\lambda M(pq_1+(1-p)q_0)}{\lambda m\left(1-p_1 e^{-\frac{\gamma}{N}K}-(1-p_1)e^{\frac{\gamma}{N}K}\right)}} > 0. \tag{A.51}$$

Substituting equation $(A.51)$ into condition $(A.51)$, condition $(A.51)$ becomes

$$\frac{\lambda\left(pq_1 + (1-p)\,q_0\right)}{\lambda M\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda m+\lambda M(pq_1+(1-p)q_0)}{\lambda m\left(1-p_1 e^{-\frac{\gamma}{N}K}-(1-p_1)e^{\frac{\gamma}{N}K}\right)}} \geq 1 - e^{-\kappa}. \tag{A.52}$$

The Two Optimality Conditions:

Substituting equations A.50 and A.51 into conditions $A.43$, we arrive at the optimality condition to research an active promising field (recognizing there is, at most, one active promising field at time t)

$$\frac{1}{m}\frac{\rho+\eta+\xi+\lambda M\left(pq_1 + (1-p)\,q_0\right)}{\lambda M\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda m+\lambda M(pq_1+(1-p)q_0)}{\lambda m\left(1-p_1 e^{-\frac{\gamma}{N}K}-(1-p_1)e^{\frac{\gamma}{N}K}\right)}} \geq 1 - e^{-\kappa}. \tag{A.53}$$

Consequently, for it to be optimal to investigate to find new topics and to research any active promising topics, conditions $A.53$ and $A.52$ must be satisfied. The principal's goal is to have both constraints be as tight as possible to maximize the number of active researchers. It is intuitive that having more researchers research active promising topics (i.e., higher m) relaxes the constraint on M in condition $A.52$, while a higher M tightens it.

If $M \geq m$, then it is immediate that if condition $A.52$ is satisfied, then so is condition $A.53$. We focus on this case when m is as large as possible, i.e., $M = m$, as this slackens condition $A.52$ the most. Then, we can rewrite condition condition $A.52$ as

$$f(m) = \frac{\lambda\left(pq_1 + (1-p)\,q_0\right)}{\lambda m\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda m(1+pq_1+(1-p)q_0)}{\lambda m\left(1-p_1 e^{-\frac{\gamma}{N}K}-(1-p_1)e^{\frac{\gamma}{N}K}\right)}} \geq 1 - e^{-\kappa}. \tag{A.54}$$

Notice that $f(m)$ is hump-shaped in m, equaling zero at $m \in 0, \infty$. There are consequently either zero solutions to condition $A.54$ or one value of m, m^*, for which $f(m^*) \geq 1 - e^{-\kappa}$ and $f(m^*+1) < 1 - e^{-\kappa}$, and this m^* is decreasing in κ (and always equal to N, the maximum value of m, when $\kappa = 0$). By feasibility, the active number of researchers is the largest $m \leq N$ that satisfies condition $A.54$. It may be the case that no feasible m satisfies condition $A.54$, in which case there is no solution.

Further, if $\kappa \leq \kappa^*$, where

$$\kappa^* = -\log\left(1 - \frac{\lambda\left(pq_1 + (1-p)\,q_0\right)}{\lambda N\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda N(1+pq_1+(1-p)q_0)}{\lambda N\left(1-p_1 e^{-\frac{\gamma}{N}K} - (1-p_1)e^{\frac{\gamma}{N}K}\right)}}\right), \tag{A.55}$$

then it is optimal for all N researchers to actively research at all times.

The Participation Constraint:

Finally, the participation constraint (8) is given by

$$-e^{-\gamma c_{nT_{i-1}^n}} h_n\left(0\right) \geq -\frac{e^{-\gamma c_{nT_{i-1}^n}}}{\rho+\eta}, \tag{A.56}$$

which reduces from equation $(A.50)$ to

$$\kappa \leq \kappa^{pc} = \log\left(\frac{\rho+\eta+\lambda m\left(1 - p_1 e^{-\frac{\gamma}{N}K} - (1-p_1)\,e^{\frac{\gamma}{N}K}\right)\frac{\lambda M(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\lambda m+\lambda M(pq_1+(1-p)q_0)}}{\rho+\eta}\right). \tag{A.57}$$

Proof of Proposition 4:

We need only verify that condition $A.54$ fails when $m = 1$ and $\kappa = \kappa^*$. Substituting with κ^* from equation A.55 into condition $A.54$ when it fails, we require that

$$\begin{aligned} f\left(1\right) &= \frac{\lambda\left(pq_1 + (1-p)\,q_0\right)}{\lambda\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda(1+pq_1+(1-p)q_0)}{\lambda\left(1-p_1 e^{-\frac{\gamma}{N}K} - (1-p_1)e^{\frac{\gamma}{N}K}\right)}} \\ &\leq \frac{\lambda\left(pq_1 + (1-p)\,q_0\right)}{\lambda N\left(pq_1 + (1-p)\,q_0\right) + (\rho+\eta)\,\frac{\rho+\eta+\xi+\lambda N(1+pq_1+(1-p)q_0)}{\lambda N\left(1-p_1 e^{-\frac{\gamma}{N}K} - (1-p_1)e^{\frac{\gamma}{N}K}\right)}}, \end{aligned} \tag{A.58}$$

which is satisfied if

$$\left(1 - p_1 e^{-\frac{\gamma}{N}K} - (1-p_1)\,e^{\frac{\gamma}{N}K}\right) N \leq \frac{(\rho+\eta)\,(\rho+\eta+\xi)}{\lambda^2\left(pq_1 + (1-p)\,q_0\right)}. \tag{A.59}$$

Because the left-hand size of Equation A.59is increasing in N, fixing p_1, it follows that if $N < \tilde{N}$, where $\tilde{N}$is the N such that Equation A.59 is satisfied with equality, then zero active researchers constitutes a Nash Equilibrium.

That zero versus N active researchers constitutes a Pareto inferior equilibrium is immediate because all researchers are ex-ante identical and, by revealed preference, the principal prefers to actively

investigate and research promising topics than to not.

