Innovation Under Resource Constraints: Supercomputing in Scientific Research

Justine Boudou* and John McKeon+

March 18, 2025

Abstract

Resource constraints play a critical role in shaping both the rate and direction of innovation. In this paper, we study how constraints on a key input to frontier innovation, high-performance computing power, impact innovative output. Our empirical setting is an NSF-funded initiative called XSEDE, which allows researchers to request supercomputing resources for their research through a distributed network of supercomputers across the United States. Leveraging capacity shortages, we are able to isolate the impact of computing resource constraints on scientific production. We find that an increase in the amount of computing resources allocated to a researcher relative to what they require (i.e., a relaxing of constraints) leads to an increase in the number of papers published and alters their research direction. Specifically, additional computing resources lead researchers to study less popular and newer topics, explore topics that they have not studied in their prior work, and broaden the scope of their projects. Our findings underscore the intricate relationship between resources and innovative output, with consequences for optimizing resource allocation strategies and fostering innovation.

Keywords: *Innovation*, *Science*, *Resource Allocation*, *Constraints*, *Computing Resources*

Authors contributed equally, listed alphabetically. *Boudou: Harvard Business School (jboudou@hbs.edu) +McKeon: Boston University (jmckeon@bu.edu). The authors thank Stephen Deems and Dave Hart for data access. We are indebted to Kevin Boudreau, Shane Greenstein, Josh Krieger, Megan MacGarvie, Raviv Muricano-Goroff, Kyle Myers, Felix Poege, Maria Roche, Tim Simcoe, Jeremy Watson, and Rosemarie Ziedonis for feedback on early drafts. We thank the participants of the BU Strategy & Innovation Brown Bag, TPRI seminar series, NBER Productivity Lunch, RISE Workshop, and 2024 SMS Conference. All errors and omissions are our own. **No sources of outside research support or potential conflicts.**

1 Introduction

Resource constraints exert a significant influence on the rate and direction of innovation (Arrow 1962, Nelson 1959, Barney 1991, Agrawal et al. 2018). Constraints can impact firms' risk-taking (Krieger et al. 2022) and R&D investments (Mezzanotti & Simcoe 2023), along with investors' and firms' tolerance for failure (Nanda & Rhodes-Kropf 2017, Atanassov 2016). Prior work has shown that *access* to new resources, such as technology and data, impacts innovation by shaping researchers' effort allocation, collaboration patterns, and the types of projects pursued (Furman & Teodoridis 2020, Ding et al. 2010, Nagaraj & Tranchero 2023). While the relationship between resource access and innovation outcomes is well-studied, we still know little about how variation in the *amount* of scarce resources allocated impacts both the rate and especially the direction of innovation – particularly when it comes to the ability of researchers to expand the boundaries of their research and push the knowledge frontier. This question is crucial for policymakers and R&D managers across a range of organizations – from national labs and large R&D-intensive firms to startups – who face critical decisions about how much to invest in acquiring innovative resources and how to allocate these scarce resources among competing projects, thus directly shaping the constraints researchers face (Babina et al. 2023, Boudreau et al. 2016, Blomfield & Vakili 2023). These decisions are particularly important when they involve resources that may enable researchers to pursue cutting-edge, less conventional work. This has been difficult to study because variation in resource constraints is usually difficult to observe, coincides with variation in other resources, or is subject to a selection process where more promising projects receive more resources.

This paper offers a novel empirical setting that allows us to study how constraints on a critical input to innovation, supercomputing power, impact the innovative output of scientific researchers. In our setting, researchers submit applications for a specific quantity of supercomputing resources to support projects tackling complex, computationally intensive problems. These requests are evaluated and then reduced based on the excess overall demand relative to the available supply at the time. This creates variation in the extent to which researchers are constrained, so that we can estimate the impact of resource constraints on the rate and direction of research output. Importantly, our data enables us to observe the quantity of resources a researcher objectively requires for their project, allowing us to account for the underlying project's scale and potential, a feature which is rarely observable in other settings.

Beyond these empirical advantages, studying supercomputing (or high-performance computing,

"HPC") resources is crucial given that they are a key input to modern innovation and the production of frontier knowledge (Thompson et al. 2022, Kim 2023, Tranchero 2023). Across many fields, high-performance computing resources have become indispensable for firms' competitive advantage and for researchers striving to push the boundaries of knowledge. In the field of artificial intelligence for instance, firms increasingly rely on massive supercomputing clusters to support the training of their prediction models.[1] Advanced computational resources also facilitate the screening of potential drug candidates and the simulation of complex weather patterns, accelerating the development of new therapies and aiding in disaster preparedness. Despite its widespread importance, supercomputing power remains a scarce resource for many innovators (Lohr 2019). In light of this, recent policy discussions in the U.S. and the E.U. emphasize the importance of additional investment in HPC resources for national competitiveness, firms' competitive advantage in the global market, and national security.[2] These inequalities in resource availability underline the importance of understanding innovators' choices when constraints on their computing power vary.

We first present a conceptual framework to motivate our empirical analysis, which considers the impact of supercomputing resource constraints on the rate and characteristics of researchers' output. Intuitively, one might expect that as resource constraints are relaxed, researchers will produce additional output. However, the effect of increased high-performance computing resources on the choice of research topic - and whether researchers shift towards more innovative work - is less clear. For instance, if researchers do not act strategically based on these constraints, there would be no change in a given researcher's choice of topic. On the other hand, given the heterogeneous nature of innovative projects and the incentives researchers face, researchers may prioritize certain projects over others as resource constraints are relaxed, thereby altering the characteristics of the output produced (Azoulay et al. 2011). Thus, how computing constraints shape the direction of innovation and in particular researchers' ability to pursue different research trajectories is an open question. As suggested by prior literature, we consider how variations in supercomputing resources may affect researchers' choice to pursue less popular topics, research lines they haven't worked on before, and projects broader in scope.

[1] For example, the "Stargate" project led by Arm, Microsoft, NVIDIA, Oracle, and OpenAI, involves investments of $500 billion which include the construction of a new supercomputer. See https://fortune.com/2024/04/02/microsoft-openai-stargate-100-billion-ai-supercomputer-star-wars-sdi/

[2] See report from the NSA-DOE https://www.nitrd.gov/nitrdgroups/images/b/b4/nsa_doe_hpc_techmeetingreport.pdf, the think-thank ITIF https://www2.itif.org/2016-high-performance-computing.pdf, and the European Investment Bank https://www.eib.org/attachments/pj/financing_the_future_of_supercomputing_en.pdf

We empirically examine how variations in supercomputing resource constraints affect researchers' innovative output in the novel context of an NSF-funded initiative known as *The Extreme Science and Engineering Discovery Environment* (henceforth "XSEDE"). XSEDE allows researchers to apply each quarter for access to a distributed network of high-performance computing resources across the United States. XSEDE supported over $6 billion of funded research over a ten-year period, playing a pivotal role in enabling researchers to tackle computationally intensive scientific challenges that were previously beyond reach.[3] Our dataset combines internal XSEDE data on 710 projects and details about their computing allocations with information on the publication output associated with each project.

In order to examine how constraints on supercomputing resources impact researchers' output, we exploit the allocation process of XSEDE which provides us with plausibly exogenous variation in the degree to which researchers are constrained through the amount of computing resources they receive. Researchers seeking computing allocations on XSEDE supercomputers first request an allocation for their project, which is assessed by a committee. The committee provides a recommended allocation based on the amount of computing resources the project *would need* if XSEDE resources were unconstrained. Thus, we are able to observe the amount of resources a project objectively requires. Throughout the period we study, demand for computing resources consistently exceeds supply. Hence, on a quarterly basis, the aggregate computing demand is then reconciled with available supply by reducing each recommendation based on an algebraic formula and allocating researchers this reduced amount. Our empirical framework leverages this unique resource allocation process of XSEDE in order to create variation in the quantity of computing resources received by researchers relative to what they would need, hence how constrained they are.

We find that increased resource allocation (i.e., a relaxing of constraints) leads to a large increase in the number of papers published within a project. Specifically, a 10 percentage point increase in allocation leads to an 11% increase in the number of papers written. Our results also show that resource constraints impact the direction of innovation. On average, we find that as resource constraints are relaxed, researchers tend to explore less studied (i.e., "unconventional") research avenues, as measured by the popularity among other researchers of the research topics they chose to study. In addition, researchers receiving more computing resources tend to broaden the scope of

[3]See NSF Awards 1053575 and 1548562.

their research, pursuing projects that span a wider array of topics. Further analysis shows that the observed changes are driven respectively by the *least* conventional and the *broadest* papers within a researcher's output, indicating significant shifts at the boundaries of researchers' work. We also find that researchers explore new research streams among their *most* exploratory work, studying topics they have not previously studied. In addition, among the papers addressing the *newest* topics, researchers increasingly focus on topics that emerged more recently. These results suggest that the relaxing of constraints on resources enables researchers to push the boundaries of their work, advancing into different territories. They do this by studying newer topics that have not yet gained traction in the scientific community, working in areas they have not previously worked on, and creating broader papers. We also find evidence that the relaxing of constraints leads to a decrease in the number of citations but has no impact on journal quality, suggesting the papers *may* struggle to accrue citations due to the topics they study rather than their inherent quality.

Taken together, our study sheds light on the intricate relationship between resource allocation and innovative output, underscoring the importance of effective resource allocation strategies among managers and policymakers in promoting innovation. We emphasize the importance of understanding the changes in the attributes of output produced when the amount of resources varies. At a high level, our findings underscore the importance of considering not only the rate but also the direction of research output when evaluating resource allocation strategies at R&D intensive firms, innovative startups, or publicly funded programs that allocate scarce resources across projects. In particular, when decision-makers are allocating scarce resources across competing project lines, a key consideration should be the potential benefits from the less conventional, or "more exploratory" work that may emerge from each project as a result of increased resource allocation.

This paper makes three distinct contributions. First, we contribute to the literature studying resource constraints and innovation by shedding light on how constraints on non-financial resources impact the direction of innovative output. Much of the literature focuses on the effect of funding and financial constraints on firms' innovation output (Hoegl et al. 2008), showing for example that financing impacts firms' quantity and quality of innovations (Atanassov 2016), the type of firms that investors are willing to fund (Nanda & Rhodes-Kropf 2017, Conti et al. 2019) as well as the composition of their R&D investments (Mezzanotti & Simcoe 2023). While financial resources often allow for substitution across funding streams and managerial discretion in allocating inputs, our study focuses on a less fungible resource that directly shapes the observable output. We pay

particular attention to research direction and the boundaries of researchers' work (e.g., Furman & Teodoridis 2020, Baruffaldi & Gaessler 2021) and complement work on other non-fungible constraints such as geography (Singh & Marx 2013, Catalini 2018), policy (Glennon 2023, Conti 2018), environmental regulation (Berrone et al. 2013) or even luck (Hill 2019) on innovation, as well as the role of other scarce inputs such as attention (Fini et al. 2022).

Second, this paper contributes to our understanding of the *intensive margin* of resource allocation. Similar to prior work that studies the returns of R&D subsidies directly on firms' patenting and revenue (Bronzini & Piselli 2016, Howell 2017) or through the spillovers they generate (Azoulay et al. 2019, Myers & Lanahan 2022), our study quantifies how variations in the amount of allocated resources – beyond the presence or absence of access – affect scientific production. This is usually hard to estimate given that variation in resources usually comes from shocks that vary *whether* firms or individuals get access to inputs or technology (Kong et al. 2022, Jacob & Lefgren 2011, Nagaraj & Tranchero 2023, Furman & Teodoridis 2020, Ding et al. 2010) but not necessarily *how much* they get access to. In our context, it is crucial to understand this margin as the relevant constraint to pursuing frontier research is not necessarily whether a researcher has access to *any* computing power, but rather *how much* computing power they have access to.

Third, we contribute to the literature on firms' and researchers' innovation strategies by providing insight into how resource constraints influence researchers' choices. This literature explores risk-taking (Mandler 2017, Krieger et al. 2022, Franzoni & Stephan 2023), effort allocation (Blomfield & Vakili 2023), search (Fleming & Sorenson 2004, 2001, Fleming 2001), collaboration networks (Fry 2023, Fry et al. 2020) and other changes in innovation direction such as the departure from one's existing line of research (Myers 2020, Nagle & Teodoridis 2020) or the expansion of existing projects (Myers & Tham 2023). Building upon this literature, our study investigates how high-performance computing resource allocation influences researchers' innovation strategies. By analyzing the impact of increased resource allocation on research output and particularly research direction, we offer insights into the mechanisms driving innovation. Our findings provide valuable contributions to understanding the complex interplay between resource allocation decisions, researchers' strategic choices, and the pursuit of innovative and cutting-edge scientific research.

Section 2 lays out a simple conceptual framework to ground our empirics. Section 3 describes the XSEDE setting and Section 4 describes our data. Section 5 discusses our empirical approach. Section 6 outlines our findings, and Section 7 concludes.

2 Researcher Strategies Under Resource Constraints

2.1 Do researchers produce more output as resource constraints are relaxed?

We begin by examining how an increase in the amount of computing resources allocated to researchers (i.e., the relaxing of constraints) impacts the quantity of output they produce.[4] In practice, it is not obvious that more resources necessarily lead to an increase in output. Researchers could potentially be less efficient with additional resources, or they might leverage external alternatives, which could offset the expected gains. However, based on previous work and the limited availability of outside options, we might expect that an increase in the amount of computing resources leads to an increase in output.

First, existing literature tends to find a positive correlation between the availability of resources and scientific or innovation-related output, typically measured by publications or patents. For instance, access to data or to technology such as Information Technology (IT) is associated with an increase in the number of publications (Ding et al. 2010, Nagaraj & Tranchero 2023). Mezzanotti & Simcoe (2023) indicates that investments in basic research decline when funding costs increase. Additional work shows that tax credits stimulate additional R&D activity by reducing the cost of R&D (e.g., Hall & Van Reenen 2000). In the scientific context, research suggests that government funding leads to increased research output (Payne et al. 2003, Arora & Gambardella 2005, Ganguli 2017). These insights indicate that enhancing resource availability generally leads to greater research output.

Importantly, prior work suggests that the impact of getting more resources on research output largely depends on the availability of outside options. For instance, Jacob & Lefgren (2011) find that being funded by the National Institute of Health (NIH) has a positive but only small effect on the number of publications and citations and show that this is consistent with researchers being able to shift to other sources of funding. In contrast, high-performance computing resources typically lack direct substitutes, due to their specialized nature and the significant infrastructure required to support them which includes high costs and substantial space requirements (Dongarra et al. 2011, Millett & Fuller 2011). Indeed, unlike other research inputs that are fungible, HPC systems are uniquely designed to handle large-scale simulations, complex data analysis, and computationally

[4] Our conceptual framework discusses the 'relaxing' of constraints to align with our empirical framework. Consistent arguments (in the opposite direction) could be made considering the 'imposing' of constraints.

intensive tasks. Further, large simulations often cannot be split up onto multiple resources but require an allocation within one environment. Hence, outside options are limited.

As a result, we expect that an increase in the amount of computing resources will have a positive impact on researchers' output.

2.2 Do researchers produce a different type of output as resource constraints are relaxed?

While we establish that the quantity of output is likely to increase with additional computing resources, the type of innovation produced when resource constraints are relaxed might also differ. Prior literature suggests that technological, spatial, institutional, and fungible constraints have meaningful consequences on innovation characteristics (Catalini 2018, Boyabatlı et al. 2016, Cerqueiro et al. 2017) and highlights mechanisms such as risk-aversion, competition, or task reallocation (Krieger et al. 2022, Mezzanotti & Simcoe 2023, Furman & Teodoridis 2020). Since supercomputing resources represent a critical input for researchers and are also subject to availability constraints, an increase in these resources could similarly impact the nature of the science pursued. Typically, incentives in science lead researchers to prioritize projects and activities that are more likely to advance their careers by yielding tangible, recognized outputs (Stephan 1996, Azoulay et al. 2007). Receiving additional resources may increase researchers' willingness to take on projects with different characteristics (Manso 2011) by reducing the risks associated with undertaking projects that are more uncertain or less "conventionally incentivized" (e.g., Azoulay et al. 2011). Additional resources can provide researchers with more slack to select projects with greater uncertainty or can allow them to pursue both conventional work and more uncertain projects. Overall, the specific way in which computing resources shape the type of science pursued by researchers remains under-explored and warrants further investigation. We examine three potential pathways that researchers may reasonably pursue with more resources: (i) exploring less-studied research avenues (ii) venturing beyond their expertise and (iii) broadening their research scope. While these strategies are not exhaustive, they offer a view into crucial ways in which computing resources might influence the type of science researchers engage in.

First, researchers may decide to use additional computing resources to pursue less-studied research avenues, either by studying topics that are inherently less popular or by focusing on topics that are younger and have appeared more recently in the scientific literature. The prevailing norms

and incentives within the scientific community tend to prioritize the pursuit of topics that yield tangible returns, such as papers and citations (Dasgupta & David 1994). While studying less-studied areas might allow researchers to establish priority on a specific topic, it may also come at the cost of having a smaller audience and making the peer-review process more difficult (Krieger et al. 2023), ultimately increasing the perceived risk of pursuing these topics. As suggested by Azoulay et al. (2011), incentives such as the tolerance for early failure can lead researchers to change their research direction and to choose less-traveled scientific avenues. By allowing researchers to bear more risk, getting additional computing resources may similarly lead researchers to study less-traveled research avenues.

Second, the relaxing of resource constraints may lead researchers to embark on research endeavors that are 'new-to-them' and to venture beyond their existing areas of expertise (Hill et al. 2021). Along these lines, a large body of innovation literature studies the exploration-exploitation trade-off, its underlying mechanisms, as well as its boundary conditions (Luger et al. 2018, Marino et al. 2015, Danneels & Sethi 2011, Sidhu et al. 2007, Choi et al. 2016, Fitzgerald et al. 2021, Zhuo 2022). While various incentives in the modern scientific community tend to encourage specialization (Jones 2009, Foster et al. 2015), researchers could use additional resources to pursue projects outside their specialty - which are inherently more uncertain - as constraints are relaxed.[5] Hence, an increase in computing resources might increase researchers' exploration of topics that are not part of their existing knowledge base.

