Shortcuts to Innovation: The Use of Analogies in Knowledge Production

Soomi Kim[†]

July 2, 2023

Abstract

Old ideas serve as critical inputs into new ideas, but how do knowledge workers innovate when there are only few existing ideas to build on? In this paper, I explore how analogical reasoning—and technologies that automate it—can serve as "shortcuts" that allow innovators to import knowledge from an adjacent domain, bypassing the need to build knowledge from the ground up. Yet, because analogies require the availability of other domains as templates, they may also constrain the direction of innovation towards areas with available templates. Using the setting of structural biology, I document a tradeoff: while the arrival of an analogy-based technology increased the rate of innovation, it led to workers herding around solving less impactful problems.

[†] Columbia Business School. Email: sk5261@columbia.edu.

1. Introduction

Knowledge production is cumulative (Romer 1990; Scotchmer 1991). Innovators use existing ideas to produce new ideas—from mechanical engineers relying on the foundation of Newtonian physics to applied economists turning to canonical econometric models. But how do knowledge workers innovate in new domains where there are only few existing ideas to build on?

In this paper, I explore how technologies can be used as "shortcuts" that speed up the process of acquiring foundational knowledge—and even circumvent it altogether. Much of the prior literature on research technologies has focused on vertical shortcuts, those that help innovators more quickly understand and apply existing knowledge to climb to the frontier. What is less obvious is how to innovate in domains where there is no foundation yet. I argue that horizontal shortcuts—specifically, analogies—allow innovators to import knowledge from another domain, bypassing the need to build knowledge from the ground up.

Consider the early history of aviation, when the physics of aerodynamics were not yet discovered. Inventors looked to birds as analogies to design devices that imitate the motion of flapping wings. But all of these attempts failed—until George Cayley, a 19th-centry British inventor, made a breakthrough. In place of flapping wings, Cayley envisioned the first prototype of modern-day airplanes that was later built by the Wright Brothers: a device with fixed wings, which glides to sustain lift.

This history of aviation illustrates both the power and pitfalls of analogies. Not only do analogies identify unexpected connections across knowledge domains, they can also circumvent the need to build foundational knowledge by borrowing insights from another domain. Yet, because analogies require the availability of templates, they may narrow the line of inquiry. Birds provided guidance behind the mechanics of flight, but since they were the only known templates for flight (and the underlying physics were not yet known), early inventors focused on flapping wings and did not consider fixed-wing machines (Spenser 2008; Pollack 2014).

Analogical reasoning is ubiquitous in both research and managerial practice. A wide range of scientific and engineering breakthroughs have been sparked by analogies (Gentner, Holyoak, and Kokinov 2001), from Velcros inspired by plant burrs to Ernest Rutherford's model of the atom as a miniature solar system. In strategy, managers often face strategic problems that are well-suited for analogical reasoning (Gavetti, Levinthal, and Rivkin 2005; Bingham and Kahl 2013). Entrepreneurs commonly conceive of new ventures by adopting insights from one industry

to another, such as the over hundred startups that claim to be the next "Uber for X" (Madrigal 2019), from Instacart (Uber for grocery deliveries) to Wag (Uber for dog walkers).

Although analogical reasoning has played a central role in innovation, automation has made it increasingly easier to deploy. In particular, algorithms based on supervised machine learning can be seen as the automation of analogical reasoning. Such algorithms mine patterns from known training templates and apply those patterns to new areas, mirroring human ability for finding patterns that explain the unfamiliar in terms of the familiar. For example, drugmakers refer to well-known compounds to identify promising candidates among unexplored compounds, while managers rely on their experiences with past employees to screen applicants—and drug discovery and hiring algorithms now routinely conduct these types of analogical reasoning on their behalf. By taking analogical reasoning out of an individual mind and outsourcing it to machines, some believe that analogies can be harnessed at scale with data-reliant technologies (Kittur et al. 2019).

However, while analogies can help innovators make progress in uncharted terrains, their need for templates can also restrict the direction of innovative activities. The automation of analogical reasoning makes this tradeoff especially salient: analogy-based technologies like machine learning can only be employed in areas with training data. The availability and location of training data can thus shape the direction of innovation towards some areas, while neglecting others.