Proof of Proposition 5:

Step 1: Principal's Indirect Utility Function

Given the linearity of the principal's objective in Y_t and X_{nt}, and that researchers are ex-ante identical up to their status, we can solve for the principal's problem researcher-by-researcher. Let $e^{-\rho t}U_{Pn}(A, X, S)$be the per capital value the principal derives from researcher n, where $S = \emptyset$ if there is no active topic, $S = 1$ if there is an active promising topic, and $S = 0$ if there is an active speculative topic. The Hamilton-Jacobi-Bellman Equation governing the principal's value when there is no active topic is

$$0 = \lambda N \left(\left(pq_1 + (1-p) q_0 \right) U_{Pn}(A, X, 1) + \left(1 - pq_1 - (1-p) q_0 \right) U_{Pn}(A, X, 0) - U_{Pn}(A, X, \emptyset) \right) \tag{A.60}$$
$$+ \frac{1}{N}(1 - \phi_Y) AK - \phi_X X + \eta \left(U_{Pn}(A, 0, \emptyset) - U_{Pn}(A, X, \emptyset) \right) - \rho U_{Pn}(A, X, \emptyset).$$

Its value when there is an active promising topic similarly has law of motion

$$\begin{aligned} 0 = {} & \lambda N \left(p_1 U_{Pn}(A+1, X + \log(1+N), \emptyset) + (1-p_1) U_{Pn}(A-1, X, \emptyset) - U_{Pn}(A, X, 1) \right) \\ & + \frac{1}{N}(1-\phi_Y) AK - \phi_X X \\ & + \partial_X U_{Pn}(A, X, 1) N + \eta \left(U_{Pn}(A, 0, 1) - U_{Pn}(A, X, 1) \right) - \rho U_{Pn}(A, X, 1). \end{aligned} \tag{A.61}$$

If speculative topics are not researched, we can conjecture that

$$0 = \lambda N \left(p_0 U_{Pn}(A+1, X + \log(1+N), \emptyset) + (1-p_0) U_{Pn}(A-1, X, \emptyset) - U_{Pn}(A, X, 0) \right) \tag{A.62}$$
$$+ \frac{1}{N}(1-\phi_Y) AK - \phi_X X$$
$$+ \partial_X U_{Pn}(A, X, 0) N + \eta \left(U_{Pn}(A, 0, 0) - U_{Pn}(A, X, 0) \right) - \rho U_{Pn}(A, X, 0).$$

We now consider two cases depending on whether only promising or all topics are researched. This is determined by each researcher's incentive compatibility constraint such that if career incentives, ϕ_X, are sufficiently high, then all topics are researched, while if it is not, then only promising topics are researched to maturity.

Case 1: Suppose researchers only pursue promising topics, and prefer to investigate a new topic than to research a speculative topic. Let us conjecture that

$$U_{Pn}(A, X, S) = u_0(S) + u_A(S) A + u_X(S) X,$$

and $U_{Pn}(A, X, 0) = U_{Pn}(A, X, \emptyset)$. Substituting these conjectures into equations $A.60$ and $A.61$, we find

$$u_A(\emptyset) = u_A(1) = \frac{1}{\rho N}(1 - \phi_Y) K,$$
$$u_X(\emptyset) = u_X(1) = -\frac{\phi_X}{\rho + \eta},$$

and

$$U_{Pn}(A, X, S) = -\frac{u(s)}{\rho}\left(\frac{\phi_X}{\rho + \eta} N + \frac{\phi_X}{\rho + \eta}\lambda N p_1 \log(1 + N) + \lambda N (1 - 2p_1)\frac{1}{\rho N}(1 - \phi_Y) K\right) + \frac{1}{\rho N}(1 - \phi_Y) AK - \frac{\phi_X}{\rho + \eta} X,$$

where

$$u(\emptyset) = \frac{\lambda N (p q_1 + (1 - p) q_0)}{\rho + \lambda N (1 + p q_1 + (1 - p) q_0)},$$
$$u(1) = \frac{\rho + \lambda N (p q_1 + (1 - p) q_0)}{\rho + \lambda N (1 + p q_1 + (1 - p) q_0)}.$$

If there is no active topic and we initialize all statuses at zero, then the principal's indirect utility function is

$$U_{Pn}(A, 0, \emptyset) = -\frac{1}{\rho}\frac{\lambda N (p q_1 + (1 - p) q_0)}{\rho + \lambda N (1 + p q_1 + (1 - p) q_0)}\left(\frac{\phi_X}{\rho + \eta} N + \frac{\phi_X}{\rho + \eta}\lambda N p_1 \log(1 + N)\right) - \frac{1}{\rho}\frac{\lambda N (p q_1 + (1 - p) q_0)}{\rho + \lambda N (1 + p q_1 + (1 - p) q_0)}\lambda N (1 - 2p_1)\frac{1}{\rho N}(1 - \phi_Y) K + \frac{1}{\rho N}(1 - \phi_Y) AK,$$

Case 2: Suppose researchers pursue both promising and speculative topics, and prefer to research them to investigating a new topic. We again conjecture that

$$U_{Pn}(A, X, S) = u_0(S) + u_A(S) A + u_X(S) X,$$

and substitute this conjecture in equations $A.60$-$A.62$ to find

$$u_A(\emptyset) = u_A(1) = u_A(0) = \frac{1}{\rho N}(1-\phi_Y)K,$$
$$u_X(\emptyset) = u_X(1) = u_A(0) = -\frac{\phi_X}{\rho+\eta},$$

and

$$U_{Pn}(A,X,S) = \frac{1}{\rho N}(1-\phi_Y)AK - \frac{\phi_X}{\rho+\eta}X + u_0(s),$$

where

$$u_0(\emptyset) = -\frac{1}{\rho}\frac{\lambda N}{\rho+2\lambda N}\left(\frac{\phi_X}{\rho+\eta}N + \lambda N\frac{\phi_X}{\rho+\eta}p\log(1+N) + \lambda N(1-2p)\frac{1}{\rho N}(1-\phi_Y)K\right),$$
$$u_0(1) = \frac{-\frac{\phi_X}{\rho+\eta}\lambda N p_1\log(1+N) - \lambda N(1-2p_1)\frac{1}{\rho N}(1-\phi_Y)K - \frac{\phi_X}{\rho+\eta}N}{\rho+\lambda N} + \frac{\lambda N}{\rho+\lambda N}u_0(\emptyset)$$
$$u_0(0) = \frac{-\frac{\phi_X}{\rho+\eta}\lambda N p_0\log(1+N) - \lambda N(1-2p_0)\frac{1}{\rho N}(1-\phi_Y)K - \frac{\phi_X}{\rho+\eta}N}{\rho+\lambda N} + \frac{\lambda N}{\rho+\lambda N}u_0(\emptyset).$$