Third, researchers may broaden the scope of their research by pursuing more interdisciplinary or cross-cutting topics that traverse traditional disciplinary boundaries. Breadth, or scope, and its relationship with firm performance or innovation has been studied in various streams of literature, such as organizational learning (Dutt & Lawrence 2022, Ferguson & Carnabuci 2017). While prior research often uses knowledge breadth as a mechanism that impacts innovation (Nagle & Teodoridis 2020, Zhou & Li 2012), the scope of research may also be endogenous to the constraints researchers face. However, our understanding of how resource constraints impact the breadth of researchers' projects remains limited. Because scientific norms tend to reward researchers' expertise on a specific piece of knowledge (Jones 2009), researchers might not have strong incentives to pursue broad projects at first. In addition, broader projects might structurally require more resources to

[5]This mechanism is in line with Krieger et al. (2022) where, when firms receive cash windfalls, they invest in riskier drug candidates outside of their existing portfolio.

get completed. By shifting researchers' incentives towards undertaking more uncertain projects and/or allowing them to allocate more resources to a project, we posit that getting additional computing resources might increase the research space of researchers' projects.

Overall, we expect that an increase in the amount of computing resources will lead researchers to pursue a different type of science, the nature of which we explore empirically.[6]

3 Setting

To investigate how constraints on supercomputing power impact innovation, we examine XSEDE, an NSF-funded initiative that provides a unique and valuable context for our study.

3.1 XSEDE

XSEDE is an initiative funded by the National Science Foundation (NSF) in 2009 for a total of $257M.[7] XSEDE is one of the most advanced, powerful, and robust collections of integrated digital resources and services in the world for open scientific research and is designed to provide researchers across various disciplines with access to advanced computing resources. XSEDE establishes a distributed network of high-performance computing resources located across the United States. These supercomputers enable researchers to tackle computationally intensive scientific challenges that were previously beyond reach.

Figure 1 shows the location of the main supercomputers used as part of the program. Table A1 shows the ten most used XSEDE supercomputers during our study period. Their associated total capacity allocated during the period is expressed in the common metric used by XSEDE called Service Unit (SU), which can be roughly equated to "core-hours," and allows one to compare computing power across XSEDE resources whose computational specifications vary.[8]

Our interviews with several PIs highlighted how critical XSEDE is for their research agenda because it cannot be easily substituted by other options. One researcher we spoke with, who runs

[6] Our discussion focuses on researchers' altering the projects they pursue when provided with additional resources. We cannot rule out the possibility that additional resources serve to reveal the pre-existing prioritization of projects without influencing choices. However, the core implications hold true even if one interprets the scenario as revealing an underlying project ordering.

[7] XSEDE's predecessor was called TeraGrid and started in 2001. Following its success, XSEDE was replaced in 2022 by "ACCESS" (Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support).

[8] XSEDE converts machine specific core-hours to SUs using resource-specific conversion factors that are based on the LINPACK benchmark, a commonly accepted measure that captures a resource's ability to quickly solve problems.

simulations to understand the impact of pollution and heat dispersion, emphasized that XSEDE is critical for their work because typical university clusters are too small. Another researcher, focused on using genetic material to unravel evolutionary history, noted that the computational demands of their work are so high that only XSEDE can provide the necessary resources, making it indispensable for their research. As such, XSEDE's computational resources have led to significant research across various disciplines by enabling complex and computationally intensive projects. For example, XSEDE has supported investigations into protein structures, where computational simulations facilitate the understanding of complex biological processes and disease mechanisms, aiding in the development of therapeutic strategies. Another example is in neuroscience, where researchers have utilized XSEDE to analyze large-scale single-cell RNA sequencing data to map neural circuitry and understand brain function, providing insights into cellular diversity and neural circuit mechanisms. Furthermore, researchers have used XSEDE resources to study how genetic diversity affects the survival rates of threatened animal species from youth to adulthood. These examples highlight XSEDE's critical role in advancing scientific understanding and innovation across an array of fields.

3.2 Resource Allocation Process

Research allocations are available to researchers affiliated with U.S.-based organizations, with the vast majority of these researchers being associated with universities or other research institutions. Specifically, in our sample, 94.1% of researchers are associated with a university, 5.6% with a non-university research institution, and 0.3% with a commercial entity. Each quarter, the lead researcher - hereafter referred to as the Principal Investigator (PI) - can submit a proposal, which aims to tackle a specific scientific problem. These scientific problems are a bundle of related projects within a researcher's agenda. Proposals include a scientific background, research objectives, a computing allocation request, and a resource usage plan covering code performance timings, resource usage details, scaling information, and benchmark runs on the requested supercomputers (see an example of a proposal in the field of Biophysics in Appendix G).

For each of the quarterly windows of submission, the XSEDE Resource Allocation Committee (XRAC) is responsible for reviewing requests. On average, around 70% of projects are accepted and thus are awarded some computing power.[9] This committee includes about 60 computational

[9]Rejected projects are typically accepted in a future quarter, so there are very few outright rejections. This paper

researchers with expertise in various fields covering the diversity of fields of science encountered at XSEDE, with the most represented fields being more populated. Reviewers serve a three-year term. The role of this committee is to conduct an impartial and independent review of each resource allocation request in order to make a *recommendation* regarding the amount of resources that the project (or "grant") would need to be completed as proposed. Importantly for our empirical design, reviewers are asked to gauge requests only on their own merit (i.e., what does the project require?) and are specifically not informed of the overall quantity of resources available or the other requests across XSEDE's systems.[10]

In our study period, aggregate demand generally exceeds available supply. Hence, after each meeting, the "Allocations Team" meets to reconcile the recommendations made by the XRAC with the overall available supply of resources. This process starts with a "balancing" phase, where recommendations from over-recommended resources are moved to similar ("architecturally equivalent") resources that are less over-recommended until the excess demand is balanced equally across similar computing platforms.[11] Next, this is followed by a "reconciliation" phase where recommendations are reduced through a formulaic solver in order to determine final *allocations*. The formula gives priority to NSF-funded projects, which represent about 40% of the proposals. Importantly, the only project-related characteristics considered in the reduction phase are (i) the recommendation made by the XRAC and (ii) the share of the project funding that is supported by the NSF. Otherwise, the reductions are a mechanical function of the level of over-subscription in each time period.[12]

Following an initial allocation, a PI can request additional resources in following quarter(s) via a "renewal." Renewal proposals are evaluated in the same way but also ensure that the prior allocation(s) led to some progress. We discuss this in more detail in section 5. Figure A1 summarizes the XSEDE allocation process. Figure 2 shows aggregate requests (i.e., demand), recommendations, and final allocations (i.e., reduced recommendations after considering supply available) for each quarter of our study period. Appendix H discusses more details on the allocation process.

---

focuses on accepted proposals.

[10] Appendix H.1 provides excerpts from the *XSEDE Reviewer Manual* describing the review criteria for requests.

[11] "Architecturally equivalent" refers to resources which have similar computational architectures and are thus near substitutes for the research that would be performed on them. For XSEDE, there are broadly two categories of architecturally similar resources: CPU-oriented systems and GPU-oriented systems.

[12] See Appendix H.2 for more details on the reconciliation formula.

4 Data

To study the impact of resource constraints on innovation within the XSEDE program, we use data about allocation awards covering each quarter between 2015 and 2022.[13] For each proposal, the data contains information at the supercomputer level about the initial request submitted by each PI, the recommended amount suggested by the XRAC, and the final allocation that is awarded. We drop resources corresponding to "storage" and we convert request, recommended, and allocated information associated with different supercomputers into standardized Service Units using conversion factors provided by XSEDE. We also observe information about the proposal field of science, as well as PI first name, last name, organization, and the percent of the funding coming from the NSF. Since both the initial request and any subsequent renewals are tied to the same project, we treat them collectively as a single XSEDE grant, associated with a PI. We therefore collapse information at the XSEDE grant (i.e., PI) level, keeping only grants for which we have complete information about their resources and for which we observe their first request. Each row of our dataset contains information about the total amount of requested resources made by the PI, the corresponding total amount recommended by the XRAC, and the total amount of allocated resources approved by the XSEDE allocation team. In theory, we could have kept our dataset at the request (i.e., quarter) level instead of collapsing it at the PI level. However, the scientific output that we observe can only be attributed to PIs and their XSEDE grants, and not to individual quarterly requests, making *grants* the relevant unit of observation. This corresponds to 1,062 unique grants.

In order to match each XSEDE grant to its scientific research production, we use two sources of data. We first perform a fuzzy match of our population of PIs to the *Dimensions AI* database using first name, last name and institution name.[14] Once PIs are identified, we retrieve their papers and keep those that reference XSEDE in the acknowledgments section.[15] Dropping book chapters, this gives us 3,659 papers. We also obtained access to the XSEDE internal publication database, where PIs can report the publications associated with their grant. From this, we retrieve 875 additional papers.

[13] We are missing complete information about proposals for the years before 2015.

[14] See https://www.dimensions.ai/. In recent years Dimensions AI has been widely used in economics and management research, see e.g., Tian & Smith (2024), Fry & MacGarvie (2024), Myers et al. (2023), Myers & Tham (2023).

[15] From our discussions with PIs who were part of the XSEDE program, acknowledging XSEDE in the papers produced is the norm. We may be missing papers that attempt to acknowledge XSEDE without specifically mentioning a variation of "XSEDE" or "Extreme Science and Engineering Discovery Environment," but to be conservative we capture papers that we are confident originate from XSEDE computing allocations.

We then retrieve information about each paper using Dimensions AI. Our empirical analysis relies heavily on the notion of "concepts" from the Dimensions database. More precisely, the Dimensions AI database uses machine learning techniques to derive concepts from papers' abstracts and rank them based on their relevance on a scale from 0 (not relevant) to 1 (very relevant).[16] In all analyses, we only consider concepts with a score above 0.5, in order to assign the most relevant topics to a given paper. Appendix I provides more detail on Dimensions concepts and our use of them.

Among the initial 1,062 grants, we observe scientific output for 719 of them. Among the 343 unmatched grants, 112 come from PIs we were unable to match or unable to uniquely match to Dimensions. For the remaining 231 grants, 1 grant has papers acknowledging XSEDE, but cannot be uniquely assigned to that grant because the associated PI is linked to multiple grants and 230 grants have PIs who did not report any paper to the internal XSEDE database nor acknowledge XSEDE in any of their publications.[17] Of the 719 grants with scientific output, we drop 9 grants whose ratio is strictly higher than 1 due to XSEDE data entry errors (see Table A2 for the successive sampling restrictions). Our final sample hence includes 710 unique grants. Section 5 discusses additional checks we perform on our data construction.

5 Empirical Strategy

Empirically, we want to estimate the impact of resource constraints on scientific output (rate and characteristics). In the ideal experiment, we would assign each researcher (i.e., grant) a random number of computing hours that is weakly less than what they require (i.e., randomly assign a level of constraint). We would then observe the subsequent output associated with the grant, which would allow us to estimate the effect of the marginal computing resource (i.e., marginal relaxing of constraint) on each researcher's output. While we cannot run such an experiment, we leverage the allocation process of XSEDE in order to get plausibly causal estimates.

In most similar settings one could study, the typical threat to a causal interpretation is that projects and researchers with a higher expected productivity will receive higher computing allocations (will be less constrained), which could bias estimates. For instance, these projects would

[16] Several papers use MeSH terms to analyze the research topics of scientific papers (Azoulay et al. 2011, Myers 2020). MeSH terms are used for publications in PubMed which covers the biomedical literature. Conceptually, the use of Dimensions AI concepts allows us to extend this analysis to all the fields of science in our dataset.

[17] We drop these 230 grants from our data since we cannot identify which of these grants are "true zeroes" (i.e., did not produce any papers) versus those that did produce papers but did not cite XSEDE.

produce more papers regardless of the quantity of resources they receive, biasing our quantity estimate upwards. We overcome this by exploiting the fact that, conditional on the recommendation from the XRAC and the percentage of project funding that is supported by the NSF, each grant's final allocation is a function of the overall excess demand at the time of application, which is plausibly exogenous to the expected productivity of the grant.[18]

Our regressions of interest are of the form:

$$Y_i = \beta_0 + \beta_1 \, \%Allocated_i + \delta \, \ln(Recommended_i) + \gamma \, \%NSF\,Funding_i + \mathbf{X}_i + \varepsilon_i \tag{1}$$

where i indexes grants. Our treatment variable is $\%Allocated_i = \frac{Allocated_i}{Recommended_i}$ which captures the fraction of resources allocated to a grant relative to what it would have objectively needed as assessed by the XRAC – higher values indicate a grant is less constrained. This specification allows us to interpret the coefficient β_1 as the impact of receiving more computing resources relative to what a researcher would have needed.[19] Note that by definition of the XSEDE Allocation process, this ratio is less than or equal to one (i.e., all projects are weakly constrained). We control for the recommendation made by the XRAC with the logarithm of $Recommended_i$. One can think of this variable as capturing a combination of project quality and scale, where larger projects will inherently exhibit different output quantity and characteristics than smaller projects. The variable $\%NSF\,Funding_i$ accounts for the share of the project funding that is supported by the NSF, since this is part of the reconciliation formula. The vector $\mathbf{X}_i$ represents a set of fixed effects. We include field of science fixed effects, since some fields might apply for specific resources which are deferentially more or less busy and these fields might also have different research norms (e.g., more papers or different research characteristics).[20] We also include grant start year fixed effects in order to control for changes in overall supply over time as well as the potential for changing research output and characteristics over time. Our model includes the interaction of field of science and grant start year fixed effects which control for field specific time trends. We use robust standard errors clustered at the grant start year × field level because the reconciliation process is performed

[18] Intuitively, excess demand is driven by both the quantity of recommendations made to other researchers and supercomputers being introduced and retired over time.

[19] In Appendix E, we confirm all results are robust to using the level of allocation, $\ln(Allocated_i)$, as the treatment variable. Appendix F shows results where we allow for $\%Allocated_i$ to have a non-linear effect.

[20] Our data has three different levels of field of science aggregation: directorate (least granular), division, and field (most granular). Our baseline results use field but our results are robust to using the other levels of aggregation.

across architecturally similar machines.[21]

We perform several tests to assess whether our specification provides us with plausibly causal estimates. First, we confirm that observable characteristics of researchers do not correlate with the level of constraint a project has. In Table A3, we show that the main independent variable of interest, $\%Allocated$, is uncorrelated with pre-determined covariates related to the PI: seniority, institution, number of papers published before XSEDE, and number of citations received before XSEDE. Second, given that the XRAC has information about the amount of computing resources requested by researchers, one might worry that recommendations are "anchored" on researchers' requests, leading some researchers to receive higher recommendations because they requested more resources than what their projects needed. In order for this to create endogeneity issues, asking for more should be correlated with both (i) the outcomes (i.e., PIs who ask for more should be either more or less productive than the average, or produce research with different characteristics) and (ii) the main explanatory variable (i.e., PI asking for more should have a higher or lower $\%Allocated$ than the average). While we cannot observe which PIs strategically asked for more than their true need, we create a proxy for the share of over-requested resources by measuring the "percent over-request" ($\frac{requested_i - recommended_i}{recommended_i}$), since larger gaps between the recommendation and request *may* suggest that the request has been inflated by the researcher. We then examine the relationship between this variable and $\%Allocated$, including the controls. Table A4 shows no significant correlation.[22]

6 Results

6.1 Descriptive Statistics

Table 1 shows descriptive statistics about our sample of 710 grants. On average, researchers request 61.6 million service units (SU). The XRAC makes recommendations equal to 41.7 million SU on average and researchers gets a final average allocation of 35.8 million SU, with large variation in

[21] Fields of science at a given point in time often require particular types of computing resources that are architecturally similar, leading to a reconciliation process that may be correlated within field-year.

[22] We also revisit checks on our data construction in light of our empirical framework. Note that a PI's decision to report to the internal XSEDE database is voluntary. We confirm that there is no correlation between our treatment variable ($\%Allocated$) and papers identified through the use of the internal database. Additionally, in order to assess whether grants in our final population are a selected sample, we construct an indicator equal to 1 if the grant is part of the 710 grants that constitute our main sample and 0 otherwise and we regress this measure on the main independent variable, $\%Allocated$, and the main controls. Table A5 shows no statistically significant relationship.

this number. On average, the share of allocated resources compared to the recommendation made by the review committee is 90%, but this variable is skewed. This can be seen in Figure 3 which shows the histogram of the treatment variable (*% Allocated*). On average, each grant is associated with 6.3 papers. Figure A2 shows the 15 most common fields of science in our sample. The most represented field is Materials Research with 56 grants, followed by Biophysics (49 grants) and Materials Engineering (42 grants).

6.2 Quantity of Science

Our conceptual framework discusses the relationship between the amount of computing resources available to a researcher and the *quantity* and *type* of science they produce. In order to study the impact of additional resources on *quantity*, our outcome of interest is the log of the total number of publications associated with each grant.[23] Table 2 shows the results, where we regress this outcome on the share of allocated resources received by grants. Column (1) shows a naïve regression including only %*Allocated* and year fixed effects and displays a strong positive relationship, significant at the 5% level. Columns (2) and (3) successively add controls for the amount of computing resources recommended by the XRAC and the share of funding coming from the NSF, both of which enter the reconciliation formula. Column (4) also controls for field of science fixed effects and Column (5) adds grant start year $\times$ field fixed effects. Using Column (5) as our preferred specification, results show a strong positive relationship significant at the 1% level: a 10 percentage point increase in the percentage of allocated resources leads to approximately an 10.9% increase in the number of papers produced. Equivalently, an increase from the 10^{th} to the 75^{th} percentile in the share of allocated resources leads to a 52.9% increase in the number of papers produced.[24] Column (6) confirms that results are robust to using a Poisson specification with the count of papers as the outcome. This result implies that scientific output increases as resources constraints are relaxed.

6.3 Type of Science

In what follows, we explore the consequences of an increase in the share of allocated computing resources on the type of science that researchers pursue. We successively study whether such increase leads to: (i) researchers pursuing less traveled research avenues (ii) researchers studying

[23] Since our outcome does not contain zeroes (each grant in our sample is associated with at least one paper), we avoid issues described in Chen & Roth (2024).

[24] The 10th percentile of %*Allocated* equals 59% and the 75th percentile of %*Allocated* equals 1.

topics beyond their expertise and (iii) researchers broadening the scope of their research.

6.3.1 Less traveled research avenues

First, an increase in the share of computing resources could impact researchers' propensity to work on less-studied topics. For each paper, we calculate the number of times each concept it entails has been used by other PIs in the same field, excluding all papers from the focal PI, and average this measure at the paper level.[25] The higher this number, the more the paper entails concepts that have been used by other researchers too – these are "more studied" concepts. The lower this number, the more the paper entails concepts that have been less studied by other researchers – these are "less frequently-studied" concepts. We then collapse this to the grant level using different measures of interest.[26] First, we look at the average among papers within grant, which gives us information about how well-studied the topics of the grant are on average. We also gain further insights about the overall distribution by considering the minimum among papers within grant, i.e., the average use of concepts by other researchers for the least "well-studied" (most "unconventional") paper. Finally, we also consider the maximum among papers within grant, i.e., the average use of concepts by other researchers for the most "well-studied" paper.