In addition to providing a framework for how analogies can serve as shortcuts in innovation, the goal of this paper is to empirically study the tradeoff of relying on analogies. Although analogical reasoning has been a topic of great interest in cognitive psychology (Gentner, Holyoak, and Kokinov 2001; Hofstadter and Sander 2013), along with fewer but important studies by management scholars (e.g., Gavetti, Levinthal, and Rivkin 2005), much of the prior work on analogies has been based on laboratory experiments or case studies.

The scarcity of empirical studies based on real-world, large-scale data is perhaps unsurprising. Analogical reasoning, while pervasive, is a mental shortcut that often goes unnoticed (Dunbar 1999). Above all, analogies—and analogy-based technologies—are challenging to study empirically. In order to investigate whether the use of an analogy-based technology shifts the direction of innovation, it is important to be able to observe what ideas could be pursued in the absence of the technology. Finding such a setting is difficult because the counterfactual of ideas that could have been pursued—but were not—is often unobservable. There also needs to be a credible way to measure distance between ideas, as analogies work by identifying similarities between disparate domains. Finally, the analogy-based technology must differentially treat only

some areas of the setting, such that outcomes in the areas where the technology was introduced can be compared to the areas without the technology.

I focus on the setting of structural biology, a field with empirical features ideally suited for this paper. Structural biology studies the 3D structures of proteins and has contributed to more than a dozen Nobel prizes, as proteins play vital roles in virtually every biological process. Elucidating a protein structure at atomic resolution—or "solving" the structure—can reveal the protein's function, which helps with applications such as designing vaccines that train human antibodies to recognize the spike proteins of SARS-CoV-2. Importantly, structural biology has several empirical features that allow me to identify how the introduction of an analogy-based technology may have shaped the subsequent rate and direction of innovation.

First, unlike many settings where only realized ideas are observable, structural biology provides a window into the entire idea landscape. Using a database of all known proteins, I observe which proteins structural biologists explored versus could have explored but neglected. For instance, of the approximately 20,000 human proteins, just one-third of them have had their structures experimentally determined as of 2020. In addition, while most settings do not have an easy way to quantify the similarity between each potential idea, the distance between ideas can be measured in structural biology (Hill and Stein 2020; 2021): proteins are composed of sequence of amino acids, so they can be grouped based on their sequence similarity. In other words, it is possible to map out the idea landscape of all known proteins and see which areas of the landscape have been explored (which I term "bright" clusters of proteins) and which areas remain unexplored and thus do not have built-up knowledge (which I term "dark" clusters).

Second, structural biology is a prime setting for studying analogy-based technologies. Solving a protein structure involves deep knowledge of biology, physics, and statistics, but many of the steps have now become automated. The specific technology I examine is the software program Phaser, released in 2003, which automates a method called molecular replacement (MR). Instead of solving a structure from scratch, MR borrows structure information from previously solved proteins that are similar to the unknown structure that the scientist is trying to elucidate. MR can therefore be viewed as an analogy-based technology since it helps knowledge workers make progress in areas of research without existing knowledge (i.e., proteins whose structures are unknown) by importing structure templates from neighboring proteins.

Finally, this analogy-based technology differentially treated some parts of the idea landscape. Since MR needs data on previously solved structures, MR only works for bright clusters of proteins (i.e., clusters with previously solved structures), and does not work for dark clusters. This allows

me to employ a difference-in-differences design where bright and dark clusters serve as my treatment and control groups. By matching data from Swiss-Prot (a database of all known proteins) to the Protein Data Bank (a database of all protein structures), I examine the quantity and quality of structures solved in bright clusters after the arrival of MR, relative to dark clusters.

My first set of results focuses on the rate of innovation. Since MR reduced the cost of solving unknown structures in bright clusters, one would expect more structures to be solved in those clusters. Indeed, I find that bright clusters experienced a relative increase in the total number of solved structures after MR was introduced. This effect was sustained throughout the entire sample period: bright clusters got brighter and brighter.

I then turn to how MR impacted quality and distinguish between two dimensions of quality: execution and importance. In any type of innovative activity, the innovation should be well-executed, but it should also solve an important problem. In the case of structural biology, execution refers to how meticulously a structure was solved (e.g., the resolution or the level of detail found in the structure), while importance refers to whether the structure led to a novel understanding of a biological process. The goal of structural biology is not to solve structures for the sake of solving them; the goal is to learn the functional roles the proteins might play by elucidating their structures. I find that while bright clusters received more well-executed structures, these structures were less scientifically important. They provided fewer functional annotations about the proteins and had lower publication and patent footprint.