If there is no active topic and we initialize all statuses at zero, then the principal's indirect utility function is

$$U_{Pn}(A,0,\emptyset) = -\frac{1}{\rho}\frac{\lambda N}{\rho+2\lambda N}\left(\frac{\phi_X}{\rho+\eta}N + \lambda N\frac{\phi_X}{\rho+\eta}p\log(1+N) + \lambda N(1-2p)\frac{1}{\rho N}(1-\phi_Y)K\right)$$
$$+\frac{1}{\rho N}(1-\phi_Y)AK.$$

Step 2: Optimal Choice of Contract Loadings

The principal maximizes its objective subject to the participation constraint, equation (8), which will bind in equilibrium. Let Ξ_n be the Lagrange multiplier on the participation constraint. Given the linearity of the principal's optimization program in ϕ_A, we can take the first-order necessary condition for the optimal choice of ϕ_A when only promising topics are pursued, recognizing the participation constraint binds and substituting with it,

$$0 = -\frac{1}{\rho N}AK + \frac{1}{\rho N}\frac{\lambda N(pq_1+(1-p)q_0)}{\rho+\lambda N(1+pq_1+(1-p)q_0)}\frac{\lambda N}{\rho}(1-2p_1)K$$
$$+\Xi_n\frac{\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\lambda N\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}}{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)}\partial_{\phi_A}\theta, \tag{A.63}$$

where

$$\partial_{\phi_A}\theta = -\gamma K\left(p_1 - p_0\right)\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}\left(pq_1+(1-p)\,q_0\right)\left(e^{\gamma\phi_A K}+2^{-\gamma\phi_X}e^{-\gamma\phi_A K}\right) + \gamma K\left(1-p_0\right)e^{\gamma\phi_A K},$$

and when all topics are pursued

$$-\frac{1}{\rho N}AK+\frac{1}{\rho}\frac{\lambda N}{\rho+2\lambda N}\lambda N\left(1-2p\right)\frac{1}{\rho N}K+\Xi_n\frac{\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\lambda N}{\rho+\eta+\lambda N\left(1-\omega\right)}\partial_{\phi_A}\omega=0, \tag{A.64}$$

where

$$\partial_{\phi_A}\omega = \gamma K\left(pq_1+(1-p)\,q_0\right)\frac{\lambda N\left(-p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+\left(1-p_1\right)e^{\gamma\phi_A K}\right)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}.$$

The optimal choice of $\phi_A$in both scenarios trades off the marginal present-value cost of rewarding researchers with additional output, which includes the expected growth in output from researching, with the relaxation of the participation constraint through raising researchers' expected lifetime utility.

Finally, we can take the first-order condition to find the optimal ϕ_X when only promising topics are pursued, recognizing the participation constraint binds and substituting with it,

$$\begin{aligned}0 = &-\frac{1}{\rho}\frac{\lambda N\left(pq_1+(1-p)\,q_0\right)}{\rho+\lambda N\left(1+pq_1+(1-p)\,q_0\right)}\left(\frac{1}{\rho+\eta}N+\frac{1}{\rho+\eta}\lambda N p_1\log\left(1+N\right)\right)\\ &+\Xi_n\frac{\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\lambda N}{\rho+\eta+\lambda N\left(1-\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\theta\right)}e^{\kappa}\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}\left(\partial_{\phi_X}\theta-\frac{\gamma N\theta}{\rho+\eta+\gamma\phi_X N+\lambda N}\right)\\ &+\Xi_n\frac{\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\frac{\lambda N}{\left(\rho+\eta+\gamma\phi_X N+\lambda N\right)^2}\lambda N}{1+\frac{\lambda N}{\rho+\eta+\gamma\phi_X N+\lambda N}},\end{aligned} \tag{A.65}$$

where Ξ_n is given by Equation A.63and

$$\partial_{\phi_X}\theta = -\left(p_1-p_0\right)\left(1+\frac{\gamma N\xi}{\left(\rho+\eta+\xi+\gamma\phi_X N+\lambda N\right)^2}\right)\left(pq_1+(1-p)\,q_0\right)\left(e^{\gamma\phi_A K}-2^{-\gamma\phi_X}e^{-\gamma\phi_A K}\right).$$

and when all topics are pursued

$$\begin{aligned}0 = &-\frac{1}{\rho}\frac{\lambda N}{\rho+2\lambda N}\left(\frac{1}{\rho+\eta}N+\lambda N\frac{1}{\rho+\eta}p\log\left(1+N\right)\right)\\ &+\Xi_n\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\frac{\frac{\lambda N(pq_1+(1-p)q_0)}{(\rho+\eta+\xi+\gamma\phi_X N+\lambda N)^2}}{1+\frac{\lambda N(pq_1+(1-p)q_0)}{\rho+\eta+\xi+\gamma\phi_X N+\lambda N}}+\Xi_n\frac{\frac{e^{-\gamma\phi_A A_t K}}{\rho+\eta}\lambda N}{\rho+\eta+\lambda N\left(1-\omega\right)}e^{\kappa}\partial_{\phi_X}\omega,\end{aligned} \tag{A.66}$$

where Ξ_n is given by Equation A.64and

$$\partial_{\phi_X}\omega = -\gamma N\left(pq_1 + (1-p)\,q_0\right)\frac{\lambda N\left(p_1 2^{-\gamma\phi_X}e^{-\gamma\phi_A K} + (1-p_1)\,e^{\gamma\phi_A K}\right) + \xi}{\left(\rho+\eta+\xi+\gamma\phi_X N+\lambda N\right)^2}.$$

Step 3: Minimizing Cost to the Principal

Step 2 provides the necessary conditions for the optimal contract loadings when the contract incentivizes researchers to research only promising topics and all topics, respectively. However, the question remains which equilibrium the principal would prefer.

For this, we recognize that given our parameter restrictions on $p_0$and p_1, the principal would not want to reward research in speculative topics if it is not necessary because such research has a negative net present value and the career incentives required to sustain it, i.e., a high ϕ_X, accrue a high cost in compensating researchers. Based on our results from Proposition 1, if researcher effort costs are sufficiently high, the principal must pay for all research; otherwise, it will restrict ϕ_X to be small enough as to not induce the researching of speculative topics.

Step 4: Avoiding Coordination Failure

Since researchers are ex-ante symmetric, either all of them exert effort to research or there is free-riding. For active researching to be a dominant strategy to free-riding, it must be the case that the expected utility to researching is higher. Notice that the only difference in compensation between active researchers and free-riders is that based on status, X_{it}, indexed by the contract loading ϕ_X.