Table 3 shows the results, with Column (1) using the average as the outcome, Column (2) using the minimum as the outcome and Column (3) using the maximum as the outcome. All columns control for the amount of resources recommended by the XRAC, as well as the share of funding coming from the NSF. All specifications include grant start year fixed effects, field fixed effects, and grant start year $\times$ field fixed effects. Column (1) displays a negative and statistically significant coefficient, showing that the higher the share of allocated resources, the less popular are the topics studied on average. The coefficient implies that a 10 percentage point increase in the share of allocated resources leads to the study of concepts that are about 6.3% less studied by other researchers. Such a relationship could be driven by changes at the tails (e.g., the papers that study the least (or most) popular topics explore more niche topics) which would inform us about the boundaries of researchers' work. Column (2) shows that this result is driven by papers in the left tail of the distribution: for the paper with the least studied topics, a 10 percentage point increase in allocation leads to the study of concepts that are about 16.3% less studied by other

[25] We normalize this number to account for publication year following Perry & Reny (2016).

[26] We use the log of this measure. In Table A6, we instead create indicator variables equal to one for grants above median of this measure and zero otherwise which allows us to keep observations whose value is zero. We find consistent results.

researchers. Column (3) shows no change at the right tail of the distribution. Overall, we find that an increase in the share of computing resources leads researchers to explore less-studied concepts, driven particularly by shifts in papers that focus on the least-studied topics.

Thus far, results consider topic usage both before and after a given grant's start date. However, this negative relationship could be driven by (i) the study of topics that were initially less popular but rising in popularity (ii) the study of topics that were popular but losing attractiveness or (iii) the study of topics that have a persistently lower level of use by others. In order to investigate this, we begin by calculating for each paper the number of times each concept it entails has been used by other PIs *prior to*, and separately *after*, the paper publication year and average this measure at the paper level. This allows us to study a relative trend in use by others before versus after publication year to disentangle (i)-(iii). We then reiterate our previous analysis, considering as outcome the mean value of this measure across all papers associated with a grant, the minimum value of this measure across all papers associated with a grant, and the maximum value of this measure across all papers associated with a grant. Results are presented in Columns (1) to (3) of Table A7 using the '*use prior*' measure and show similar results. Results presented in Columns (4) to (6) of Table A7 use the '*use after*' measure and again show similar findings. This suggests that when researchers are allocated a higher share of resources, they study topics that are persistently less popular – i.e., less conventional.

We next study whether the concepts used are more or less recently introduced to the scientific literature as researchers get allocated additional computing resources – e.g., are these "less studied" concepts also "newer to the world?" We take each concept associated with a paper and calculate the first year we see this concept appearing among XSEDE researchers.[27] We then average this measure at the paper level. This gives us information about the average 'vintage' of the concepts used in a paper. The higher this measure, the more recent the concepts it includes are. The lower this measure, the older the concepts it studies are. Column (4) of Table 3 uses as an outcome the average of this value across all papers associated with a grant. This measure proxies how recent the concepts used by a grant are on average. Columns (5) and (6) use as the outcome respectively the minimum and maximum of this value across all papers associated with a grant, which gives

[27]Since we do not observe the *universe* of scientific articles in our data, we cannot observe with certainty when the topic was initially studied overall. We use the first time we observe a topic in XSEDE as a good proxy for its appearance in the computational scientific literature. Further, we have no reason to suspect a correlation between when concepts appear among XSEDE researchers relative to the universe and a grant's computing allocation.

us information about the paper with the oldest and the most recent concepts, respectively. All specifications control for the amount of resources recommended by the XRAC, the share of funding coming from the NSF, grant start year fixed effects, field fixed effects, and grant start year $\times$ field fixed effects. Column (4) shows no significant effect, implying that an increase in the share of computing resources does not impact the *average* age of concepts associated with a grant. We also do not find any significant result in Column (5), implying that an increase in the share of computing resources does not impact the age of concepts associated with the "oldest vintage" papers. However, Column (6) shows a positive and statistically significant estimate at the 5% level, implying that an increase in the share of computing resources leads to the use of more recent concepts associated with the "newest vintage" papers. This suggests that an increase in the share of computing resources leads researchers to advance their work into newer territories, as it is not the "oldest vintage" papers that change, but rather the "newest vintage" ones that explore even newer ideas. More specifically, Column (6) shows that a 10 percentage point increase in $\%Allocated$ leads to the study of topics that are about 0.47 years newer on average.

Taken together, we find strong evidence consistent with the fact that as resource constraints are relaxed, researchers choose to study topics that are less common and newer to the world.

6.3.2 Topics beyond expertise

An increase in the share of computing resources could also impact researchers' propensity to "pivot," i.e., to work on topics that they haven't studied previously. To analyze this, we calculate for each paper in our sample the share of concepts that have not been used by the focal PI in the years preceding their entry into XSEDE. We then study the average, minimum, and maximum of this measure at the grant level. Conceptually, we are capturing the degree of "pivoting" or "exploration" by each researcher, studying topics that they are less familiar with. Using the average allows us to capture the general tendency of researchers to explore new topics across all papers within the grant as computing resources increase. Using the minimum within a grant helps us see what happens for papers that study the most familiar topics. Using the maximum gives us information about papers that focus on the most unfamiliar topics, i.e., the most "exploratory" papers.

We use the pool of papers written by the researcher in the two years preceding their entry into XSEDE to define the initial pool of papers with which we compare papers written as part of

XSEDE.[28] Table 4 shows the results. Column (1) examines the mean share within grant. Column (2) examines the minimum and Column (3) examines the maximum. We find no significant effects on the mean nor minimum measures but strong positive results on the maximum value within grant. More precisely, a 10 percentage point increase in the share of allocated resources leads to about a 1.2 percentage point increase in the share of new concepts. Overall, our findings imply that as resource constraints are relaxed, researchers explore 'new-to-them' topics among the most exploratory of their papers. This finding is consistent with the idea that the relaxing of computing constraints allows researchers to pursue more "exploratory" work that they do not seem to pursue in more constrained environments.

6.3.3 Research breadth

Finally, an increase in the share of computing resources could impact the breadth of research projects pursued by researchers. To explore this, we study the effect of an increase in the share of allocated computing resources on the number of concepts used by papers associated with a grant. We first calculate the number of unique concepts associated with each paper, which informs us about the diversity of ideas it embodies. Appendix I shows an example of this measure. Our first outcome averages this measure at the grant level. We use this as a proxy for the breadth of research associated with a grant: the higher the average number of concepts, the broader the research on average. We also use as an outcome the minimum of this value across all papers associated with a grant. We use this as a proxy for the breadth of the "narrowest" paper associated with a grant. Similarly, we use as an outcome the maximum of this value across all papers associated with a grant. We use this as a proxy for the breadth of the "broadest" paper associated with a grant.[29]

Table 5 shows the results. Column (1) studies the average within grant and shows a significantly positive coefficient, implying that a 10 percentage point increase in the share of allocated resources leads to a 4.5% increase in the number of concepts associated with a grant. This indicates that a higher share of allocated resources leads researchers to expand their research breadth. Column (2) studies the minimum within grant and shows no effect. Column (3) uses as an outcome the maximum within grant and shows that the result in Column (1) is driven by the "broadest" papers becoming broader. This implies that as resource constraints are relaxed, researchers work on

[28] Table A8 reiterates this analysis using a 3-year and 5-year pre-XSEDE period and finds similar results.

[29] In contrast to the journal breadth measure used in Azoulay et al. (2011), our measure more directly captures researchers' pursuit of broad topics separately from the impact in the scientific community.

broader papers and specifically expand the broadest ones.

6.4 Journal Quality and Citations

Thus far, we find evidence that as researchers receive more computing resources, they write more papers. We also find that researchers with additional resources study less popular and newer topics that go beyond their prior expertise, and broaden the scope of their projects (see Figure 4 for a summary of the results). These findings provide insight into how computing resources influence researchers' strategies, as they reflect deliberate choices undertaken by researchers. In this section, we explore downstream outcomes that the researchers have less control over. We use journal placement as our measure of *ex-ante* quality and normalized forward citations as our measure of scientific impact.

First, we study the impact of additional resources on the quality of journal placement for papers associated with a grant. Specifically, we exploit information about the SNIP (Source Normalized Impact per Paper) value of the journal associated with each paper. SNIP measures the average citation impact of the publications in a given journal, which we use as a proxy of journal quality independent of the citations accrued to any given paper.[30] Column (1) of Table 6 examines the average SNIP value of papers associated with a grant. Columns (2) and (3) repeat this exercise by considering respectively the minimum and maximum SNIP value of papers associated with a grant. Overall, we do not find any significant estimate, suggesting that an increase in computing resources may not have a significant impact on the *ex-ante* quality distribution. This suggests that an increase in the share of allocated resources does not lead to a change in inherent quality.

Next, we study the impact of an increase in the share of computing resources allocated to researchers on citations. We use an adjusted citation measure that normalizes by publication year and the field of science associated with each paper.[31] Since this measure already accounts for field differences, we do not control for field fixed effects or field $\times$ year fixed effects in these analyses.[32] This essentially allows us to compare the citations received by papers in our sample relative to the *universe* of papers, while normalizing by publication year and field. This measure alleviates

[30] SNIP values are normalized to account for differences in citation norms over years.

[31] This measure is provided by Dimensions AI and calculates the scientific influence of a paper by dividing the number of citations it received by the average number of citations received by papers published in the same year and field of research. A value of one means that a paper receives exactly the average number of citations expected for its publication year and field.

[32] We still control for grant start year FE since the Dimensions normalization is by paper publication year, which is different than grant start year.

potential concerns about truncation issues when studying papers published relatively recently, since papers are compared within field and publication year. We then collapse this measure at the grant level, calculating the mean, minimum, and maximum across papers. Results are presented in Table 6. Column (4) uses as the outcome the average number of citations received by papers associated with each grant. Columns (5) and (6) repeat this exercise by using the number of citations associated with the least (respectively most) cited paper associated with a grant. Columns (4) and (6) show non-statistically significant estimates, while Column (5) shows a strong negative relationship. This suggests that, while an increase in the share of computing resources does not lead to a significant change in average scientific impact or reception, it does lead to a significant decrease in the number of citations received by the least-cited papers associated with a grant.[33] Given the results in columns (1)-(3) and our findings about changes in research direction, we are cautious to interpret this finding as representative of lower quality papers *per se*. While this is possible, it is also logical to interpret this finding as a consequence of the changing characteristics of the work produced which may correlate with less follow-on work and thus citations by other researchers.

To further explore this finding, we shed light on the association between the directional changes we estimate and this citation decrease. Table A9 regresses normalized citations on our different measures capturing research direction. Column (1) shows that an increase in the use of conventional topics is associated with an increase in citations. Column (2) shows that an increase in the use of more recent topics is associated with a decrease in citations. Column (3) shows that an increase in the share of new concepts is associated with a decrease in citations. Given that our previous results concluded that an increase in the share of allocated resources leads to the study of less conventional, more recent, and more exploratory topics, Columns (1), (2) and (3) are consistent with the idea that these changes in research direction following an increased resource allocation lead to less citations. Column (4) shows that an increase in breadth is associated with an increase in citations. Given that our previous results concluded that an increase in the share of allocated resources leads to broader research, this finding suggests that broader research does not contribute to the observed decline in citations.

[33] In unreported analyses, we construct an alternative field-year normalized citation measure that only uses our sample rather than the universe of papers and find similar results.

7 Discussion

In this paper, we study how constraints on supercomputing resources, a key input to modern innovation, affect the quantity and type of innovation produced. At a high level, innovation is a critical driver of organizations' competitive advantage and overall economic growth (Romer 1990, Schumpeter 1934, Aghion et al. 2005) that results from the combination of various inputs (Henderson & Clark 1990, Fleming 2001). Understanding how constraints on key inputs, such as computing power, influence the rate and direction of innovation is essential for optimizing resource allocation and investments decisions, both in academic and industrial contexts. By examining the allocation of supercomputing resources among researchers, this paper provides some of the first empirical evidence into how variation in resource constraints on the intensive margin can influence both the quantity and the characteristics of innovative output.

Our conceptual framework posits that as resource constraints are relaxed – i.e., as researchers receive more computing resources – output will increase and the type of science that researchers pursue may change. We then empirically investigate the relationship between resource constraints and research output in the context of XSEDE. By leveraging its specific allocation process, we are able to observe variation in the share of computing resources that researchers receive compared to what they would have needed, creating variation in how constrained they are.

Our findings indicate that an increase in the share of allocated resources leads to an increase in research output, as evidenced by the increase in the number of papers published. We then explore changes to research direction. Our findings indicate that a higher share of allocated resources leads researchers to work on less popular and newer topics. We also find evidence that researchers shift toward topics that they have not previously worked on themselves. Finally, we find that researchers who receive more computing resources tend to broaden the scope of their research. We also find a decrease in citations but no significant change in journal placement quality, suggesting that these directional changes may be associated with less follow-on work by other researchers but no inherent change in quality. Importantly, these directional shifts are particularly salient for the *least* conventional papers, those that study the *newest* topics, the *most* exploratory papers, and the *broadest* papers, respectively – implying the relaxing of these constraints actually drives researchers to "spread their wings" and push the boundaries of their work.

Our findings suggest that computing resources, a key constrained input to the innovation pro-

cess, significantly impacts both the rate and direction of innovation. Our study expands our understanding of the role of resource constraints on the characteristics of innovation and highlights the multifaceted nature of the relationship between resource allocation, research strategies, and characteristics of the output produced. In particular, our findings underscore the importance of considering the nuanced impacts of resource allocation strategies on research outcomes and the need for continued exploration of the underlying mechanisms driving these dynamics.

Future work could delve into the relationship between researcher characteristics and the impact of constraints. Additionally, exploring how other types of resource constraints impact the characteristics of innovative output could provide a more comprehensive understanding of how our findings apply to other scarce resources. Finally, future work could consider how access to complementary resources, such as data, interacts with the effects we highlight.

Practically, this work has clear managerial and policy implications. Our study highlights the need for R&D managers and government programs with limited resources to carefully consider how resource allocation decisions affect the characteristics of research output. For R&D managers, particularly those in firms where innovation is critical to competitive advantage, our study provides valuable insights into how resource allocation decisions can influence the direction of research. Managers should be aware that providing less resources can lead to shifts in research focus, with researchers engaging in different types of projects that are less novel and narrower. The implications of our findings are also particularly relevant for policymakers who are involved in the allocation of resources to innovators. Our study underscores the importance of recognizing the impact of computing resource allocations on the type of science and innovation ultimately produced. Since firms also often build on academic research as the foundation for new technologies, the type of science produced in academia directly influences the innovations that firms can develop (Marx & Fuegi 2020, Arora et al. 2018, Aghion et al. 2008, Cohen et al. 2002). This makes the direction of scientific output a crucial factor in shaping future technological advancements.

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R. & Howitt, P. (2005), 'Competition and innovation: An inverted-U relationship', *The Quarterly Journal of Economics* **120**(2), 701–728.

Aghion, P., Dewatripont, M. & Stein, J. C. (2008), 'Academic freedom, private-sector focus, and the process of innovation', *The RAND Journal of Economics* **39**(3), 617–635.

Agrawal, A., Catalini, C., Goldfarb, A. & Luo, H. (2018), 'Slack time and innovation', *Organization Science* **29**(6), 1056–1073.

Arora, A., Belenzon, S. & Patacconi, A. (2018), 'The decline of science in corporate R&D', *Strategic Management Journal* **39**(1), 3–32.

Arora, A. & Gambardella, A. (2005), 'The impact of NSF support for basic research in economics', *Annales d'Economie et de Statistique* pp. 91–117.

Arrow, K. (1962), Economic Welfare and the Allocation of Resources for Invention, Princeton University Press, Princeton, pp. 609–626.

Atanassov, J. (2016), 'Arm's length financing and innovation: Evidence from publicly traded firms', *Management Science* **62**(1), 128–155.

Azoulay, P., Ding, W. & Stuart, T. (2007), 'The determinants of faculty patenting behavior: Demographics or opportunities?', *Journal of Economic Behavior & Organization* **63**(4), 599–623.

Azoulay, P., Graff Zivin, J. S., Li, D. & Sampat, B. N. (2019), 'Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules', *Review of Economic Studies* **86**(1), 117–152.

Azoulay, P., Graff Zivin, J. S. & Manso, G. (2011), 'Incentives and creativity: evidence from the academic life sciences', *The RAND Journal of Economics* **42**(3), 527–554.

Babina, T., He, A. X., Howell, S. T., Perlman, E. R. & Staudt, J. (2023), 'Cutting the innovation engine: how federal funding shocks affect university patenting, entrepreneurship, and publications', *The Quarterly Journal of Economics* **138**(2), 895–954.

Barney, J. (1991), 'Firm resources and sustained competitive advantage', *Journal of Management* **17**(1), 99–120.

Baruffaldi, S. & Gaessler, F. (2021), 'The returns to physical capital in knowledge production: Evidence from lab disasters', *Max Planck Institute for Innovation & Competition Research Paper* (21-19).

Berrone, P., Fosfuri, A., Gelabert, L. & Gomez-Mejia, L. R. (2013), 'Necessity as the mother of 'green' inventions: Institutional pressures and environmental innovations', *Strategic Management Journal* **34**(8), 891–909.

Blomfield, M. & Vakili, K. (2023), 'Incentivizing effort allocation through resource allocation: Evidence from scientists' response to changes in funding policy', *Organization Science* **34**(1), 100–128.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R. & Riedl, C. (2016), 'Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science', *Management Science* **62**(10), 2765–2783.

Boyabatlı, O., Leng, T. & Toktay, L. B. (2016), 'The impact of budget constraints on flexible vs. dedicated technology choice', *Management Science* **62**(1), 225–244.

Bronzini, R. & Piselli, P. (2016), 'The impact of R&D subsidies on firm innovation', *Research Policy* **45**(2), 442–457.

Catalini, C. (2018), 'Microgeography and the direction of inventive activity', *Management Science* **64**(9), 4348–4364.

Cerqueiro, G., Hegde, D., Penas, M. F. & Seamans, R. C. (2017), 'Debtor rights, credit supply, and innovation', *Management Science* **63**(10), 3311–3327.