A potential identification concern is that bright clusters may have been evolving on different trends than dark clusters before the introduction of MR. I conduct several robustness analyses to address this concern. First, I show that there are no pre-trends in the corresponding event studies. Second, I control for predicted brightness; the idea is to compare clusters that share ex-ante similar traits, but some clusters just happened to be actually bright while other clusters happened to be dark. MR only works when the cluster is actually bright regardless of whether the cluster is predicted to be bright or dark, and I verify that only actual brightness matters when estimating the impact of MR.

Taken together, my results suggest a tradeoff: while the arrival of MR increased the rate of innovation, it also led to knowledge workers solving less impactful problems. These results from structural biology illustrate an inherent limitation of analogies: analogies may serve as shortcuts for making progress in new domains (i.e., proteins whose structures are unknown), but they are also constrained by the need for templates and thus may be employed in domains with neighbors—

that is, potentially crowded areas (i.e., bright, already well-explored clusters) that may not be the most impactful.

This paper contributes to several literatures. I first build on a body of evidence that examines how technologies can both advance and constrain knowledge production. Much of this prior work studies technologies that can be classified as those that help innovators digest and apply *existing* knowledge (Teodoridis 2018; Furman and Teodoridis 2020; Anthony 2021; Miric, Ozalp, and Yilmaz 2021; Mannucci 2017). In contrast, by introducing the idea of analogies, this paper aims to shed light on technological shortcuts that help knowledge workers innovate in domains where little is known. In other words, I distinguish between technologies that reduce the "burden of knowledge" (Jones 2009)—the problem of innovators facing an increasing educational burden as knowledge accumulates—and technologies that alleviate a different problem of innovators lacking existing ideas that can serve as inputs in the production of new ideas.

By focusing on analogy-based technologies, I also contribute to the emerging literature on AI and data-driven exploration. While the literature on AI has extensively documented how algorithmic bias can arise from poor-quality training data (Cowgill and Tucker 2020; Cowgill et al. 2020; Choudhury, Starr, and Agarwal 2020), this paper joins a smaller literature that focuses on how the very availability of training data can dictate where innovations take place (Cockburn, Henderson, and Stern 2018; Hoelzemann et al. 2022).

Lastly, I leverage the setting of structural biology, which was first brought to the attention of social scientists by Hill and Stein (2020; 2021). The authors assess the costs of the priority reward system in science, which tends to only recognize the first discoverer.[1] While I do not study priority races and instead examine the impact of an analogy-based technology by exploiting a novel identification strategy, I follow Hill and Stein (2020; 2021) in highlighting the strengths of the setting. As a field with both rich scientific achievements and empirical features, structural biology is an attractive setting for investigating broader questions of how to manage innovation.

The rest of the paper proceeds as follows. Section 2 provides a taxonomy of shortcuts and an overview of the key features and tradeoffs of analogies. Section 3 introduces the institutional context and empirical features of structural biology. Section 4 lists the main data sources. Section 5 describes the difference-and-differences design that underpins this study's empirical strategy.

[1] Specifically, Hill and Stein (2020; 2021) document the effects of being "scooped" on subsequent career outcomes, as well as how competition leads to rushing and lower-quality science. Additionally, a recent paper by Zhuo (2022) estimates a model of lab decision-making on resource allocation in structural biology.

Section 6 presents my main results, along with robustness analyses. Section 7 discusses the implications of my results and conclusions.

2. Shortcuts to Innovation

The cumulative nature of knowledge production typically characterizes the innovation process as a sequence of old ideas generating new ideas (Romer 1990; Scotchmer 1991). This section discusses how there can be shortcuts (particularly shortcuts enabled by technologies) that can speed up—or even bypass—this sequence of knowledge accretion.

2.1 Prior Literature on Research Technologies: Vertical Shortcuts

As knowledge accumulates, every new generation of innovators faces a greater educational burden. This "burden of knowledge" has several implications, including increased training length and specialization as innovators struggle to learn the growing body of knowledge (Jones 2009). The prior literature on research technologies has primarily focused on technologies that I consider as "vertical" shortcuts, which mitigate this burden. These are technologies that allow innovators to more quickly acquire existing foundational knowledge, such that they can use the knowledge as stepping stones to climb to the frontier of knowledge and produce new ideas.