From Proposition 4, the incentive to free-ride arises when researchers are compensated based only on output, i.e., $\phi_A > 0$ and $\phi_X = 0$. Notice this does not depend on the value of ϕ_A, but reflects that when individual effort is not rewarded, it need not be incentivized. Consequently, to avoid a coordination failure, $\phi_X > 0$.

Because researchers are awarded for self-citations, there exists a $\underline{\phi}_X$ such that if $\phi_X \geq \underline{\phi}_X$, then zero active researchers is not an equilibrium, i.e., one researcher will find it incentive compatible to research alone. However, if one researcher finds it optimal to research alone, then a second researcher must find it optimal to research with the first because this increases both citations and the likelihood of arrival and maturation of topics. By inductive reasoning, the only equilibrium is all N researchers actively research.

Notice that $\underline{\phi}_X$ is higher than the minimum ϕ_X required to incentivize all N researchers to work

together instead of zero. If the effort cost, κ, is sufficiently high, then the principal induces researchers to research all topics, even if researching only promising topics is optimal in the absence of the potential for a coordination failure.

B Appendix B: Extension with Continuous Learning

In this online appendix, we consider how continually learning about an active topic's significance over time impacts researcher behavior. We characterize a more realistic version of our model by assuming that signals about the viability of a topic constantly arrive over time. As a topic is researched, researchers are likely to learn more about whether it will turn out to be significant. This change gives rise to a time-varying probability of the topic being significant. In this setting, all researchers research an active topic until the probability it is significant falls below some critical threshold, and this threshold is decreasing in the importance of career concerns, ϕ_X.

Suppose now that there are no signals at inception about whether a topic is likely significant. Instead, all topics have an initial prior probability $p_0 = .5$ of being significant. As researchers work on the topic, however, they receive signals about whether the topic is likely to be significant. Specifically, at each instant, the new projects in topic j of field n generate a signal S^n_{jt} about the topic's significance π^n_j as long as it is active

$$dS^n_{jt} = \pi^n_j dt + \sigma dZ^n_{jt} + \left(\pi^n_j - S^n_{jt-}\right) dB^n_t, \tag{B.1}$$

where Z^n_{jt} is a standard Wiener process independent of other $Z^{n'}_{j't}$ and σ is the diffusion of the signal. It is immediate that the common knowledge posterior about the quality of the current topic in field n given $S^n_{jk_{k\in[t_j,t]}} \in \mathcal{F}^c_t$ for $t \geq t_j$ is summarized by $p_{nt} = \mathbb{E}\left[\mathbf{1}_{\pi^n_j=1} \mid \mathcal{F}^c_t\right]$ and is given by the Wonham Filter

$$dp_{nt} = \frac{p_{nt}\left(1-p_{nt}\right)}{\sigma} d\hat{Z}^n_{jt} + \left(\pi^n_j - p_{nt-}\right)\left(dB^n_t - \lambda m^n_t dt\right), \tag{B.2}$$

where $d\hat{Z}^n_{jt} = dZ^n_{jt} + \frac{1}{\sigma}\left(\pi^n_j - p_{nt}\right) dt$ is a $\mathcal{F}^c_t$-standard Wiener process by Girsanov's Theorem, and $dB^n_t - \lambda \mu^n_t dt$ is a martingale jump process that reveals the true π^n_j once the topic matures.

Based on our analysis in the baseline model, it is clear that if there is an active topic, all researchers will focus on working on it. However, there is now the possibility that they may abandon an active topic to investigate a new one if the likelihood that it is significant, p_{nt}, is sufficiently low. Because there will be at most one topic on which researchers are working at a given instant, we drop the n subscript from p_{nt}, and just refer to the probability the current active topic is significant as p_t. Our key result is that the threshold probability p^* at which a topic is abandoned is decreasing in the reward for career advancement, ϕ_X. This is summarized in the following proposition.

Proposition 6. *All researchers research any active field as long the probability it is significant, p_t,*

exceeds a threshold p^*

$$p^* = \frac{e^{\gamma\phi_A K} - \frac{1+\gamma\phi_X/\lambda-1/N}{\rho+\eta+\lambda}e^{\kappa} - \frac{\lambda}{\rho+\eta+\lambda}\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\gamma\phi_X N+2\lambda N}\left(1+\frac{1}{2}\left(2^{-\gamma\phi_X}e^{-\gamma\phi_A K}+e^{\gamma\phi_A K}\right)+\frac{\rho+\eta}{\lambda N}\right)}{e^{\gamma\phi_A K}-2^{-\gamma\phi_X}e^{-\gamma\phi_A K}}. \tag{B.3}$$

This threshold is decreasing in the reward for career advancement, $\gamma\phi_X$.

Proof of Proposition 6:

In what follows, we assume the effort cost, κ, is sufficiently small It is immediate from our analysis in Proposition 1 that researchers will all research an active field before they all investigate a new one provided that the probability it is significant is sufficiently high. Consequently, to prove our claim, we focus on the case in which there is one active topic with current probability p_t of being significant, and researchers must choose whether to research it or investigate a new topic.

Let us conjecture that researcher n's value function takes the form

$$V_{nT_{i-1}^n}\left(t, X_{nt}, A_t, p_t\right) = -e^{-(\rho+\eta)\left(t-T_{i-1}^n\right)-v_a A_t - v_n X_{nt}} f_n\left(p_t\right),$$

with the convention that $p_t = 0$ if there is no active topic (i.e., $t = 0$) or the researchers abandon an active topic.