Chen, J. & Roth, J. (2024), 'Logs with zeros? Some problems and solutions', *The Quarterly Journal of Economics* **139**(2), 891–936.

Choi, B., Kumar, M. S. & Zambuto, F. (2016), 'Capital structure and innovation trajectory: The role of debt in balancing exploration and exploitation', *Organization Science* **27**(5), 1183–1201.

Cohen, W. M., Nelson, R. R. & Walsh, J. P. (2002), 'Links and impacts: the influence of public research on industrial R&D', *Management Science* **48**(1), 1–23.

Conti, A. (2018), 'Entrepreneurial finance and the effects of restrictions on government R&D subsidies', *Organization Science* **29**(1), 134–153.

Conti, A., Dass, N., Di Lorenzo, F. & Graham, S. J. (2019), 'Venture capital investment strategies under financing constraints: Evidence from the 2008 financial crisis', *Research Policy* **48**(3), 799–812.

Danneels, E. & Sethi, R. (2011), 'New product exploration under environmental turbulence', *Organization Science* **22**(4), 1026–1039.

Dasgupta, P. & David, P. A. (1994), 'Toward a new economics of science', *Research Policy* **23**(5), 487–521.

Ding, W. W., Levin, S. G., Stephan, P. E. & Winkler, A. E. (2010), 'The impact of information technology on academic scientists' productivity and collaboration patterns', *Management Science* **56**(9), 1439–1461.

Dongarra, J., Beckman, P., Moore, T., Aerts, P., Aloisio, G., Andre, J.-C., Barkai, D., Berthou, J.-Y., Boku, T., Braunschweig, B. et al. (2011), 'The international exascale software project roadmap', *The International Journal of High Performance Computing Applications* **25**(1), 3–60.

Dutt, N. & Lawrence, M. (2022), 'Learning to manage breadth: experience as repetition and adaptation', *Organization Science* **33**(4), 1300–1318.

Ferguson, J.-P. & Carnabuci, G. (2017), 'Risky recombinations: Institutional gatekeeping in the innovation process', *Organization Science* **28**(1), 133–151.

Fini, R., Perkmann, M. & Ross, J.-M. (2022), 'Attention to exploration: The effect of academic entrepreneurship on the production of scientific knowledge', *Organization Science* **33**(2), 688–715.

Fitzgerald, T., Balsmeier, B., Fleming, L. & Manso, G. (2021), 'Innovation search strategy and predictable returns', *Management Science* **67**(2), 1109–1137.

Fleming, L. (2001), 'Recombinant uncertainty in technological search', *Management Science* **47**(1), 117–132.

Fleming, L. & Sorenson, O. (2001), 'Technology as a complex adaptive system: Evidence from patent data', *Research Policy* **30**(7), 1019–1039.

Fleming, L. & Sorenson, O. (2004), 'Science as a map in technological search', *Strategic Management Journal* **25**(8-9), 909–928.

Foster, J. G., Rzhetsky, A. & Evans, J. A. (2015), 'Tradition and innovation in scientists' research strategies', *American Sociological Review* **80**, 875–908.

Franzoni, C. & Stephan, P. (2023), 'Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals', *Research Policy* **52**(3), 104706.

Fry, C. & MacGarvie, M. (2024), 'Author country of origin and attention on open science platforms: Evidence from covid-19 preprints', *Management Science* **70**(8), 5426–5444.

Fry, C. V. (2023), 'Crisis and the trajectory of science: Evidence from the 2014 ebola outbreak', *Review of Economics and Statistics* **105**(4), 1028–1038.

Fry, C. V., Cai, X., Zhang, Y. & Wagner, C. S. (2020), 'Consolidation in a crisis: Patterns of international collaboration in early covid-19 research', *PloS one* **15**(7), e0236307.

Furman, J. L. & Teodoridis, F. (2020), 'Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering', *Organization Science* **31**(2), 330–354.

Ganguli, I. (2017), 'Saving soviet science: The impact of grants when government R&D funding disappears', *American Economic Journal: Applied Economics* **9**(2), 165–201.

Glennon, B. (2023), 'How do restrictions on high-skilled immigration affect offshoring? Evidence from the H-1B program', *Management Science* **70**(2), 907–930.

Hall, B. & Van Reenen, J. (2000), 'How effective are fiscal incentives for R&D? A review of the evidence', *Research Policy* **29**(4-5), 449–469.

Henderson, R. M. & Clark, K. B. (1990), 'Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms', *Administrative Science Quarterly* **35**(1), 9–30.

Hill, R. (2019), Searching for superstars: Research risk and talent discovery in astronomy, Technical report, Working Paper. Cambridge, MA: Massachusetts Institute of Technology.

Hill, R., Yin, Y., Stein, C., Wang, D. & Jones, B. (2021), 'Adaptability and the pivot penalty in science', *Working Paper* .

Hoegl, M., Gibbert, M. & Mazursky, D. (2008), 'Financial constraints in innovation projects: When is less more?', *Research Policy* **37**(8), 1382–1391.

Howell, S. T. (2017), 'Financing innovation: Evidence from R&D grants', *American Economic Review* **107**(4), 1136–1164.

Jacob, B. A. & Lefgren, L. (2011), 'The impact of research grant funding on scientific productivity', *Journal of Public Economics* **95**(9-10), 1168–1177.

Jones, B. F. (2009), 'The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?', *Review of Economic Studies* **76**, 283–317.

Kim, S. (2023), 'Shortcuts to Innovation: The Use of Analogies in Knowledge Production', *Working Paper* .

Kong, D., Lin, C., Wei, L. & Zhang, J. (2022), 'Information accessibility and corporate innovation', *Management Science* **68**(11), 7837–7860.

Krieger, J. L., Myers, K. R. & Stern, A. D. (2023), 'How important is editorial gatekeeping? Evidence from top biomedical journals', *Review of Economics and Statistics* pp. 1–33.

Krieger, J., Li, D. & Papanikolaou, D. (2022), 'Missing novelty in drug development', *The Review of Financial Studies* **35**(2), 636–679.

Lohr, S. (2019), 'At tech's leading edge, worry about a concentration of power', *The New York Times* .

Luger, J., Raisch, S. & Schimmer, M. (2018), 'Dynamic balancing of exploration and exploitation: The contingent benefits of ambidexterity', *Organization Science* **29**(3), 449–470.

Mandler, M. (2017), 'The benefits of risky science', *The Economic Journal* **127**(603), 1495–1526.

Manso, G. (2011), 'Motivating innovation', *The Journal of Finance* **66**(5), 1823–1860.

Marino, A., Aversa, P., Mesquita, L. & Anand, J. (2015), 'Driving performance via exploration in changing environments: Evidence from formula one racing', *Organization Science* **26**(4), 1079–1100.

Marx, M. & Fuegi, A. (2020), 'Reliance on science: Worldwide front-page patent citations to scientific articles', *Strategic Management Journal* **41**(9), 1572–1594.

Mezzanotti, F. & Simcoe, T. (2023), 'Research and/or development? Financial frictions and innovation investment', *Working Paper* .

Millett, L. I. & Fuller, S. H. (2011), *The future of computing performance: game over or next level?*, National Academies Press.

Myers, K. (2020), 'The Elasticity of Science', *American Economic Journal: Applied Economics* **12**(4), 103–134.

Myers, K. R. & Lanahan, L. (2022), 'Estimating spillovers from publicly funded R&D: Evidence from the

US department of energy', *American Economic Review* **112**(7), 2393–2423.

Myers, K. R., Tham, W. Y., Thursby, J., Thursby, M., Cohodes, N., Lakhani, K., Mural, R. & Xu, Y. (2023), 'New facts and data about professors and their research', *Working Paper* .

Myers, K. & Tham, W. Y. (2023), 'Money, time, and grant design', *Working Paper* .

Nagaraj, A. & Tranchero, M. (2023), 'How does data access shape science? Evidence from the impact of US Census's research data centers on economics research', *Working Paper* .

Nagle, F. & Teodoridis, F. (2020), 'Jack of all trades and master of knowledge: The role of diversification in new distant knowledge integration', *Strategic Management Journal* **41**(1), 55–85.

Nanda, R. & Rhodes-Kropf, M. (2017), 'Financing risk and innovation', *Management Science* **63**(4), 901–918.

Nelson, R. R. (1959), 'The economics of invention: A survey of the literature', *The Journal of Business* **32**(2), 101–127.

Payne, A. A., Siow, A. et al. (2003), 'Does federal research funding increase university research output?', *Advances in Economic Analysis & Policy* **3**(1), 1–24.

Perry, M. & Reny, P. J. (2016), 'How to count citations if you must', *American Economic Review* **106**(9), 2722–2741.

Romer, P. (1990), 'Endogenous Technological Change', *Journal of Political Economy* **98**(5, part 2), S71–S102.

Schumpeter, J. A. (1934), 'The theory of economic development'.

Sidhu, J. S., Commandeur, H. R. & Volberda, H. W. (2007), 'The multifaceted nature of exploration and exploitation: Value of supply, demand, and spatial search for innovation', *Organization Science* **18**(1), 20–38.

Singh, J. & Marx, M. (2013), 'Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity', *Management Science* **59**(9), 2056–2078.

Stephan, P. E. (1996), 'The economics of science', *Journal of Economic Literature* **34**(3), 1199–1235.

Thompson, N. C., Ge, S. & Manso, G. F. (2022), 'The importance of (exponentially more) computing power', *Working Paper* .

Tian, T. Y. & Smith, E. B. (2024), 'Stretched thin: How a misalignment between allocation and valuation underlies the paradox of diversity achievement in higher education', *Administrative Science Quarterly* **69**(3), 711–746.

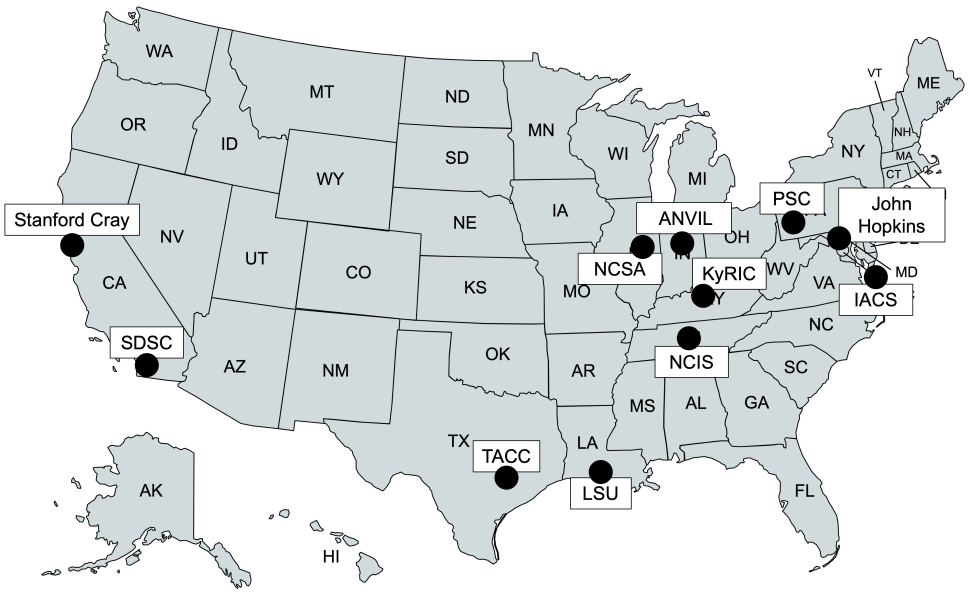
Tranchero, M. (2023), 'Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation', *Working Paper* .

Zhou, K. Z. & Li, C. B. (2012), 'How knowledge affects radical innovation: Knowledge base, market knowledge acquisition, and internal knowledge sharing', *Strategic Management Journal* **33**(9), 1090–1102.

Zhuo, R. (2022), 'Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs', *Working Paper* .

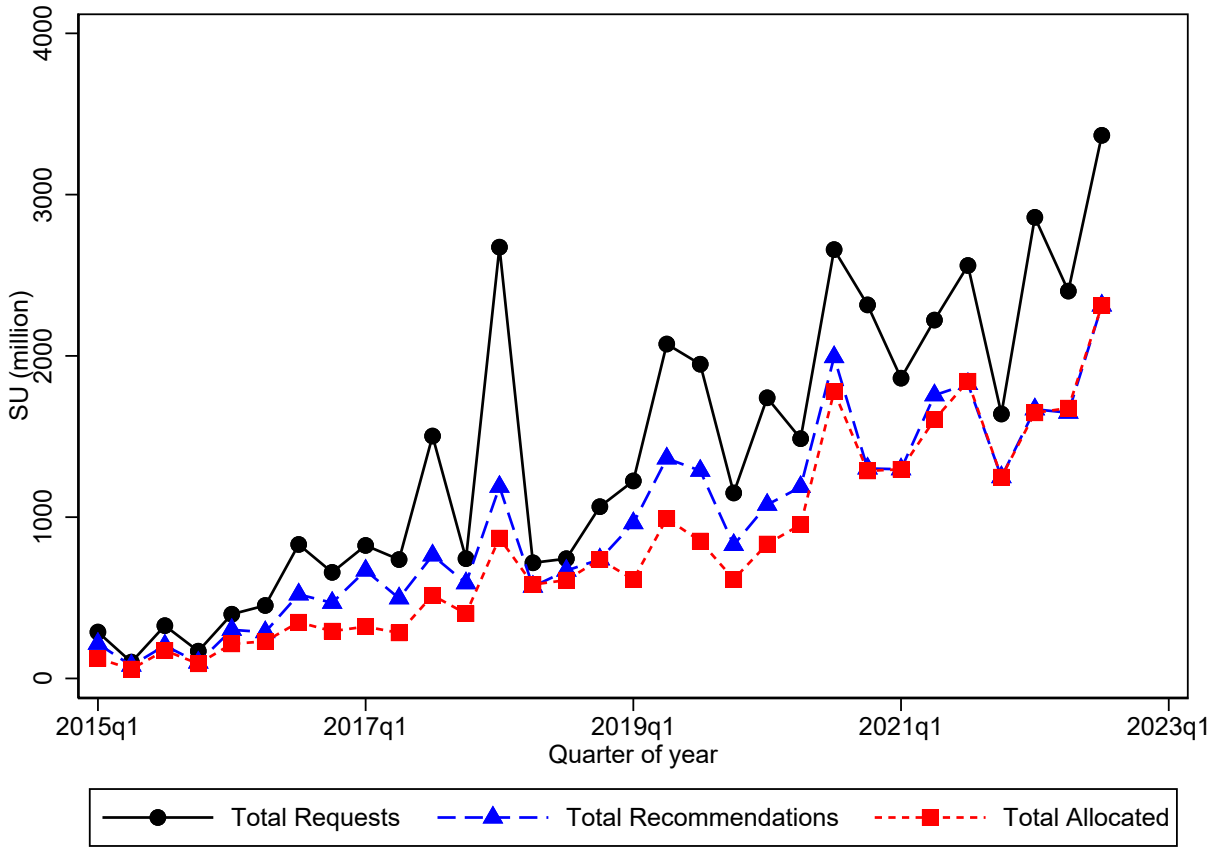
Figures

Figure 1: Location of XSEDE resources



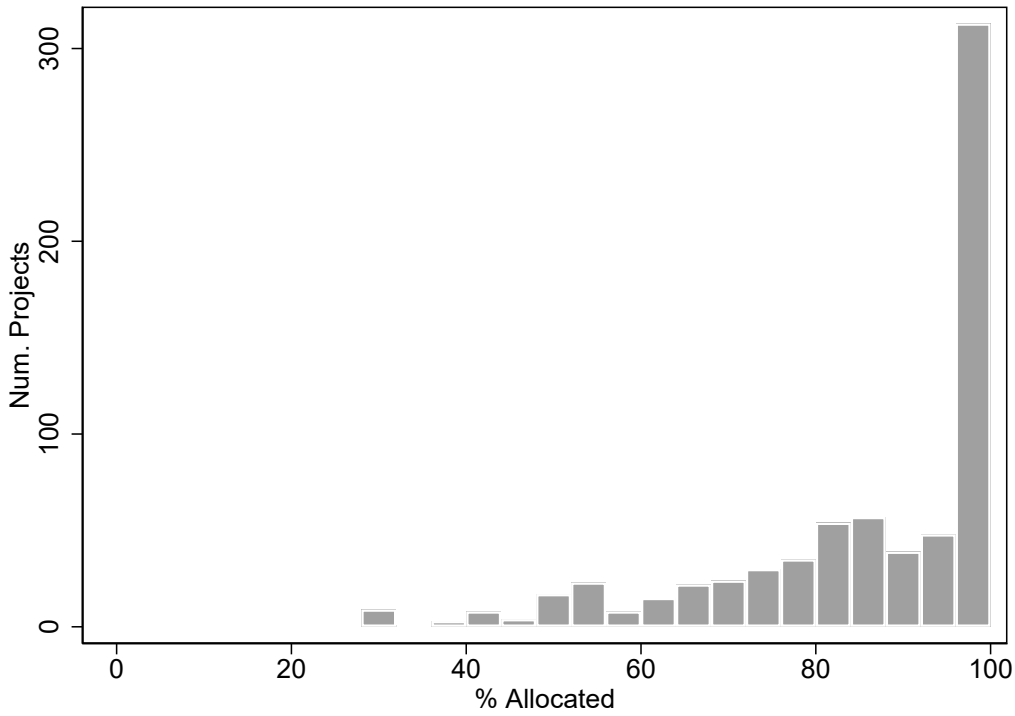
Notes: This figure shows the location of the main supercomputers that are associated with XSEDE.