Let r_t^n be the indicator whether researcher n researchers the active topic and e_t^n the indicator whether he investigates a new topic. Factoring out the $e^{-(\rho+\eta)\left(t-T_{i-1}^n\right)-v_a A_t - v_n X_{nt}}$ from the value function, the Hamilton-Jacobi-Bellman Equation for researcher n is then

$$\begin{aligned} 0 \geq \sup_{r^n, e^n, (r_t^n+e_t^n)'\iota_L \leq 1 \ \forall\ t} \Bigg\{ & - e^{v_a A_t + v_n X_{nt} - \gamma\phi_A A_t K - \gamma\phi_X X_{nt} + \kappa(r_t^n+e_t^n)'\iota_L} + v_n f_n\left(p_t\right)\sum_l \left(R_{-n,lt}+r_{lt}^n\right) r_{lt}^n \\ & - \lambda\left(R_{-n,t}+r_t^n\right)\left(\left(p_t\left(1+r_t^n\right)^{-v_n}e^{-v_a}+\left(1-p_t\right)e^{v_a}\right) f_n\left(0\right) - f_n\left(p_t\right)\right) - \frac{1}{2}f_n''\left(p_t\right)\left(\frac{p_{nt}\left(1-p_{nt}\right)}{\sigma}\right)^2 \\ & - \sum_{l\in\mathcal{E}_t}\lambda\left(E_{-n,lt}+e_{lt}^n\right)\left(f_n\left(\frac{1}{2}\right) - f_n\left(p_t\right)\right) + \left(\rho+\eta\right) f_n\left(p_t\right)\Bigg\}. \end{aligned} \tag{B.4}$$

It is immediate that $v_n = \gamma\phi_X$ and $v_a = \gamma\phi_A K$, confirming the conjecture. The Hamilton-Jacobi-

Bellman Equation (B.4) consequently reduces to

$$0 \geq \sup_{r^n, e^n, (r_t^n + e_t^n)' \iota_L \leq 1 \ \forall \ t} \Bigg\{ - e^{\kappa (r_t^n + e_t^n)' \iota_L} + \gamma \phi_X f_n \left(p_t \right) \sum_l \left(R_{-n,lt} + r_{lt}^n \right) r_{lt}^n - \frac{1}{2} f_n'' \left(p_t \right) \left(\frac{p_{nt} \left(1 - p_{nt} \right)}{\sigma} \right)^2$$
$$- \lambda \left(R_{-n,t} + r_t^n \right) \left(\left(p_t \left(1 + r_t^n \right)^{-\gamma \phi_X} e^{-\gamma \phi_A K} + \left(1 - p_t \right) e^{\gamma \phi_A K} \right) f_n \left(0 \right) - f_n \left(p_t \right) \right)$$
$$- \sum_{l \in \mathcal{E}_t} \lambda \left(E_{-n,lt} + e_{lt}^n \right) \left(f_n \left(\frac{1}{2} \right) - f_n \left(p_t \right) \right) + \left(\rho + \eta \right) f_n \left(p_t \right) \Bigg\}. \tag{B.5}$$

We now compare the value to the researcher of investigating a current topic along with the other $N - 1$ researchers versus abandoning it to research a new one by himself. Both choices incur the effort cost κ, and therefore it is irrelevant. Similarly, new information about the current topic will arrive regardless, so it is irrelevant as well. As shown in the proof of Proposition 1, researchers are indifferent to investigating in the same or in different fields. We assume they all investigate in the same field without loss.

Researcher n will consequently pursue the active topic if

$$\left(\gamma \phi_X N + \lambda \left(N - 1 \right) \right) f_n \left(p_t \right) - \lambda N \left(p_t 2^{-\gamma \phi_X} e^{-\gamma \phi_A K} + \left(1 - p_t \right) e^{\gamma \phi_A K} \right) f_n \left(0 \right) \geq -\lambda f \left(\frac{1}{2} \right). \tag{B.6}$$

It follows from equation (B.6) that there is a critical p_t, p^*, at which an active topic will be abandoned if $p_t \leq p^*$.

Consequently, assuming conducting research to not is always optimal, we have two regions for the Hamilton-Jacobi-Bellman Equation. For $p_t \geq p^*$, the Hamilton-Jacobi-Bellman Equation from (B.5) in the research region is

$$\left(\rho + \eta + \gamma \phi_X N + \lambda N \right) f_n \left(p_t \right) = e^{\kappa} + \lambda N \left(p_t 2^{-\gamma \phi_X} e^{-\gamma \phi_A K} + \left(1 - p_t \right) e^{\gamma \phi_A K} \right) f_n \left(0 \right)$$
$$+ \frac{1}{2} f_n'' \left(p_t \right) \left(\frac{p_{nt} \left(1 - p_{nt} \right)}{\sigma} \right)^2, \tag{B.7}$$

and in the investigate region when $p_t < p^*$, which we designate as $p_t = 0$, is

$$f_n \left(0 \right) = \frac{e^{\kappa} + \lambda N f_n \left(\frac{1}{2} \right)}{\rho + \eta + \lambda N}. \tag{B.8}$$

Substituting equation (B.8) into (B.7), we arrive at

$$\left(\rho+\eta+\gamma\phi_X N+\lambda N\right) f_n\left(p_t\right)=e^{\kappa}+\lambda N\left(p_t 2^{-\gamma\phi_X} e^{-\gamma\phi_A K}+\left(1-p_t\right) e^{\gamma\phi_A K}\right) \frac{e^{\kappa}+\lambda N f_n\left(\frac{1}{2}\right)}{\rho+\eta+\lambda N} + \frac{1}{2} f_n''\left(p_t\right)\left(\frac{p_{nt}\left(1-p_{nt}\right)}{\sigma}\right)^2. \tag{B.9}$$

Notice that $f_n\left(\frac{1}{2}\right)$ in equation is a constant in (B.9).

Let us conjecture that $f_n\left(p_t\right)$ is linear in p_t, or

$$f_n\left(p_t\right)=a+bp_t.$$

Substituting this conjecture into equation (B.9)for $p_t=\frac{1}{2}$, we have that

$$f_n\left(\frac{1}{2}\right)=\frac{1+\frac{1}{2}\left(2^{-\gamma\phi_X} e^{-\gamma\phi_A K}+e^{\gamma\phi_A K}\right)\frac{\lambda N}{\rho+\eta+\lambda N}}{\rho+\eta+\gamma\phi_X N+\lambda N-\frac{1}{2}\lambda N\left(2^{-\gamma\phi_X} e^{-\gamma\phi_A K}+e^{\gamma\phi_A K}\right)\frac{\lambda N}{\rho+\eta+\lambda N}} e^{\kappa},$$

and

$$a=\frac{e^{\kappa}+\lambda N e^{\gamma\phi_A K}}{\rho+\eta+\gamma\phi_X N+\lambda N}\frac{e^{\kappa}+\lambda N f_n\left(\frac{1}{2}\right)}{\rho+\eta+\lambda N},$$
$$b=\frac{\lambda N\left(2^{-\gamma\phi_X} e^{-\gamma\phi_A K}-e^{\gamma\phi_A K}\right)}{\rho+\eta+\gamma\phi_X N+\lambda N}\frac{e^{\kappa}+\lambda N f_n\left(\frac{1}{2}\right)}{\rho+\eta+\lambda N},$$

confirming the conjecture. Consequently, we have for $p_t \geq p^*$ that

$$f_n\left(p_t\right)=\frac{1}{\rho+\eta+\gamma\phi_X N+\lambda N}\frac{e^{\kappa}+\lambda N f_n\left(\frac{1}{2}\right)}{\rho+\eta+\lambda N}\left(e^{\kappa}+\lambda N e^{\gamma\phi_A K}+\lambda N\left(2^{-\gamma\phi_X} e^{-\gamma\phi_A K}-e^{\gamma\phi_A K}\right) p_t\right) \tag{B.10}$$

Substituting our expression for $f_n\left(p_t\right)$ from equation (B.10) and (B.8) into condition (B.6) at the critical p^*, i.e., when it holds with equality, we find

$$\left(\rho+\eta+\lambda\right) f_n\left(p^*\right)=\lambda f\left(\frac{1}{2}\right)+\frac{e^{\kappa}+\lambda N f_n\left(\frac{1}{2}\right)}{\rho+\eta+\lambda N} e^{\kappa} \tag{B.11}$$

from which follows that

$$p^*=\frac{e^{\gamma\phi_A K}-\frac{1+\gamma\phi_X/\lambda-1/N}{\rho+\eta+\lambda} e^{\kappa}-\frac{\lambda}{\rho+\eta+\lambda}\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\gamma\phi_X N+2\lambda N}\left(1+\frac{1}{2}\left(2^{-\gamma\phi_X} e^{-\gamma\phi_A K}+e^{\gamma\phi_A K}\right)+\frac{\rho+\eta}{\lambda N}\right)}{e^{\gamma\phi_A K}-2^{-\gamma\phi_X} e^{-\gamma\phi_A K}}. \tag{B.12}$$

It is immediate from rewriting p^* from equation (B.12) as

$$p^* = \frac{e^{\gamma\phi_A K} - \frac{1+\gamma\phi_X/\lambda-1/N}{\rho+\eta+\lambda}e^{\kappa} - \frac{\lambda}{\rho+\eta+\lambda}\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\gamma\phi_X N+2\lambda N}\left(1+e^{\gamma\phi_A K}+\frac{\rho+\eta}{\lambda N}\right)}{e^{\gamma\phi_A K} - 2^{-\gamma\phi_X}e^{-\gamma\phi_A K}},$$
$$+\frac{\lambda}{\rho+\eta+\lambda}\frac{\rho+\eta+\gamma\phi_X N+\lambda N}{\rho+\eta+\gamma\phi_X N+2\lambda N}$$

that p^* is decreasing in ϕ_X.

C Appendix C: Alternative Measure of Scientific Importance

In this appendix, we replicate the main analyses using an alternative proxy for the scientific importance of a gene. Whereas the baseline measure relies on atheoretical genome-wide association studies (GWAS) to capture the number of diseases associated with each gene, here we instead employ data on differential gene expression. Differential expression refers to systematic differences in the level of expression of a gene between diseased and healthy tissues. Because gene expression regulates when, where, and how much a gene product is produced, changes in expression levels provide molecular evidence of a gene's potential biological relevance. This evidence is used in biomedical research to identify candidate biomarkers, therapeutic targets, and gene signatures for diagnostics (Rodriguez-Esteban and Jiang, 2017; Richardson et al., 2024).

The probability of disease-related expression for each human gene is obtained from Northwestern University's *Find My Understudied Genes* (FMUG) database (`https://fmug.amaral.northwestern.edu/`). This probability captures the average likelihood that a gene is expressed in human disease contexts, based on pooled evidence from high-throughput transcriptomic studies. Differential expression data have been widely used to prioritize genetic targets in pharmaceutical and clinical research, even though not all expression changes translate into consequential biological activity (Stoeger et al., 2018). Importantly, the FMUG measure offers a complementary dimension of biological evidence, independent of publication activity, making it suitable as an alternative indicator of ex ante scientific importance.

Using this proxy yields results that are consistent with the main analysis. Panel (a) of Figure C1 plots the probability of disease expression for each gene, ordered by the cumulative number of publications. The figure shows that the genes most frequently studied in the literature are not necessarily those with the highest probability of differential expression in human disease. Panel (b) of Figure C1 replicates the main binscatter analysis, underscoring that papers focusing on genes with higher disease expression are only weakly associated with higher citation counts. By contrast, the strongest predictor of citations remains whether the gene is already heavily studied, reinforcing the evidence of a systematic citation penalty for less-studied genes.

Table C1 presents regression estimates that replicate the main specifications. Across all models, papers on less-studied genes continue to receive significantly fewer citations, even after controlling for the gene's probability of differential expression. The coefficient on scientific importance is positive and significant, but the magnitude of the citation penalty for understudied genes remains

large. These findings confirm that the misalignment between citation incentives and scientific importance is not specific to the GWAS-based measure but is robust to an alternative proxy rooted in gene expression biology. Taken together, the results in this appendix strengthen the conclusion that the observed citation penalty does not reflect underlying differences in biological importance. Instead, it reflects the crowding of attention onto a narrow set of already popular genes, consistent with the mechanism highlighted in the main text.

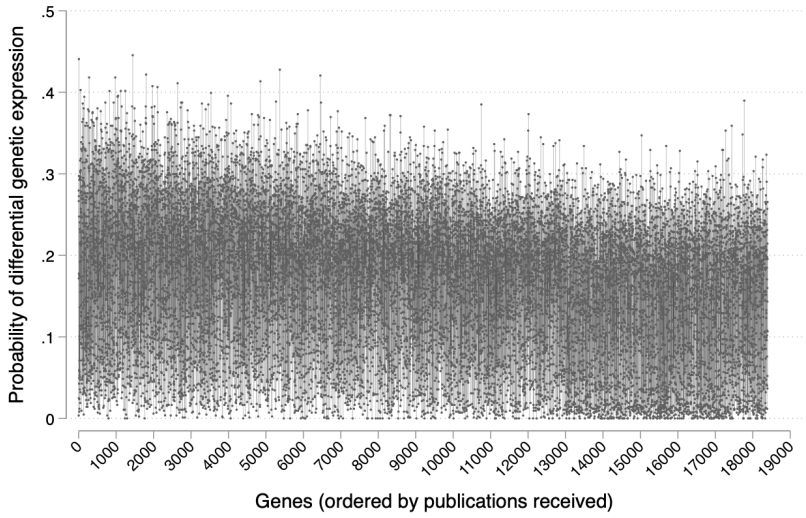
Table C1: Citation Penalty to Papers Focusing on Less Studied Genes.