Figure 2: Total Request, Recommendation and Supply Over Time



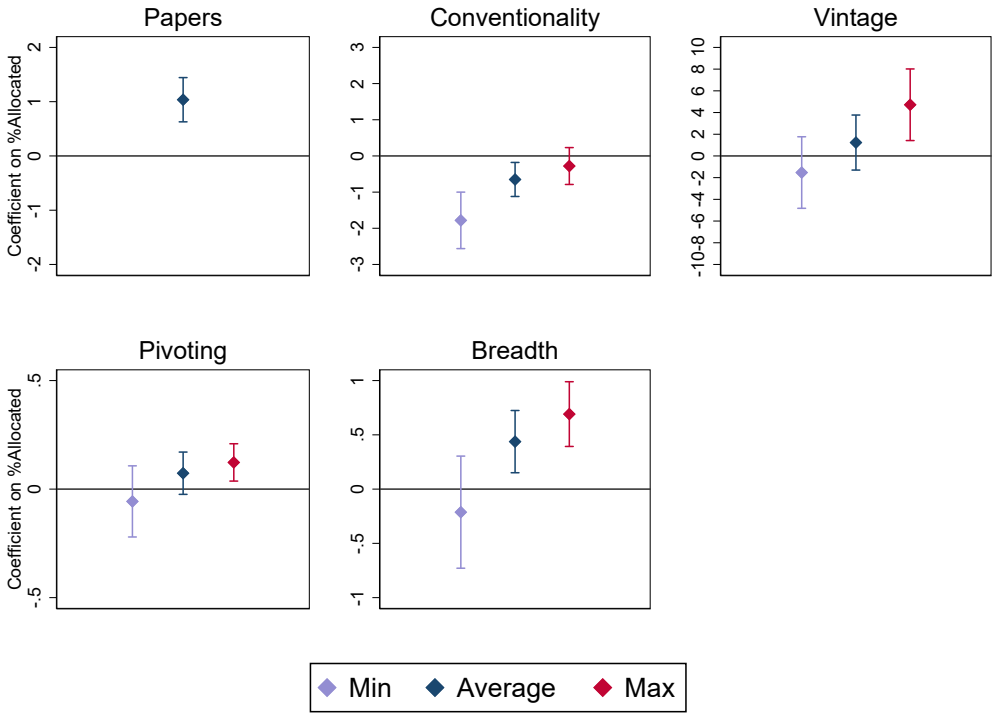
Notes: The x-axis shows each quarter between 2015 and 2022. 'Total Requests' represents the total amount of requested resources, 'Total Recommendations' represents the total amount of recommendations made by the XRAC committee, and 'Total Allocated' represents the total amount of allocated resources after reconciliation of supply and demand. Series show aggregate demand across all XSEDE machines. Several resources (Indiana Jetstream, Kentucky Research Informatics, PSC Bridges-2, Purdue Anvil) were launched in 2021, leaving the XSEDE supply less constrained and explaining the near-overlap between the 'Total Recommendations' and the 'Total Allocated' in the latest quarters.

Figure 3: Histogram of the share of resources allocated



Notes: This figure reports the distribution of the treatment variable %*Allocated*, which represents the ratio of computing resources allocated to a grant compared to the amount of resources recommended by the XRAC.

Figure 4: Coefficient Estimates - Summary



Notes: Each point represents the coefficient estimate on *%Allocated*. Regressions are run using the specification of Equation 1. Each dependent variable used is first calculated at the paper level and is described in Section 6. 'Average' takes the average among papers within grant. 'Min' takes the minimum among papers within grants. 'Max' takes the maximum among papers within grants.

Tables

Table 1: Summary statistics

	Mean	SD	Min	Max
Requested (Mil. SU)	61.6	144.3	0.0	1627.6
Recommended (Mil. SU)	41.7	91.2	0.1	930.7
Approved (Mil. SU)	35.8	80.2	0.1	862.9
% Allocated	0.9	0.2	0.3	1.0
Grant start year	2017.9	2.0	2015.0	2022.0
%NSF Funding	0.4	0.5	0.0	1.0
R1 universities	0.8	0.4	0.0	1.0
Junior PI	0.2	0.4	0.0	1.0
Number of papers	6.3	9.2	1.0	97.0
log(use)	1.0	0.8	0.0	11.8
Mean concept year	2004.4	4.7	1980.0	2023.0
Share of new concepts (2yr)	0.8	0.2	0.0	1.0
Average number of unique concepts	24.8	9.7	1.0	66.7
Observations	710			

Notes: All summary statistics are at the grant level. *Log (use)* is defined as the average number of times concepts used in papers associated with a grant are used by other researchers and is described in greater detail in section 6.3.1. *Mean concept year* is defined as the average year in which concepts used in papers associated with a grant are first observed and is described in greater detail in section 6.3.1. *Share of new concepts (2yr)* is constructed by calculating for each paper the share of concepts that have not been used by the focal PI in the two years preceding their first entry into XSEDE and is described in greater detail in section 6.3.2. *Average number of unique concepts* is the average number of unique concepts used in papers associated with a grant and is described in greater detail in section 6.3.3.

Table 2: Quantity of papers

	DV = ln(papers)					DV = Num. Papers
	(1)	(2)	(3)	(4)	(5)	(6)
$\%Allocated_i$	0.532**	0.962***	0.940***	0.725***	1.036***	0.789**
	(0.209)	(0.199)	(0.208)	(0.217)	(0.246)	(0.365)
$\ln(Recommended_i)$		0.277***	0.275***	0.283***	0.285***	0.315***
		(0.022)	(0.023)	(0.027)	(0.036)	(0.044)
$\%\text{NSF } Funding_i$			0.038	0.030	-0.072	-0.043
			(0.075)	(0.085)	(0.104)	(0.152)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE				Yes	Yes	Yes
Field × Year FE					Yes	Yes
Specification	OLS	OLS	OLS	OLS	OLS	Poisson
Dep Var Mean	1.29	1.29	1.29	1.31	1.35	6.58
R-Sq	0.05	0.24	0.24	0.36	0.46	
Observations	710	710	710	710	710	710

Notes: The unit of observation is a grant. In Columns (1)-(5), the outcome is the log number of papers associated with a grant. Column (6) shows a Poisson specification where the outcome is the number of papers associated with a grant. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects. Columns (4) to (6) add grant field fixed effects. Columns (5) and (6) add grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Less traveled research avenues

	DV = log(Use)			DV = Concept Vintage		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Least Conv.	Most Conv.	Avg. Paper	Oldest	Newest
$\%Allocated_i$	-0.650**	-1.781***	-0.279	1.234	-1.529	4.717**
	(0.284)	(0.470)	(0.308)	(1.531)	(1.988)	(1.991)
$\ln(Recommended_i)$	0.054	-0.134**	0.154***	-0.360**	-1.278***	0.443**
	(0.048)	(0.059)	(0.051)	(0.179)	(0.244)	(0.219)
%NSF $Funding_i$	-0.022	-0.087	-0.050	-0.615	-0.626	-1.016
	(0.118)	(0.146)	(0.135)	(0.470)	(0.680)	(0.675)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	-0.41	-1.27	0.07	2003.85	2000.22	2007.54
R-Sq	0.67	0.55	0.63	0.44	0.44	0.39
Observations	613	529	613	665	665	665

Notes: The unit of observation is a grant. In Columns (1) to (3), we first calculate for each paper the average number of times the concepts it entails have been used by other PIs in the same field. Column (1) takes the mean of this value at the grant level, logged. Column (2) takes the min of this value at the grant level, logged. Column (3) takes the max of this value at the grant level, logged. In Columns (4) to (6), we calculate for each concept associated with a paper the first year we see it appearing and take the average at the paper level. Column (4) averages this measure at the grant level. Column (5) takes the min of this measure at the grant level. Column (6) takes the max of this measure at the grant level. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Topics beyond expertise

	DV = Share of new concepts (vs prev. 2y)		
	(1)	(2)	(3)
Paper w/i grant:	Avg. Paper	Lowest Share New	Highest Share New
$\%Allocated_i$	0.073	-0.057	0.123**
	(0.059)	(0.099)	(0.052)
$\ln(Recommended_i)$	-0.002	-0.043***	0.014**
	(0.008)	(0.015)	(0.007)
%NSF $Funding_i$	0.006	0.021	-0.006
	(0.026)	(0.044)	(0.023)
Year FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes
Dep Var Mean	0.78	0.63	0.88
R-Sq	0.45	0.37	0.42
Observations	600	600	600

Notes: The unit of observation is a grant. For each paper, we calculate the share of 'new-to-the-researcher' concepts, i.e., concepts that have not been previously used by the PI in the 2 years that precede XSEDE entry. Column (1) takes the average of this measure at the grant level. Column (2) takes the min of this measure at the grant level. Column (3) takes the max of this measure at the grant level. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Research Breadth

	DV = log(Unique Concepts)		
	(1)	(2)	(3)
Paper w/i grant:	Avg. Paper	Narrowest Paper	Broadest Paper
$\%Allocated_i$	0.438**	-0.213	0.691***
	(0.173)	(0.312)	(0.180)
$\ln(Recommended_i)$	-0.009	-0.258***	0.047*
	(0.023)	(0.047)	(0.024)
%NSF $Funding_i$	0.037	0.150	0.015
	(0.046)	(0.125)	(0.053)
Year FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes
Dep Var Mean	3.09	2.32	3.37
R-Sq	0.35	0.41	0.36
Observations	653	653	653

Notes: The unit of observation is a grant. We calculate for each paper the number of unique concepts it entails. Column (1) averages this measure at the grant level, logged. Column (2) takes the min of this measure at the grant level, logged. Column (3) takes the max of this measure at the grant level, logged. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Journal quality and number of normalized citations

	DV = log(SNIP)			DV = log(Norm. Cites)		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Lowest SNIP	Highest SNIP	Avg. Paper	Least Cited	Most Cited
$\%Allocated_i$	-0.111	-0.256	-0.019	-0.261	-0.642***	0.115
	(0.154)	(0.176)	(0.206)	(0.221)	(0.214)	(0.274)
$\ln(Recommended_i)$	0.040**	-0.025	0.107***	0.107***	-0.020	0.208***
	(0.016)	(0.017)	(0.023)	(0.022)	(0.025)	(0.027)
%NSF $Funding_i$	-0.088*	-0.066	-0.104*	-0.011	0.004	-0.006
	(0.050)	(0.061)	(0.062)	(0.084)	(0.084)	(0.097)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes			
Field × Year FE	Yes	Yes	Yes			
Dep Var Mean	-0.12	-0.48	0.20	1.64	0.94	2.10
R-Sq	0.29	0.32	0.33	0.05	0.03	0.10
Observations	681	681	681	618	618	618

Notes: The unit of observation is a grant. In Column (1), the outcome is the average SNIP value of papers associated with a grant, logged. In Column (2), the outcome is the minimum SNIP value of papers associated with a grant, logged. In Column (3), the outcome is the maximum SNIP value of papers associated with a grant, logged. In Column (4), the outcome is the average number of citations associated with a grant, logged. In Column (5), the outcome is the number of citations received by the least cited paper associated with a grant, logged. In Column (6), the outcome is the number of citations received by the most cited paper associated with a grant, logged. Citations associated with each paper are normalized to account for differences across fields and publication year. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. Columns (1) to (3) control for grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Columns (4) to (6) control for grant start year fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX

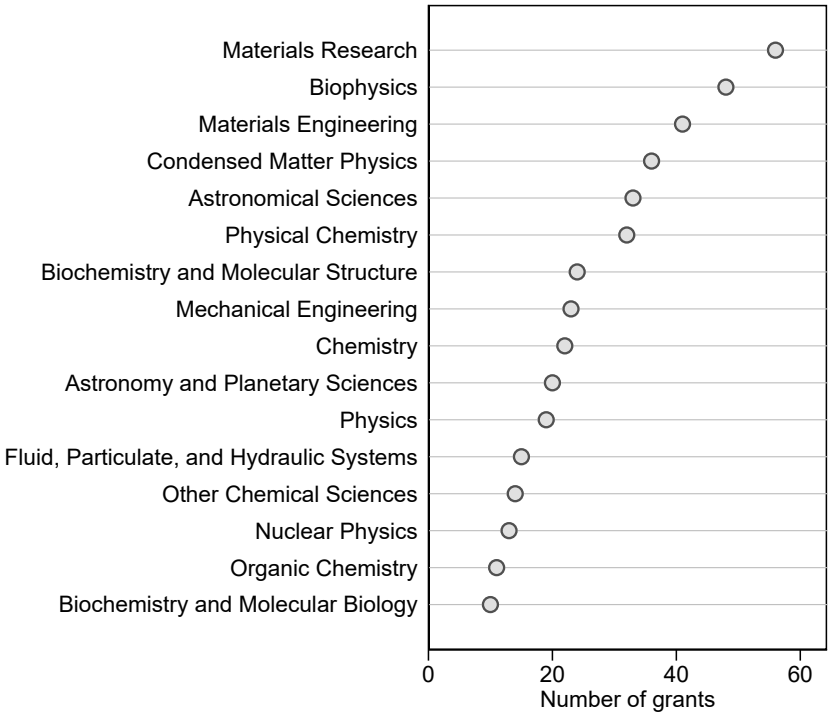
A Figures

Figure A1: XSEDE Resource Allocation Process



Notes: This figure summarizes the resource allocation process of XSEDE.

Figure A2: Most common fields of science (top 15)



Notes: This figure reports the number of grants per field of science for the most common 15 fields of science.

B Tables: Data construction and Empirical strategy

Table A1: Ten Most Used XSEDE Supercomputers

Supercomputer Name	SU (Millions)
TACC Dell/Intel Knights Landing (Stampede2)	27,454
SDSC Dell Cluster (Comet)	10,327
SDSC Expanse CPU	9,306
PSC Bridges-2 (Bridges-2)	8,493
TACC Dell PowerEdge (Stampede)	6,320
PSC Regular Memory (Bridges)	6,079
Purdue Anvil CPU	3,674
NCSA Delta GPU (Delta GPU)	2,340
IU/TACC (Jetstream)	2,264
PSC Bridges-2 GPU (Bridges-2 GPU)	1,965

Notes: This table shows the 10 most used supercomputers over the period 2015-2022. Usage is reported in Service Units (SU) which allows comparison across supercomputers.

Table A2: Sample construction

Sample	Observations
Grants	1,062
Grants with identifiable PIs	950
Grants with a 1:1 correspondence with their PIs	949
Grants with at least one associated paper	719
Grants with a ratio ≤ 1	710

Notes: This table shows the successive steps that lead to the construction of the final sample with 710 observations.

Table A3: Balance test

	DV = %Allocated				
	(1)	(2)	(3)	(4)	(5)
Junior PI	-0.002				-0.011
	(0.017)				(0.016)
R1 Univ		-0.021			-0.009
		(0.020)			(0.020)
Nb papers pre-XSEDE			-0.000		-0.000
			(0.000)		(0.000)
Nb cites pre-XSEDE				-0.000	-0.000
				(0.000)	(0.000)
%NSF $Funding_i$	0.100***	0.096***	0.101***	0.100***	0.100***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)
ln($Recommended_i$)	-0.026***	-0.027***	-0.025***	-0.025***	-0.024***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Field FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	0.85	0.85	0.85	0.85	0.85
R-Sq	0.51	0.49	0.52	0.52	0.52
Observations	686	710	675	675	675

Notes: The unit of observation is a researcher. *%Allocated* is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. *Junior* is an indicator equal to 1 the PI has less than 8 years of experience. *R1* is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Nb papers pre-XSEDE* is the number of publications published by the PI before entering XSEDE. *Nb cites pre-XSEDE* is the number of citations received by the PI for publications published before entering XSEDE. All columns control grant start year fixed effects, field fixed effects, and grant start year × field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Over-requesting test

	DV = $\%Allocated$
	(1)
$\%Over_request$	-0.001
	(0.001)
%NSF $Funding_i$	0.095***
	(0.017)
ln($Recommended_i$)	-0.028***
	(0.005)
Year FE	Yes
Field FE	Yes
Field × Year FE	Yes
Dep Var Mean	0.85
R-Sq	0.49
Observations	710

Notes: The unit of observation is a grant. For each grant, $\%Over_request$ is equal to $(Requested_i - Recommended_i)/Recommended_i$, which we use as a proxy for the share of over-requested resources. $\%NSF\ Funding$ is the share of funding coming from the NSF. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. We include grant start year fixed effects, grant field fixed effects and grant start year × field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Matched vs Unmatched Grants

	DV = 1{In-sample}
	(1)
$\%Allocated_i$	-0.002
	(0.001)
%NSF $Funding_i$	0.077**
	(0.034)
$\ln(Recommended_i)$	0.022*
	(0.013)
Field FE	Yes
Year FE	Yes
Field × Year FE	Yes
Dep Var Mean	0.67
R-Sq	0.38
Observations	1062

Notes: The unit of observation is a grant. The dependent variable is an indicator equal to 1 if the grant is part of our sample of 710 grants and 0 otherwise. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. We control for grant start year fixed effects, grant field fixed effects grant start year × field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

C Tables: Robustness

Table A6: Less traveled research avenues - Robustness

	DV = 1{Use ≥ Median}		
	(1)	(2)	(3)
Paper w/i grant:	Avg. Paper	Least Conv.	Most Conv.
$\%Allocated_i$	-0.282**	-0.578***	-0.134
	(0.110)	(0.149)	(0.120)
$\ln(Recommended_i)$	0.031*	-0.059***	0.057***
	(0.017)	(0.019)	(0.017)
%NSF $Funding_i$	-0.011	0.024	0.038
	(0.044)	(0.052)	(0.048)
Year FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes
Dep Var Mean	0.62	0.58	0.62
R-Sq	0.51	0.42	0.49
Observations	653	653	653

Notes: The unit of observation is a grant. We first calculate for each paper the average number of times the concepts it entails have been used by other PIs in the same field. We then calculate the mean of this measure at the grant level (mean use of papers within a grant), the minimum of this measure at the grant level (use of concepts by the least conventional paper), and the maximum of this measure at the grant level (use of concepts by the most conventional paper). For each of these grant-level variables, we calculate the median across all grants and create an indicator equal to 1 if the value of the variable for the focal grant is above this median. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Less traveled research avenues - Trend over time

	DV = log(Use Prior)			DV = log(Use After)		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Least Conv.	Most Conv.	Avg. Paper	Least Conv.	Most Conv.
$\%Allocated_i$	-0.616**	-1.640***	-0.217	-0.759***	-2.281***	-0.285
	(0.295)	(0.464)	(0.317)	(0.283)	(0.474)	(0.293)
$\ln(Recommended_i)$	0.034	-0.120**	0.132***	0.053	-0.034	0.170***
	(0.044)	(0.054)	(0.047)	(0.051)	(0.072)	(0.051)
%NSF $Funding_i$	0.050	-0.018	0.024	-0.055	0.026	-0.099
	(0.101)	(0.145)	(0.121)	(0.132)	(0.146)	(0.146)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	-0.38	-1.20	0.10	-0.50	-0.99	0.04
R-Sq	0.68	0.57	0.64	0.69	0.63	0.65
Observations	609	512	609	585	416	585

Notes: The unit of observation is a grant. In Columns (1) to (3), we first calculate for each paper the average number of times the concepts it entails have been used by other PIs in the same field *prior* to their entry into XSEDE. Column (1) takes the mean of this value at the grant level, logged. Column (2) takes the min of this value at the grant level, logged. Column (3) takes the max of this value at the grant level, logged. Columns (4) to (6) repeat this exercise using the average number of times concepts used in a paper have been used by other PIs in the same field *after* their entry into XSEDE. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\ln(Recommended)$ is the log of the number of SUs recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Topics beyond expertise - Robustness