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Inverse Study Rank	-.0121***	-.00919***		
	(.000537)	(.00087)		
Understudied (0/1)			-.262***	-.156***
			(.0161)	(.0257)
Scientific Importance	.401***	.358**	.465***	.409***
	(.0733)	(.111)	(.0732)	(.111)
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	Yes	No	Yes
N	790,650	604,906	790,650	604,906

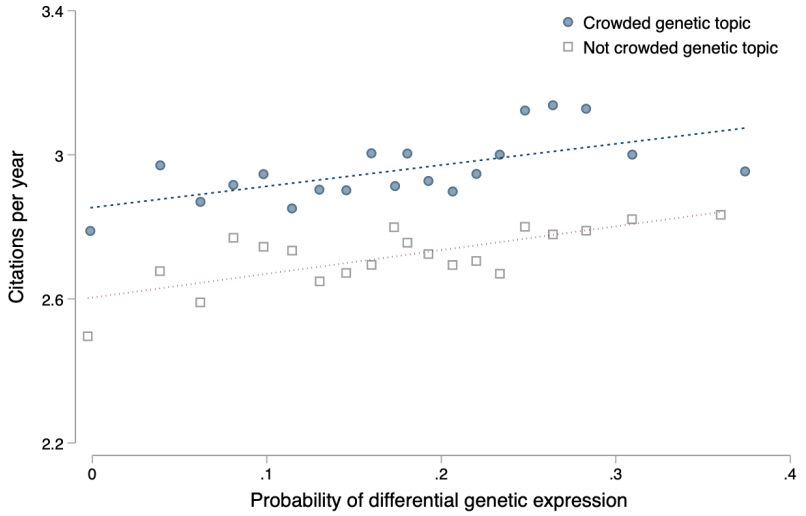
Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Citations*: average yearly scientific citations received by the publication; *Inverse Study Rank*: percentile rank of the gene by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied*: 0/1 = 1 for protein-coding human genes with a below-median number of publications; *Scientific Importance*: probability of expression of the gene in human diseases, using data from Northwestern University's Find My Understudied Genes (FMUG); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. See text for details.

Figure C1: Robustness of the Main Analyses Using Differential Disease Expression of Genes as an Alternative Metric of Scientific Importance.

(a) *Probability of Disease Expression for Each Gene*



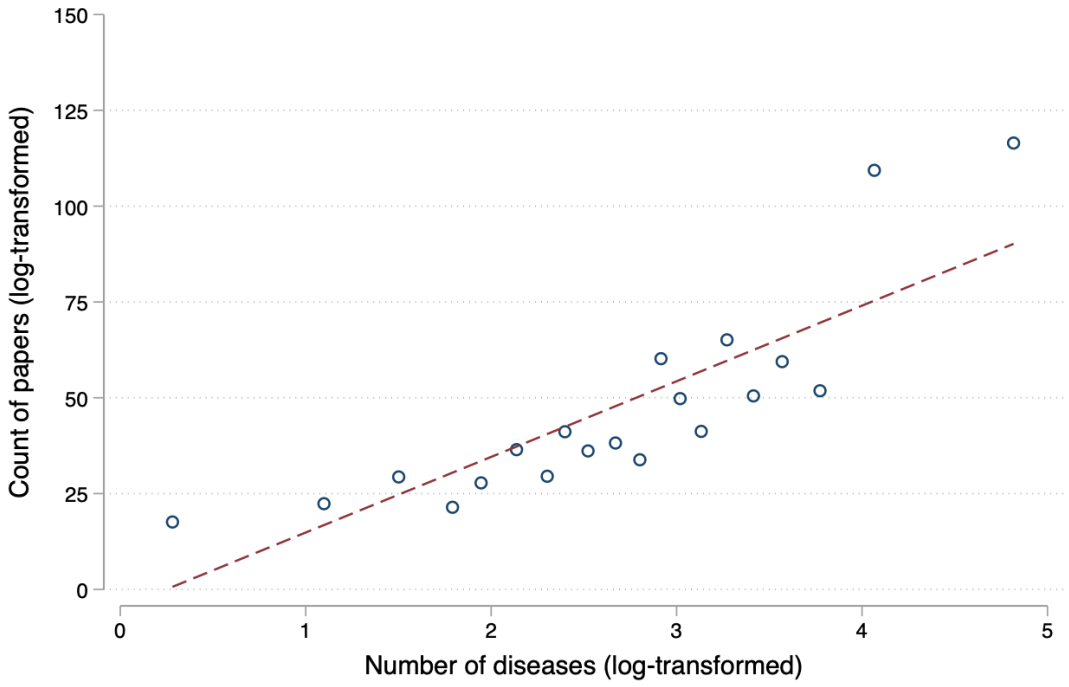
(b) *Relationship Between a Paper's Citations and the Disease Expression of the Gene Studied*



Note: The figure replicates the main results using an alternative proxy of scientific importance. Panel (a) shows the probability that a gene is expressed in a human disease on the Y axis, with genes on the X axis sorted from the most to the least studied. Genes on the X axis are sorted like in Figure 1. Panel (b) plots the relationship between yearly citations received by a publication and the biological importance of the gene it studies, proxied by the probability of being expressed in a human disease. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we residualize yearly citations and biological importance with respect to an indicator for each journal-year bin. We divide the sample into 20 equal-sized groups based on the ventiles of the biological importance measure and plot the mean of yearly cites against the mean of importance in each bin. The sample is the full analysis sample as defined in the text. See text for details.

D Additional Analysis and Robustness Checks

Figure D1: Genes Presenting Mutations Associated with More Human Diseases Receive More Publications on Average.



Note: This figure plots the relationship between the total number of papers received by a publication and the biological importance of the gene it studies. The importance of a gene is proxied by the number of diseases associated with it in unbiased GWAS studies. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we residualize yearly citations and biological importance with respect to an indicator for each journal-year bin. We divide the sample into 20 equal-sized groups based on the ventiles of the biological importance measure and plot the mean of yearly cites against the mean of importance in each bin. The sample is the full analysis sample as defined in the text. See text for details.