	DV = Share of new concepts (vs prev. 3y)			DV = Share of new concepts (vs prev. 5y)		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Lowest Share New	Highest Share New	Avg. Paper	Lowest Share New	Highest Share New
$\%Allocated_i$	0.101*	-0.010	0.147***	0.091	-0.028	0.145***
	(0.060)	(0.108)	(0.049)	(0.063)	(0.107)	(0.052)
$\ln(Recommended_i)$	-0.001	-0.042***	0.016**	-0.004	-0.045***	0.016**
	(0.008)	(0.014)	(0.007)	(0.008)	(0.014)	(0.007)
%NSF $Funding_i$	-0.000	0.011	-0.016	-0.001	0.013	-0.017
	(0.025)	(0.041)	(0.023)	(0.024)	(0.040)	(0.022)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	0.76	0.61	0.87	0.74	0.59	0.86
R-Sq	0.45	0.37	0.43	0.45	0.37	0.44
Observations	608	608	608	617	617	617

Notes: The unit of observation is a grant. For each paper, we calculate the share of 'new-to-the-researcher' concepts, i.e., concepts that have not been previously used by the PI. Columns (1) to (3) compare XSEDE papers with papers written in the 3 years that precede XSEDE entry. Columns (4) to (6) compare XSEDE papers with papers written in the 5 years that precede XSEDE entry. Columns (1) and (4) take the average of this measure at the grant level. Columns (2) and (5) take the min of this measure at the grant level. Columns (3) and (6) take the max of this measure at the grant level. $\%Allocated$ is the share of resources allocated to a grant relative to the amount of resources recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D Tables: Citations and Research Direction

Table A9: Normalized citations and Type of science

	log(Norm. Cites)			
	(1)	(2)	(3)	(4)
log(Use)	0.080**			
	(0.033)			
Concept Year		-0.020**		
		(0.008)		
Share New (2y)			-0.663***	
			(0.171)	
log(Unique Concepts)				0.234***
				(0.074)
ln(*Recommended*$_i$)	0.078***	0.095***	0.094***	0.100***
	(0.027)	(0.024)	(0.025)	(0.024)
%NSF *Funding*$_i$	-0.021	-0.022	0.003	-0.061
	(0.086)	(0.080)	(0.080)	(0.082)
Year FE	Yes	Yes	Yes	Yes
Dep Var Mean	1.67	1.65	1.66	1.65
R-Sq	0.06	0.06	0.12	0.06
Observations	542	585	524	575

Notes: The unit of observation is a grant. The outcome is the average number of citations associated with a grant, logged. Citations associated with each paper are normalized to account for differences across fields and publication year. To calculate *Log(Use)*, we measure the average number of times concepts used in a paper are used by other PIs and take the mean of this value at the grant level, logged. To calculate *Concept Year*, we calculate for each paper associated with a grant the average year we observe the concepts it entails for the first time and average this measure at the grant level. To calculate *Share New (2y)*, we calculate for each paper the share of concepts that have not been used by the PI in papers written in the 2 years that precede XSEDE entry and average this measure at the grant level. To calculate *Log(Unique Concepts)*, we calculate for each paper the number of unique concepts it entails and average this measure at the grant level, logged. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. All columns control for grant start year fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

E Tables: Main analyses using allocation levels

Table A10: Quantity of papers

	DV = ln(papers)					DV = Num. Papers
	(1)	(2)	(3)	(4)	(5)	(6)
ln(*Allocated*$_i$)	0.275***	0.683***	0.669***	0.537***	0.730***	0.583**
	(0.022)	(0.136)	(0.141)	(0.150)	(0.163)	(0.240)
ln(*Recommended*$_i$)		-0.410***	-0.398***	-0.258*	-0.451***	-0.271
		(0.135)	(0.139)	(0.149)	(0.162)	(0.225)
%NSF *Funding*$_i$			0.034	0.026	-0.076	-0.048
			(0.074)	(0.084)	(0.103)	(0.151)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE				Yes	Yes	Yes
Field × Year FE					Yes	Yes
Specification	OLS	OLS	OLS	OLS	OLS	Poisson
Dep Var Mean	1.29	1.29	1.29	1.31	1.35	6.58
R-Sq	0.23	0.24	0.24	0.36	0.46	
Observations	710	710	710	710	710	710

Notes: The unit of observation is a grant. The outcome is the log number of papers associated with a grant. ln(*Allocated*) is the log of the number of SUs allocated. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. We include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Less traveled research avenues

	DV = log(Use)			DV = Concept Vintage		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Least Conv.	Most Conv.	Avg. Paper	Oldest	Newest
ln($Allocated_i$)	-0.413**	-1.226***	-0.167	0.967	-1.001	3.285**
	(0.192)	(0.311)	(0.204)	(0.998)	(1.342)	(1.296)
ln($Recommended_i$)	0.472**	1.102***	0.324	-1.332	-0.266	-2.869**
	(0.201)	(0.315)	(0.215)	(1.014)	(1.353)	(1.291)
%NSF $Funding_i$	-0.026	-0.083	-0.053	-0.632	-0.632	-1.027
	(0.120)	(0.145)	(0.137)	(0.473)	(0.685)	(0.673)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	-0.41	-1.27	0.07	2003.85	2000.22	2007.54
R-Sq	0.67	0.55	0.63	0.44	0.44	0.39
Observations	613	529	613	665	665	665

Notes: The unit of observation is a grant. In Columns (1) to (3), we first calculate for each paper the average number of times the concepts it entails have been used by other PIs in the same field. Column (1) takes the mean of this value at the grant level, logged. Column (2) takes the min of this value at the grant level, logged. Column (3) takes the max of this value at the grant level, logged. In Columns (4) to (6), we calculate for each concept associated with a paper the first year we see it appearing and take the average at the paper level. Column (4) averages this measure at the grant level. Column (5) takes the min of this measure at the grant level. Column (6) takes the max of this measure at the grant level. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. All columns include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Topics beyond expertise

	DV = Share of new concepts (vs prev. 2y)		
	(1)	(2)	(3)
Paper w/i grant:	Avg. Paper	Lowest Share New	Highest Share New
$\ln(Allocated_i)$	0.048	-0.040	0.083**
	(0.038)	(0.063)	(0.034)
$\ln(Recommended_i)$	-0.050	-0.003	-0.069**
	(0.039)	(0.064)	(0.035)
%NSF $Funding_i$	0.007	0.022	-0.006
	(0.026)	(0.044)	(0.023)
Year FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes
Dep Var Mean	0.78	0.63	0.88
R-Sq	0.45	0.37	0.42
Observations	600	600	600

Notes: The unit of observation is a grant. For each paper, we calculate the share of 'new-to-the-researcher' concepts, i.e., concepts that have not been previously used by the PI in the 2 years that precede XSEDE entry. Column (1) takes the average of this measure at the grant level. Column (2) takes the min of this measure at the grant level. Column (3) takes the max of this measure at the grant level. ln(*Allocated*) is the log of the number of SUs allocated. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Topics beyond expertise - alternative time windows

	DV = Share of new concepts (vs prev. 3y)			DV = Share of new concepts (vs prev. 5y)		
	(1)	(2)	(3)	(4)	(5)	(6)
Paper w/i grant:	Avg. Paper	Lowest Share New	Highest Share New	Avg. Paper	Lowest Share New	Highest Share New
ln(*Allocated*$_i$)	0.064	-0.010	0.096***	0.059	-0.020	0.097***
	(0.040)	(0.072)	(0.033)	(0.042)	(0.072)	(0.034)
ln(*Recommended*$_i$)	-0.066	-0.032	-0.081**	-0.064	-0.025	-0.081**
	(0.040)	(0.072)	(0.033)	(0.042)	(0.071)	(0.035)
%NSF *Funding*$_i$	0.000	0.011	-0.016	-0.001	0.014	-0.016
	(0.025)	(0.041)	(0.023)	(0.024)	(0.040)	(0.022)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Dep Var Mean	0.76	0.61	0.87	0.74	0.59	0.86
R-Sq	0.45	0.37	0.43	0.45	0.37	0.44
Observations	608	608	608	617	617	617

Notes: The unit of observation is a grant. For each paper, we calculate the share of 'new-to-the-researcher' concepts, i.e., concepts that have not been previously used by the PI in the years that precede XSEDE entry. Columns (1) to (3) compare XSEDE papers with papers written in the 3 years that precede XSEDE entry. Columns (4) to (6) compare XSEDE papers with papers written in the 5 years that precede XSEDE entry. Columns (1) and (4) take the average of this measure at the grant level. Columns (2) and (5) take the min of this measure at the grant level. Columns (3) and (6) take the max of this measure at the grant level. ln(*Allocated*) is the log of the number of SUs allocated. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Research Breadth

	DV = log(Unique Concepts)		
	(1)	(2)	(3)
Paper w/i grant:	Avg. Paper	Narrowest Paper	Broadest Paper
ln(*Allocated*$_i$)	0.267**	-0.190	0.435***
	(0.114)	(0.199)	(0.125)
ln(*Recommended*$_i$)	-0.279**	-0.068	-0.394***
	(0.113)	(0.196)	(0.125)
%NSF *Funding*$_i$	0.042	0.157	0.020
	(0.046)	(0.124)	(0.054)
Year FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Field × Year FE	Yes	Yes	Yes
Dep Var Mean	3.09	2.32	3.37
R-Sq	0.35	0.41	0.36
Observations	653	653	653

Notes: The unit of observation is a grant. We calculate for each paper the number of unique concepts it entails. Column (1) averages this measure at the grant level, logged. Column (2) takes the min of this measure at the grant level, logged. Column (3) takes the max of this measure at the grant level, logged. ln(*Allocated*) is the log of the number of SUs allocated. ln(*Recommended*) is the log of the number of SUs recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. All columns include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

F Figures: Non-parametric specification

To construct the following figures, we bin the *% Allocated* variable into 5 quantiles: below the 10^{th} percentile, between the 10^{th} percentile and the 25^{th} percentile, between the 25^{th} percentile and the 50^{th} percentile, between the 50^{th} percentile and the 75^{th} percentile and above the 75^{th} percentile. We run similar regressions as Equation 1 using these indicators as main variables of interest and plot the coefficient estimates we obtain. The dependent variables are the same as described in the text.