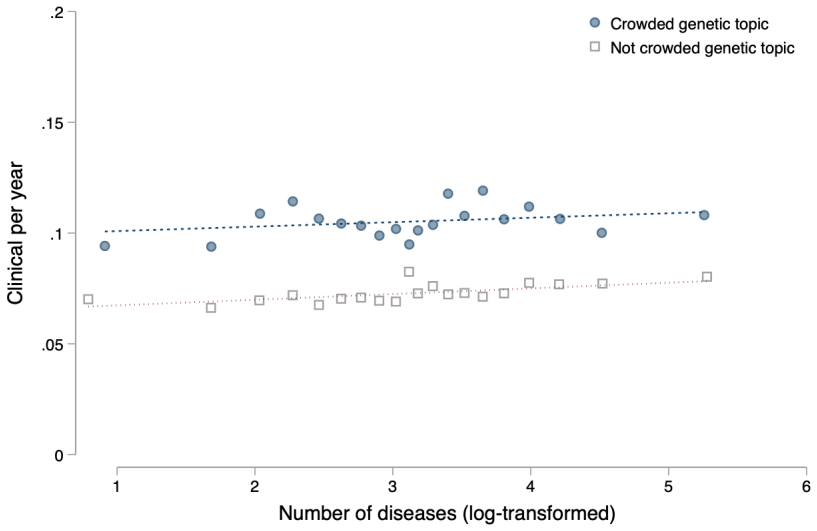
Table D1: Alternative Definitions of Senior Authors.

Seniority definition:	(1) Inverse Study Rank $\geq$ 8 years	(2) Inverse Study Rank $\geq$ 9 years	(3) Inverse Study Rank $\geq$ 10 years	(4) Inverse Study Rank $\geq$ 11 years
Senior Author (0/1)	0.299752*	0.320932*	0.231893†	0.195456†
	(0.147457)	(0.130743)	(0.118351)	(0.107484)
Scientific Importance	Yes	Yes	Yes	Yes
Journal Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	Yes	Yes	Yes	Yes
N	604906	604906	604906	604906

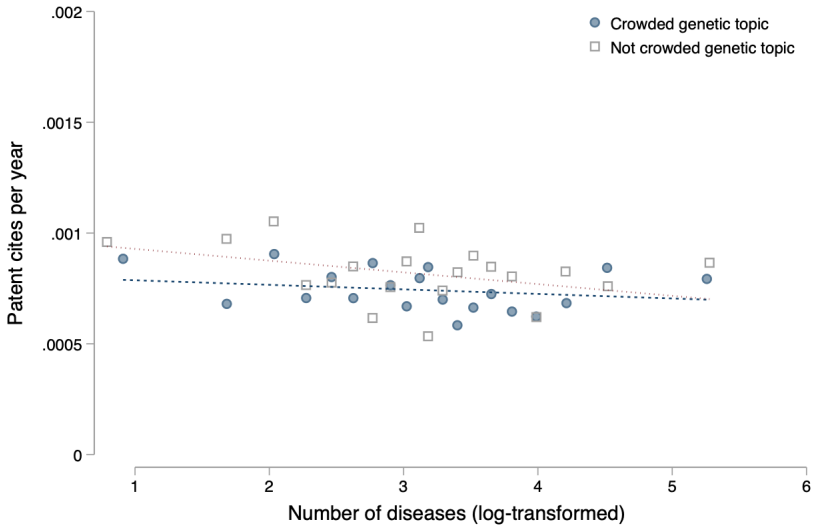
Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level. Robust std. err. in parentheses. *Inverse Study Rank*: percentile rank of the gene by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Senior Author*: 0/1 = 1 if the last author of the publication has been active in publishing for at least a certain number of years (as indicated in the column heading); *Scientific Importance*: count of diseases linked to mutations in the gene, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. See text for details.

Figure D2: Clinical Research Prioritizes Well-Studied Genes, While Corporate Innovation Does Not.

(a) *Clinical Citations to a Publication and Importance of the Gene Studied*



(b) *Patent Citations to a Publication and Importance of the Gene Studied*



Note: This figure plots the relationship between yearly citations from clinical studies (panel (a)) and USPTO patents (panel (b)) received by a publication and the biological importance of the gene it studies. The importance of a gene is proxied by the number of diseases associated with it in unbiased GWAS studies. The plot is presented as a binned scatterplot. To construct this binned scatterplot, we residualize yearly citations and biological importance with respect to an indicator for each journal-year bin. We divide the sample into 20 equal-sized groups based on the ventiles of the biological importance measure and plot the mean of yearly cites against the mean of importance in each bin. The sample is the full analysis sample as defined in the text. See text for details.

Table D2: Robustness to the Inclusion of Publications Focusing on Multiple Genes.

Panel A: Main Regression

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Inverse Study Rank (average)	-.00958***	-.00956***		
	(.000673)	(.000679)		
Understudied (share)			-.177***	-.175***
			(.0203)	(.0203)
Scientific Importance (average)		.0000486		.0000889
		(.000149)		(.000148)
Journal Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	Yes	Yes	Yes	Yes
N	1036692	1036692	1036692	1036692

Panel B: Instrumental Variable

	(1) Inverse Study Rank	(2) Citations	(3) Understudied (0/1)	(4) Citations
Mouse Homolog (share)	-.808***		-.0369***	
	(.0696)		(.00223)	
Inverse Study Rank (average)		-.149***		
		(.0363)		
Understudied (share)				-3.26***
				(.772)
F-Statistic (First Stage)	134.892		274.487	
Scientific Importance (average)	Yes	Yes	Yes	Yes
Journal Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	Yes	Yes	Yes	Yes
N	1,036,692	1,036,692	1,036,692	1,036,692

Note: †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional OLS regressions at the publication level including also articles focusing on multiple genes. Robust std. err. in parentheses. *Citations*: average yearly scientific citations received by the publication; *Inverse Study Rank (average)*: average percentile rank of the genes studied by amount of prior research, reversed so that 100 = least studied and 0 = most studied; *Understudied (share)*: share of genes studied with a below-median number of publications; *Mouse Homolog (share)*: share of gene studied with a homolog gene in the mouse, which allows them to be studied using the laboratory mouse; *Scientific Importance (average)*: average count of diseases linked to mutations in the genes studied, as identified by unbiased genome-wide association studies (GWAS); *Journal-Year FE*: fixed effect for articles published in a scientific journal in a given year; *Disease Class FE*: fixed effect for disease codes from the MeSH tree; *Principal Investigator FE*: fixed effect for the last author of the article, usually denoting the PI of the project. Panel B reports the Kleibergen-Paap F statistic for the first-stage regression. See text for details.