Figure A3: Papers

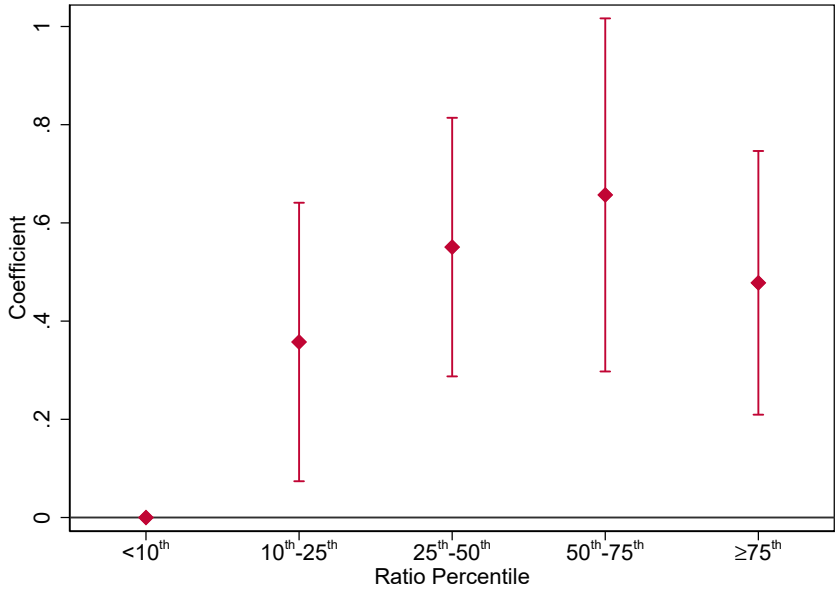


Figure A4: Use of concepts by others

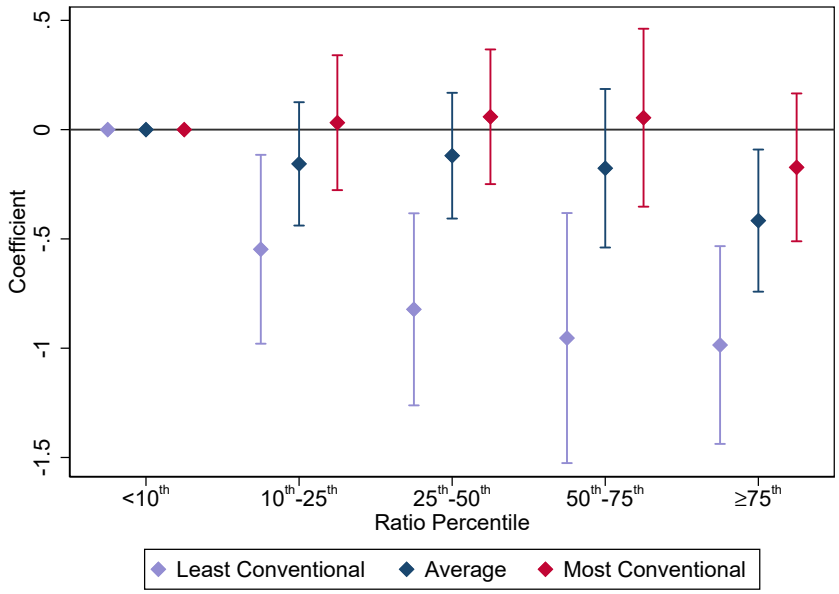


Figure A5: Concept Vintage

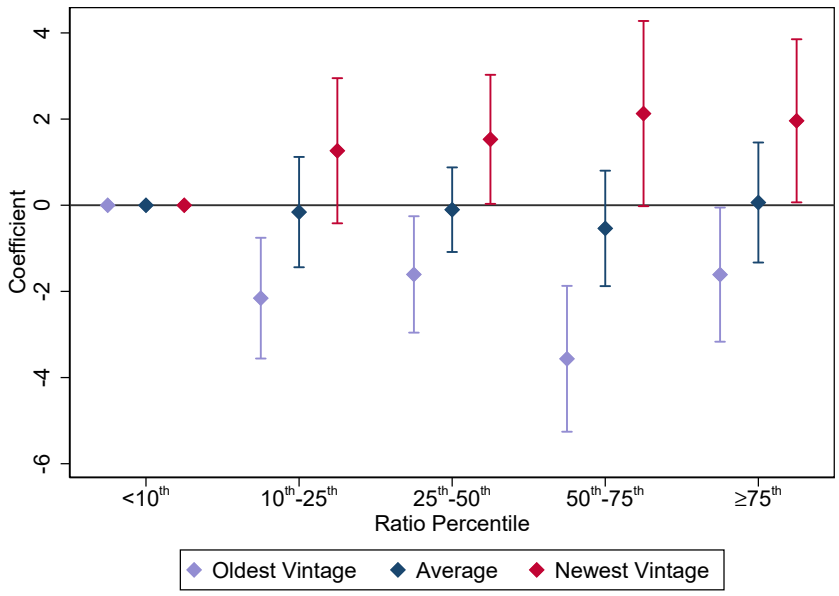


Figure A6: Share of new concepts

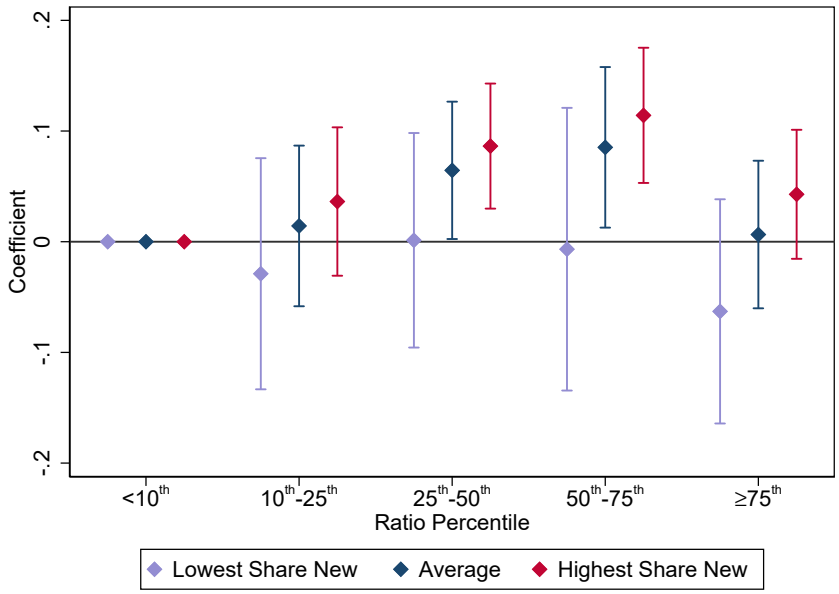


Figure A7: Research breadth

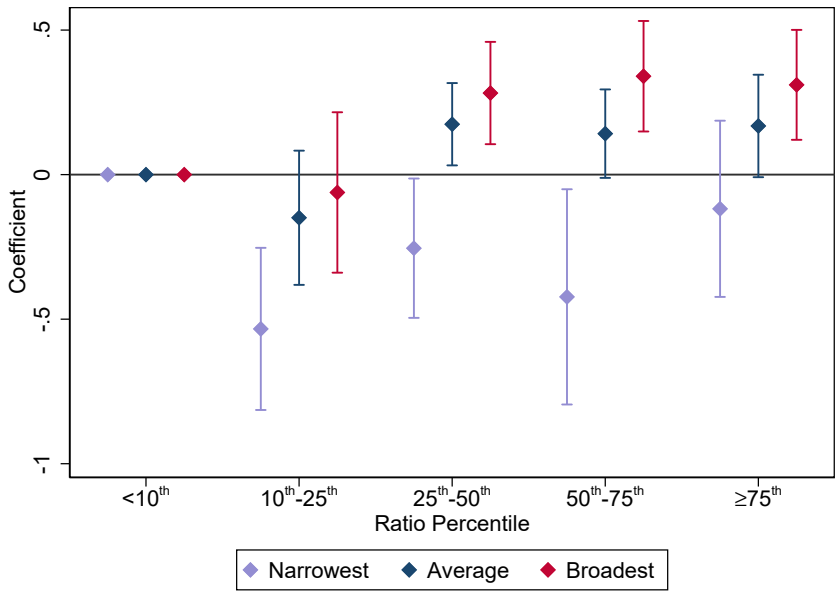


Figure A8: SNIP

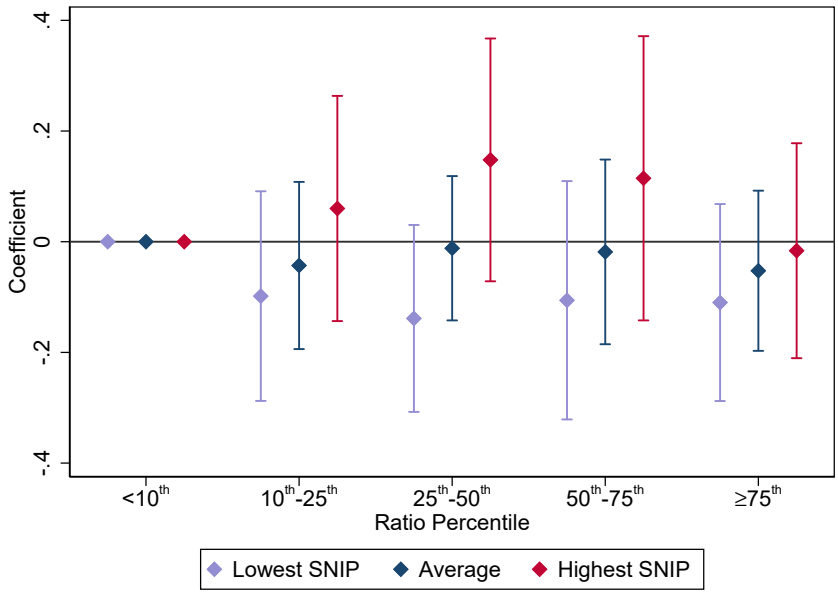
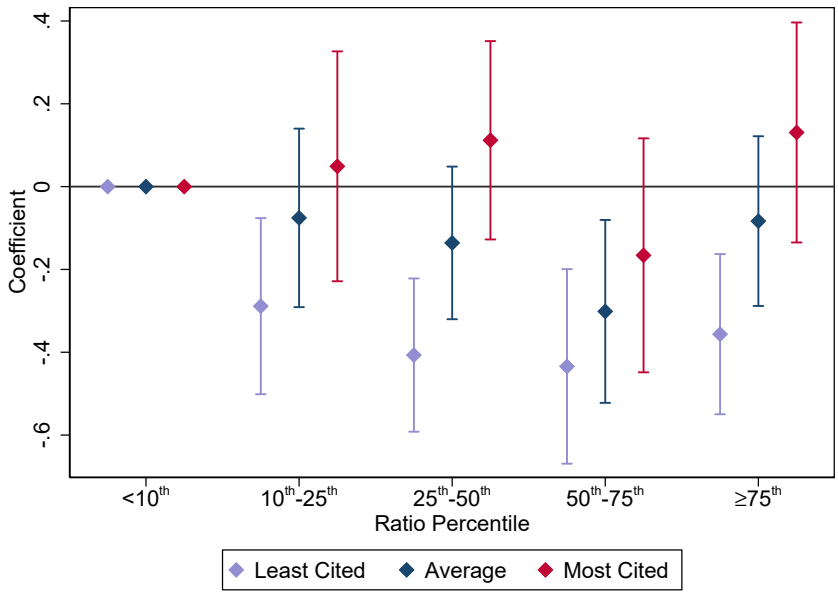


Figure A9: Citations



G Proposal Example

This is an example of a proposal in the field of Biophysics.[34]

G.1 Summary of Proposed Research

We are requesting continued use of XSEDE resources to support four independent projects that each utilize conventional and enhanced sampling molecular dynamics (MD) simulations. Specifically, these projects are:

1. The nucleosome core particle (NCP) consists of ∼147 base pairs of DNA wrapped around a histone octamer and is the fundamental unit of DNA packaging in eukaryotes. The substitution of variant proteins to the octamer has been linked to epigentic regulation in vivo, such as the increased presence of H2A.B in transcriptionally active regions of DNA. While several studies have highlighted potential mechanisms, a fully atomistic understanding of H2A.B-regulated NCP destabilization has not been described. Here, we propose a combination of conventional and free energy MD simulations to study the mechanisms of epigenetic activation by H2A.B substitution.

2. The chromatosome is an extension of the NCP which binds an additional ∼20 base pairs of DNA due to the presence of a linker histone. Experiments have shown that linker histones can bind to DNA in either an on- or off-dyad locations, the selection of which is dictated by modifications to the linker histone sequence. However, the molecular mechanisms that drive the binding mode, as well as how they are influenced by specific linker histone residues, remains poorly understood. Here, we will perform a series of conventional MD simulations of three different linker histones bound in on- and off-dyad locations to determine how specific residues influence the structure and energetics of the chromatosome.

3. Chromatin fibers consist of repeating NCP units and form the building blocks of chromosomes within the nucleus of eukaryotic cells. Therefore, the structure and dynamics of poly-NCP arrays directly influences several processes, including gene expression, DNA replication, and DNA repair. In this project, we aim to study the relationship between structure, function, and dynamics in poly-NCP arrays. More concretely, we will perform fully atomistic MD simulations of an octa-NCP construct in a compact state. Analysis will focus on the stability and conformational dynamics of the chromatin fiber, along with the importance of stacking interactions between adjacent nucleosomes and inter-NCP contacts formed by the histone tails. Furthermore, we will focus on identifying allosteric networks within poly-NCP arrays.

4. Pili are protein structures that protrude from bacterial cell walls and function to enhance their virulence. They are bio-synthesized by the C class sortase proteins, which was the subject of our previous

[34] See https://web.archive.org/web/20210414012547/https://portal.xsede.org/allocations/research#examples for more details, in-text citations from this proposal, and additional examples.

work. Structurally, pili are polymers built by covalently linking various pilin subunits. However, there is little detail in the literature about the molecular structure and dynamics of these building blocks or how their properties are potentially modified due to polymerization. Here, we propose to implement free energy MD simulations on the dimeric form of Streptococcus pneumoniae and Corynebacterium diphtheriae pilins. Simulations will be carried out both in the presence and absence of the experimentally determined intradomain isopeptide bonds that have been proposed to impart mechanical strength to pili.

G.2 Background and Specific Goals

G.2.1 Simulations of the H2A.B Variant Nucleosome (NIH 1R15GM114758)

Eukaryotic cells store their genetic material into chromosomes through a hierarchical packaging scheme whose fundamental unit is the nucleosome core particle (NCP) [1]. The NCP wraps ∼147 base pairs of DNA around a symmetric histone octamer formed by two (H2A-H2B) dimers anking an (H3-H4)2 tetramer [2]. In vivo, NCPs act as epigenetic governors of transcription and alter gene expression through a variety of chemical changes [3, 4]. For example, substitution of canonical histones with a structurally similar but sequentially deviant protein, a histone "variant", can comprehensively modulate gene expression. Our previous work was focused on a transcription repressor, the macroH2A variant, in both homogeneous (both dimers replaced) [5] and heterogeneous (one canonical dimer and one variant dimer) nucleosomes. In contrast, Barr body deficient H2A (H2A.B) is associated with transcriptionally active regions of DNA[6], but little is known about how the structure and dynamics of H2A.B containing NCPs affect their function.

Like other H2A variants, the histone fold of H2A.B is believed to follow the structural motif of three alpha helices connected by short loops and anked by exible N- and C-terminal tails. However, the sequence of this fold is only ∼50% identical to the canonical protein, making H2A.B the most sequentially divergent of the major variants [3]. Indeed, these sequence variations alter important chemical properties of the NCP, such as reducing the total number of lysine residues and eliminating the acidic patch believed to be crucial in internucleosomal interactions. Strikingly, the C-terminal sequence of H2A.B is ∼20 amino acids shorter than the canonical protein, reducing the sequence responsible for forming dimer-tetramer interactions by ∼25%. Furthermore, these C-terminal residues are located near the DNA entry/exit site in canonical H2A. Thus, the lack of C-terminal sequence may destabilize the NCP by degrading both protein-protein and protein-DNA interactions and result in an open, more accessible structure [79].

In the simulations proposed here, we will comprehensively study the eects of H2A.B substitution on NCP dynamics and stability through a combination of conventional and free energy MD simulations. Results from conventional simulations will be compared against our previous work to determine the differences in dynamics between H2A.B and canonical nucleosomes in the hundreds of nanoseconds timescale, and solution ensembles will be compared against experimental scattering profiles [7, 8]. Furthermore, free energy calculations will be

conducted for both the canonical and H2A.B NCPs to quantify the disparity in DNA binding capabilities. The success of these free energy calculations will serve to guide unwinding simulations on future NCP systems, such as other variant NCPs, the chromatosome, and post-translationally modied NCPs. These simulations, along with our previous work, will provide further understanding of the interaction between structure, dynamics, and function in variant-containing nucleosomes.

G.2.2 Simulations of the Chromatosome (NSF 1552743)

The chromatosome is an extension of the NCP that contains the same structural foundations with an additional ∼20 base pairs of DNA (∼167 total) accompanied by a highly-conserved linker histone. Colloquially known as histone H1, this nuclear protein plays a crucial role in the condensation of nucleosome chains into higher order structures [10], as well as other cellular functions [11] such as gene expression [12, 13], heterochromatin genetic activity [14], and cell differentiation [15,16], among many others [1719]. Linker histones have also been found to rapidly exchange locations throughout the chromosome [20, 21]. In avian species, H1 is replaced by the H5 isoform.

Although H1 and H5 have highly similar primary and tertiary structures, recent experiments have shown that they exhibit different binding modes on the nucleosome. Histone H5 binds centered on the dyad axis in the "on-dyad" location, whereas histone H1 binds in a skewed mode relative to the dyad axis, or off-dyad. In a recent study, Zhou et al examined the binding modes of wild type globular histone H1 (GH1), H5 (GH5), and a GH5 penta-mutant that had five residues switched from their GH5 to GH1 sequences (GH5 pMu) [22]. Results showed that these small number of mutations were able to shift the binding state from the H5 on-dyad location to the H1 off-dyad location. This suggests that these five residues may be the key determinants for linker histone binding modes on chromatosomes. However, the molecular mechanisms that orchestrate this conformational shift, the role of specific linker histone residues, and how residue modifications affects this process is poorly understood. Moreover, the extent to which these phenomenon alter greater chromatin dynamics is also unclear.

Here, we will study the atomic scale mechanisms that dictate linker histone binding modes through a series of conventional MD simulations of different chromatosomes containing GH1, GH5, and GH5 pMu. We hypothesize that each linker histone is capable of binding in both the on- and off-dyad locations, and that differences in the histones shift the free energies of the states. Therefore, we will initiate simulations of each system from both the on- and off-dyad locations. Analysis will focus on the structure and energetics of specific contacts that stabilize these two binding modes, with an emphasis on how the mutated residues in GH5 pMu create the experimentally observed shift in binding.

G.2.3 Simulations of the Canonical Octanucleosome (NSF 1552743)

In vitro, nucleosomes form long "beads on a string" like structures in which many nucleosomes are periodically positioned on a long DNA molecule. In its most compact form, poly-NCP arrays condense into structures that are composed of tetra-NCP repeats with DNA that zigzags between the individual units [23]. These tetra-NCP structures likely form the basis of chromatin fibers and thus represent a second level of gene compaction [1, 2426]. What the relationship is between the structure, dynamics, and function of these constructs, and how these properties are modulated by chromatin remodeling factors, are questions of deep scientific interest that have yet to be fully understood.

In previous work on mono-NCPs, we showed that chromatin remodeling factors influence not only the NCP structure, but also the dynamic networks that extend throughout the complex [5]. We hypothesize that the regulation of allosteric networks is a general feature of chromatin remodeling factors, and that histone variants and PTMs modulate both intra- and inter-NCP dynamics as a method of influencing chromatin stability far from their physical location. If true, these networks would likely propagate through the inter-nucleosomal DNA and histone tails, and they would link locations that affect nucleosome stability, such as the DNA entry and exit points, as well as the locations most heavily influenced by remodeling factors, such as the L1 loops and the acidic patch.

In this project, we will begin to test the hypothesis that allosteric networks link regions between nucleosomes that are known to be important for chromatin fiber regulation. To do this, we will perform a set of conventional MD simulations of an octa-NCP system. Analysis will focus on determining the strength and linchpins of inter-NCP and inter-tetraNCP dynamic networks with a combination of methods that we previously applied to mono-NCP systems [5]. In future years, we will build on these results by simulating poly-NCP constructs that include histone variants and PTMs, as well as systems in other, non-compact, conformations. We also aim to use our full atomistic simulations to validate existing Coarse Grained models that have been used to simulate mono-NCPs [27] and their reliability to be used in simulating larger poly-NCP to model realistic chromatin fibers.

G.2.4 Simulations of Pilin Subunits (NIH 1R35GM119647)

Pili, which are elongated protein structures that reside on the cell surface of bacteria, have been shown to increase bacterial virulence. They are formed by sortase C (SrtC) proteins through the polymerizing of pilin sub-units [28]. The main pilin monomer of Streptococcus pneumoniae (RrgB) is made up of four domains as shown in Figure 4. The second and third domains are arranged side to side, while the first and last units lie on top and below these middle units. In the case of Corynebacterium diphtheriae, its main building block (SpaA) is a smaller three-domain protein with domains that are placed on top of each other. Interestingly, each of the domains in both proteins contain isopeptide bonds that have been proposed to provide additional mechanical stability.

Previously, we examined the lid opening process of SrtC, and our results predicted that it likely requires the action of an external factor to function due to the high calculated free energies required for the lid opening process; a result that was later confirmed by NMR dynamics experiments carried out by our collaborators [29]. We further went on to examine the energetics of pilin sorting signals binding to SrtC (Progress Report, section 1.2). For the present project, we aim to understand the structural dynamics of the Streptococcus pneumoniae and Corynebacterium diphtheriae pilin monomers, explain how dimerization affects overall dynamics, and elucidate the molecular reasons for the intra-domain isopeptide bonds observed in these pilin molecules.

G.3 Proposed Simulations and System Details

Here, we describe the specific details of the simulations we will perform in each project. These simulations will use either NAMD, GROMACS, or AMBER [3032], and GPU acceleration will be used whenever feasible [33]. All simulations with explicit solvent environments will be performed with periodic boundary conditions, with non-bonded interactions truncated at 10 A, and long-range electrostatics treated with the particle mesh Ewald (PME) method [34]. Hydrogen containing bonds will be constrained using SHAKE [35] in AMBER and NAMD, and P-LINCS will restrain similar bonds in GROMACS simulations [36]. All simulations are to be conducted with a 2 fs timestep unless otherwise specified.

G.3.1 Simulations of the H2A.B Nucleosome

To better understand the structure-dynamics-function relationship in H2A.B containing NCPs, we are requesting XSEDE resources for conducting two complementary arrays of simulations. In the first, we will conduct conventional MD (cMD) simulations on the H2A.B NCP. In the second, we will calculate the free energy of unwinding of canonical and H2A.B NCPs via umbrella sampling (US) calculations. In both sets of simulations, the protein and nucleic acid forcefield parameters will come from FF14SB, the most recent AMBER xed point-charge forcefield [37, 38], and ions will be paramaterized according to Joung and Cheatham [39].

Table 1: A summary of the total requested resources for simulations relating to the H2A.B NCP. Conventional simulations have a benchmark speed of 300 SUs/ns on Comet with NAMD, whereas umbrella sampling calculations were benchmarked at 5 SUs/ns on XStream with AMBER.

NCP Type	Simulation Method	System Size (atoms)	Simulation Length	Number of Simulations	Comet SUs	XStream SUs
H2A.B	cMD	241,730	250 ns	3	225,000	———
H2A.B	US	24,509	250 ns	36	———	45,000
Canonical	US	25,172	250 ns	36	———	45,000
				Total:	**225,000**	**90,000**

G.3.1.1 Conventional Molecular Dynamics

Since no crystal structure of the H2A.B NCP exists, an initial conformation was constructed from the 1.9

A resolution X-ray structure of the canonical NCP (PDB: 1KX5) [40]. The canonical H2A sequence was truncated and mutated to the sequence of H2A.B in VMD and then neutralized by the addition of Cl- ions. The neutralized H2A.B NCP was then solvated in a box of TIP3P water[41] at a salt concentration of 150 mM NaCl, and the box was extended 10 A from the solute, yielding a system size of 241,730 atoms.

Simulations will be performed in the NPT ensemble at 300 K and 1 atm using the NAMD engine and a 2 fs timestep. Long range correlations, DNA-protein interactions and binding energies, and dimer-tetramer interactions will be analyzed from the resulting trajectories and will be compared against our previous results for the canonical NCP[5]. Furthermore, the H2A.B solution of states will be compared to experimental scattering data [7, 8] using our recently developed in-house protocol (see Progress Report, Section 1.4).

To determine the amount of necessary simulation time, we direct to our previous work on canonical and macroH2A variant NCPs in which our simulations required 50-60 ns in order for the extended histone tails to equilibrate and contract [5]. Therefore, in order to obtain ∼200 ns of production data, a total of 250 ns per simulation must be conducted. Three separate simulations will be employed to ensure meaningful statistics in observed dynamic behavior. Thus, a total of 750 ns of cMD will be conducted for the H2A.B NCP. To conduct these simulations, we request 225,000 SUs on Comet, in accordance with our benchmarking data. Additionally, we are requesting 0.42 TB of storage space on Data Oasis for this portion of the project.

G.3.1.2 Umbrella Sampling of DNA Unwinding

To determine the simulation protocol for our umbrella sampling calculations, we reference a recent coarse-grained simulation of DNA unwinding from the standard NCP [42]. The reported free energy profile suggests that the unwrapping of DNA occurs in an asymmetric pathway in which only one end of DNA unwinds, in agreement with other studies [43, 44]. To generate their PMF, Zhang et al. coupled harmonic restraints to the radius of gyration (Rg) of all DNA base pairs over 12 windows spanning a range of 45-72.5 A, with centers evenly spaced every 2.5 A. However, sufficient sampling for a similar protocol over such a wide range of Rg values at the atomistic level is currently unfeasible. Instead, we will utilize the asymmetric unwinding behavior to conduct two separate umbrella sampling calculations per NCP system, one for each DNA end. The reaction coordinate will be defined by the Rg of 73 base pairs extending from the dyad axis in either direction, and harmonic biases will be introduced to 18 windows spread over a range of Rg = 42.5-59.5 A. Note that we are increasing the number of windows relative to the study done by Zhang et al., as we expect that our all-atom model will have smaller fluctuations, and thus require a higher density of windows, then their course grain model. Probability densities from all windows will be converted into a 1-dimensional potential of mean force (PMF) using WHAM [45].

The accuracy of any free energy calculation relies heavily on the convergence of each simulation window. Therefore, we propose that each window will be run for ∼250 ns, in accordance to our cMD protocol. To increase computational performance, we will employ a hydrogen mass repartitioning protocol to increase our timestep two-fold (Dt = 4 fs) [46], and we will treat the solvent environment implicitly, which has been

shown to provide good interpretations of nucleosome dynamics [47]. In doing so, we will reduce each system to ∼25,000 atoms to be simulated at 300 K in the NVT ensemble with all pair-wise interactions calculated directly.

We will utilize the effcient GPU performance of the recently developed pmemd.CUDA (v16) for simulating each window. From our benchmarks on XStream, simulating the proposed windows will require 45,000 SUs per NCP type, so we request a total of 90,000 XStream SUs to conduct our umbrella sampling calculations. Furthermore, we request 0.52 TB of storage space on Data Oasis for this portion of the project.

G.3.2 Simulations of the Chromatosome

To determine the structural and energetic factors that dictate the binding states of linker histones, conventional MD simulations of chromatosomes with GH1, GH5, and the GH5 pMu linker histones will be conducted. Each simulation will be initiated in both the on- and off-dyad conformations to determine the how seemingly minor structural differences shift the equilibrium of these states. Simulations of the on-dyad state will be initiated from the 4QLC crystal structure of the H5 chromatosome system, with missing loops and tails modelled using Modeller 9.17 [48]. Currently there is no atomic-resolution structure of a chromatosome containing an off-dyad linker histone. Therefore, these simulations will be based on the cryogenic electron microscopy (cryo-EM) micrograph of a chromatin fiber which contains histone H1 in the off-dyad binding mode (EMDataBank ID emd 2600.ccp4) [49]. Using this electron density map, we will perform molecular dynamics flexible fitting (MDFF) calculations to fit linker histones to the off-dyad position [50]. These simulations will be performed with local resources.

Simulation parameters will be similar to those for our H2A.B simulations, including the use of an NPT ensemble at 300 K and 1 atm with the NAMD engine and a 2 fs timestep. Protein and nucleic acid parameters will be taken from the AMBER ff14SB force field, and TIP3P waters will be used. As in our previously performed simulations of the canonical and macroH2A-containing NCP, as well as the proposed H2A.B simulations, each simulation will be performed for 250 ns three times. Analysis will focus on the complex stability, as well as the binding energetics and dynamics of the different linker histones.

Using current resources allocated by XSEDE, we have already begun - and expect to finish - three simulations of the chromatosome containing GH5 in the on-dyad binding mode. Therefore, we are requesting resources for a total of 15 separate 250 ns simulations to complete three trajectories of the remaining five systems. In accordance to our benchmarking data, we are requesting 1,312,500 Comet hours (see Table 2). Additionally, we are requesting 2.40 TB of storage to store our trajectory files.

Table 2: A summary of the total requested resources for simulations relating to the chromatosome. A benchmark speed of 350 SUs/ns was measured on Comet.

Linker Histone (Conformation)	Simulation Method	System Size (atoms)	Simulation Length (ns)	Number of Simulations	Comet SUs
GH1 (On-Dyad)	cMD	270,000	250	3	262,500
GH1 (Off-Dyad)	cMD	270,000	250	3	262,500
GH5 (Off-Dyad)	cMD	270,000	250	3	262,500
GH5_pMu (On-Dyad)	cMD	270,000	250	3	262,500
GH5_pMu (Off-Dyad)	cMD	270,000	250	3	262,500
				Total:	**1,312,500**

G.3.3 Simulations of the Canonical Octanucleosome

To begin our study of inter-NCP allosteric networks, we are performing cMD simulations of a compact octanucleosome comprised of canonical histones. The original structure for these simulations is based on the 11 A resolution cryo-EM micrograph of a 30 nm chromatin fiber reconstituted on 12 x 177 base pairs (EMDataBank ID emd 2600.ccp4) [49]. Individual NCPs are based on the 1.9 A crystal structure by Davey et al. [40] (PDB ID: 1KX5). Each tetra-NCP is based on the crystal structure by Schalch et al.[23] (PDB ID: 1ZBB). We use the AMBER14SB force field to describe the proteins and the DNA. The system is solvated with TIP3P water molecules, and counterions have been added to neutralize the system and simulate a 150 mM NaCl environment [37, 51]. The resulting system has ∼2,800,000 atoms. We use Gromacs 5.0.6 to run these simulations. Pressure is controlled using the Parrinello-Rahman barostat [52, 53] while temperature is controlled using the Nose-Hoover thermostat [54, 55].

To map allosteric effects in the NCP, we will utilize multiple techniques. Residue-residue correlations will be calculated by utilizing mutual information methods [56, 57], while allosteric networks will be mapped with the Weighted Implementation of Sub-optimal Pathways approach [58]. In addition, the importance of individual residues to dynamic networks throughout the structure will be computed with edge-betweenness centrality calculations [59]. Finally, we will determine how packing into chromatin fibers alters the dynamics of a single NCP by computing the Kullback-Leibler divergence of dihedral angle populations between these simulations and our previous monoNCP simulations [60].

To estimate the required simulation times for convergence of these calculations, we refer to our previous NCP study where we performed three independent 250 ns simulations for each construct. Given the system size of the octaNCP, dynamics is expected to be slowed by packing. We estimate that simulations twice as long, or 500 ns in length, will be required to achieve an acceptable level of sampling. We have already performed 250 ns of cMD of a canonical octanucleosome. Preliminary results show that simulation times of ∼150 ns are required to reach equilibrium. In addition, it is important that we perform these simulations multiple times to delineate which dynamical properties are robust. Therefore, we intend to perform three 500 ns simulations. Based on the timing in the benchmark section, we therefore request 3,750,000 Stampede

hours (see Table 3). In addition, we are requesting 3.96 TB of storage to save our trajectory files (see scaling document).

Table 3: Simulations we are requesting XSEDE resources for in the octanucleosome project. A conversion rate of 2997 SUs/ns was measured on Stampede (see Benchmark data).

System	System Size (Atoms)	Simulation Time	Number of Simulations	Stampede SUs
OctaNCP	2,800,000	500 ns	2	3,000,000
OctaNCP	2,800,000	250 ns	1	750,000
			Total:	**3,750,000**

G.3.4 Umbrella Sampling Molecular Dynamics Simulations of Pilin Subunits

Using in house resources, we implemented three conventional and three accelerated MD simulations of the monomer and dimer forms of RrgB and SpaA. The results suggest that their structures are more dynamic in the absence of the intra-domain isopeptide bonds. In some of the calculations, a kink appears in the dimer with the angle between monomers approaching approximately 110°. Such structural changes were not seen in the simulations of dimers with an isopeptide bond, which could imply that isopeptide bonds are vital to pili structural integrity. However, it could also be that the calculations implemented for structures with an isopeptide bond are yet to find this kinked conformation. Clarifying this requires the use of free energy calculations. Therefore, we propose to implement umbrella sampling calculations to further extend the aMD and cMD simulations we have already performed and provide detailed molecular level information about the role of isopeptide bonds in pilin molecules.

The octahedral box containing a solvated RrgB dimer is made up of 987,650 particles. The SpaA system has a slightly smaller total number of particles due to its smaller size. To implement the umbrella sampling calculations, a dihedral collective variable will be used. The vertices of this dihedral will be made from the center of mass of groups of atoms in the first and last sub-units of RrgB and SpaA. These calculations will therefore make use of the non-GPU accelerated version of AMBER, since such groupings are yet to be implemented in the GPU accelerated module. The dihedral angle will be varied from 90° to 180° in 3° increments, leading to thirty windows per system. Each window will be run for 100 ns, which means a total time of 3.1 μs per system, and windows will be combined to create a PMF using WHAM [45]. A 100 ns simulation time was chosen since we observed equilibration times of between 20 to 40 ns in the aforementioned cMD calculations, which will allow 60 to 80 ns of sampling time for collecting statistics. Calculations will be carried out three times for RrgB and SpaA dimers in the presence and absence of isopeptide bonds, yielding a total of 12.0 μs simulation time for this project. Therefore, in accordance with our benchmarking data, we request 744,000 Comet SUs for our free energy calculations of pilin structures. We are also requesting 14.9 TB of storage for these calculations.

Table 4: Simulations we are requesting XSEDE resources for umbrella sampling of pilin proteins. A conversion rate of 60 SUs/ns was used for the simulations (see Benchmark data).

System	Number of Simulations	Length	Comet SUs
RrgB Dimer	100 ns × 31 windows	3.1 µs	186,000
RrgB Dimer with Iso	100 ns × 31 windows	3.1 µs	186,000
SpaA Dimer	100 ns × 31 windows	3.1 µs	186,000
SpaA Dimer with Iso	100 ns × 31 windows	3.1 µs	186,000
		Total:	**744,000**

H XSEDE Review & Reconciliation Process

This section provides additional details about the XSEDE Review and Reconciliation process, through which recommendations are made and then reduced to final allocations. In addition to conversations with individuals at XSEDE, much of the following (including excerpts) comes from the 2021 v1.1 version of *XSEDE Allocations Practices and Procedures* and the 2019 v1.1 version of the *XRAC Reviewer Manual*.[35]

H.1 Review Criteria:

The excerpt below from the *XRAC Reviewer Manual* summarizes the criteria reviewers should consider when making recommendations from proposals:

1. *Appropriateness of Methodology: Does the request describe appropriate tools, methods, and approaches for addressing the research objectives? These methodologies may be community codes or models, data analysis methods, or algorithmic formulations expressed in user-developed scripts or tools.*
2. *Appropriateness of Research Plan: Does the request describe necessary and sufficient computational experiments to answer the research questions posed? In some cases, the research plan may be more reasonably expressed as estimates of resource use, supported by past or early experience. Serious concerns about the research plan will be documented in reviews and may lead to reduced allocation awards.*
3. *Efficient Use of Resources: Has the request identified appropriate resources for undertaking the research plan with the proposed methodology? And will those resources be used as efficiently as is reasonably possible and in accordance with the recommended use guidelines of those resources? Is the proposed work better suited or equally well suited to other resources available for allocation?*

XRAC reviewers may also consider these "additional considerations:"

Intellectual Merit: Reviewers also consider whether the work has any supporting grants for which the science has been merit-reviewed. If so, is the allocation request consistent with the objectives of the supporting grants as described in the submission and is the scale of resource use commensurate with the level of support? If no such grants are identified, the reviewers must assess the intellectual merit of the work. This assessment of the intellectual merit will factor into their overall recommendation.

[35] The reviewer manual can be found at https://www.ideals.illinois.edu/items/111788/bitstreams/366207/data.pdf. The practices & procedures document has been removed from the internet but can be shared by the authors upon request.

In our context, the intellectual merit consideration serves to put all requests on a "level playing field." The vast majority of grants are supported by government agencies and do not warrant review of their intellectual merit.

> *Prior results: Research requests typically require some form of prior work or progress to be considered for an allocation. For New requests, demonstrated benchmarking or test runs are typically required on the requested resource in almost all cases; in some instances (such as a request for a new resource for which Startups are not yet available), demonstrated work on an architecturally similar resource may suffice. For Renewal requests, reviewers consider the provided Progress Report as well as utilization levels of prior allocations*

The prior review criteria for renewal requests ensures researchers are using their existing allocations (i.e., are not wasteful). In conversations with XSEDE directors, this criteria serves primarily to check that prior allocations are indeed "doing something." In particular, reviewers may look at if the project has produced *any* publications. If this leads grants with a prior publication to receive a higher recommendation than their *true need*, one might wonder whether this may introduce bias to our quantity (number of papers) analysis. This is not a concern because larger recommendations are reduced more (see Appendix H.2) so this feature of the renewal process would in fact bias our estimates in Table 2 downward, leading us to underestimate the true effect. This can be seen in a simple numerical example. Suppose we observe a project that had already published a paper using a previous XSEDE allocation and the PI requests a renewal. This project is recommended 100 SU and is awarded 50 SU ($\%Allocated = 50\%$). If, in reality, this project had not already published a paper, the project may have been recommended 90 SU. Because larger recommendations are reduced more (see reconciliation formula below), this project will be awarded an allocation greater than 45 SU (such that $\%Allocated > 50\%$). Hence, the project with prior publications (and higher expected productivity) has a lower ratio – biasing the results downward.

H.1.1 XRAC Recommendations and Resource Availability

The excerpt below illustrates that reviewers are instructed to make recommendations independent of available XSEDE resources.

> *In all parts of the process, reviewers are asked to consider requests only on their merits. The reviewers are not informed about the total requests on the different systems until the caucus session and only after their individual reviews have been completed. This approach is designed to ensure that reviews are based purely on the merit of the proposal. Providing the allocation request total and available units for each resource at the caucus allows them time to prepare mobility information for requests that need to be or could possibly be moved to alternate systems in an oversubscription situation. The total of the Recommended Allocations is reported to NSF,*

so that NSF will be aware of the panel's statement of the actual need for resources.

H.2 Reconciliation formula

After recommendations have been made by the XRAC for each proposal, the Allocations team meets to "reconcile" recommendations with the resources available. In the first ("balancing") phase, the team reallocates resources from over-recommended systems to computing systems which are less over-recommended but are architecturally similar, until the excess demand is balanced equally across platforms. There are generally two categories of architecturally similar resources that XSEDE considered: CPU-oriented systems and GPU-oriented systems.

After the balancing, if there is no longer excess demand for a group of architecturally similar resources, grants are allocated their full recommendation. If there is still excess demand after this balancing occurs, recommended allocations are reduced (the "reconciliation" phase). The amount of the reduction for each project is determined by the following formula:

$$\text{Award} = [(1 - G) \times R_n + (1 - G) \times F \times R_o] \times Size(R)$$

R is the Recommended amount, R_n the share of the Recommended amount supported by NSF-funding (between 0-1) and R_o the share of the Recommended amount not supported by NSF-funding $(1 - R_n)$. F is the funding priority factor which notes how much priority is given to higher NSF shares, and $Size$ is a scaling function that linearly reduces smaller awards less than larger ones.[36] In the formula, G is the global scaling factor, between 0-1, that is solved for in Microsoft Excel to maximize the total of all allocations (Awards) subject to the supply available. The term in brackets is a share less than or equal to one (this can be thought of as a "tax rate"). $Size(R)$ indicates a number of SUs (the recommendation scaled for it's size) that is multiplied by this share to get the final award value.

The process is described in the practices and procedures document as:

> *Following the discussion of all requests, if the aggregate Recommended Allocation is less than the total Available Amount for a system, the Recommended Allocations generally become the Awarded Allocations. When the total Recommended Allocation exceeds the total Available Allocation for a given resource, XSEDE and SP staff use a procedure approved by the NSF to adjust the Recommended Allocations to arrive at the Awarded Allocations. This Reconciliation adjustment is performed by the XSEDE Allocations staff and SP representatives after the XRAC meeting*

[36] Sufficiently small recommendations, typically less than one million SU, are not reduced. F takes on a value of either 1 or 0.5 depending on the level of oversubscription. Note that this differs from the XSEDE documentation which lists possible values of F as 0.5 or 2 – this is a typo according to our XSEDE contacts.

ends and formulaically adjusts all Recommended Allocations to remove oversubscription. Within two weeks of the meeting, the XSEDE allocations staff complete the Reconciliation process, summarize the panel discussions, craft administrative comments or notes to PIs as needed, and send notifications of final awards to all PIs, co-PIs and Allocation Managers of requesting projects, as well as to all affected SPs.

I Concepts

Definition

Dimensions AI categorizes and organizes vast amounts of academic research data using machine learning techniques. One of its key features that we utilize in this paper is its ability to extract and identify "concepts" from scientific texts.

Concept Identification: Dimensions AI scans through publications' abstracts and identifies "concepts" associated with each paper. These concepts are meant to represent key themes or topics within a paper. They are derived from the text using machine learning models that recognize important keywords, phrases, and themes commonly associated with specific scientific domains.

Relevance Scoring: Each concept identified by Dimensions is assigned a relevance score, which indicates how central the concept is. While the exact scoring scheme is proprietary, it is a function of the field of study of the paper.

Differences with other methods: The algorithm of Dimensions AI differs from other algorithms such as LDA where the number of topics has to be imputed. This allows us to have variation in breadth. In that sense, Dimensions AI concepts are closer in idea to Hierarchical Dirichlet Process (HDP) that have been used previously in the literature, where the number of topics assigned to a document is not pre-determined.[37]

We present two examples to illustrate the use of concepts.

Paper 1: Defect-engineered graphene chemical sensors with ultrahigh sensitivity

- Abstract: *We report defect-engineered graphene chemical sensors with ultrahigh sensitivity (e.g., 33% improvement in NO2 sensing and 614% improvement in NH3 sensing). A conventional reactive ion etching system was used to introduce the defects in a controlled manner. The sensitivity of graphene-based chemical sensors increased with increasing defect density until the vacancy-dominant region was reached. In addition, the mechanism of gas sensing was systematically investigated via experiments and density functional theory calculations, which indicated that the vacancy defect is a major contributing factor to the enhanced sensitivity. This study revealed that defect engineering in graphene has significant potential for fabricating ultra-sensitive graphene chemical sensors.*
- https://doi.org/10.1039/C5CP04422G
- Concepts: chemical sensors, conventional reactive ion, density functional theory calculations, ultrahigh sensitivity

[37] See for instance Furman & Teodoridis (2020).

Paper 2: Hybrid massively parallel fast sweeping method for static Hamilton–Jacobi equations

- Abstract: *The fast sweeping method is a popular algorithm for solving a variety of static Hamilton–Jacobi equations. Fast sweeping algorithms for parallel computing have been developed, but are severely limited. In this work, we present a multilevel, hybrid parallel algorithm that combines the desirable traits of two distinct parallel methods. The fine and coarse grained components of the algorithm take advantage of heterogeneous computer architecture common in high performance computing facilities. We present the algorithm and demonstrate its effectiveness on a set of example problems including optimal control, dynamic games, and seismic wave propagation. We give results for convergence, parallel scaling, and show state-of-the-art speedup values for the fast sweeping method.*
- https://doi.org/10.1016/j.jcp.2016.06.023
- Concepts: Hamilton-Jacobi equation, algorithm, computer architecture, dynamic game, equations, example problems, fast sweeping method, heterogeneous computer architectures, high performance, hybrid parallel algorithm, optimal control, parallel algorithm, parallel computing, parallel method, parallel scaling, popular algorithms, seismic wave propagation, speedup values, static Hamilton–Jacobi equations, sweeping algorithm, sweeping method, wave propagation