As shown in Figure 1, vertical shortcuts can be thought of as aiding in either understanding or applying foundational knowledge. *Summaries*—from textbooks[2] to Wikipedia—help with understanding existing knowledge by providing a short synopsis of a given domain, saving knowledge workers from having to read every research article or replicate every experiment. *Calculators* help with applying foundational knowledge by executing a pre-programmed menu of instructions based on such knowledge. Consider programs like Stata, which is embedded with canonical econometric models. Stata allows even a college first-year with little training in econometrics to run regressions by simply entering "reg y x."

A large body of prior work on research technologies can be conceptualized as vertical shortcuts, particularly calculators of varying sophistication. Calculators are closely related to the idea of modularity, where "information hiding" (Parnas 1972) within modules glued together by standardized interfaces can facilitate a division of innovative labor (Baldwin and Clark 1997; Sanchez and Mahoney 1996; Simcoe 2015). Examples studied in prior work range from financial

[2] A recent work by Greenblatt (2021) on medical guidelines illustrates how summaries can spur innovation.

spreadsheet technology (Anthony 2021) to animation toolkit (Mannucci 2017) to videogame "middleware" (Miric, Ozalp, and Yilmaz 2021). Although this prior literature does not, for the most part, explicitly discuss the burden of knowledge, notable exceptions are Teodoridis (2018), Furman and Teodoridis (2020), and Nagle and Teodoridis (2020). The authors examine the arrival of an automating motion-sensing technology and suggest the role technologies can have in reducing the burden of knowledge.

2.2 Analogies: Horizontal Shortcuts

While vertical shortcuts can assist knowledge workers in innovating in domains with deep foundation, what about in new domains without such foundation? This paper complements the burden of knowledge literature by focusing on a different problem: how to innovate when there are few existing ideas that can serve as inputs into new ideas.

In new domains, knowledge workers face the challenge of having to build foundational knowledge from scratch—and analogical reasoning can be used to circumvent this challenge. In this section, I describe the key features and tradeoffs of analogies.

2.2.1 Key Features of Analogies

Analogical reasoning has been extensively studied by cognitive psychologists as a crucial component of human cognition. Analogy has been broadly described as "the ability to identify similarities in *relations* that hold within domains" (Gentner 1982; Gentner, Holyoak, and Kokinov 2001; Holyoak and Thagard 1996), even if the individual objects are distinct (e.g., how sound propagates through the air is analogous to how water waves travel in a pond, even though sound and water are not alike).

While the concept of analogy has been employed in diverse disciplines, ranging from linguistics to philosophy,[3] I focus on a simple definition of analogy adapted from cognitive science: the importing of patterns from one knowledge domain to another. This definition leads to three key features of analogies, with respect to their role in knowledge production.

(i) Analogies can serve as shortcuts. Analogical reasoning can be viewed as a shortcut because it can serve as an alternative to other ways in which innovators build knowledge in new domains, such as trial-and-error (Thomke 1998) or by generating a theory (Fleming and Sorenson

[3] Even before the rise of modern cognitive science, the word analogy had been widely used, prompting the 19th-century political economist John Stuart Mill to once remark, "There is no word . . . which is used more loosely, or in a greater variety of senses, than Analogy" (Mill 1843/1974).

2004). As an example, suppose a drugmaker is trying to create a drug for a new disease. One approach would be to screen through millions of compounds to detect pharmacological activity through trial-and-error. Another approach would be to start by building a theory of how the disease operates at the molecular level and then design drugs that target the molecular action (Henderson 1994). However, trial-and-error can require extensive resources and does not guarantee a solution, while uncovering underlying causal principles is challenging and not always possible.

Instead of building knowledge from the ground up through trial-and-error or theory development, I focus on how innovators can take a "horizontal" shortcut by importing intuition and insights from a neighboring field. With the case of drug development, in lieu of brute-force screening or rational drug design, drugmakers can rely on pattern recognition by identifying drug candidates for a new disease based on already approved drugs for similar diseases.

This strength of analogical reasoning in helping innovators quickly make progress in new domains has been highlighted in prior work. Psychology studies have shown that scientists frequently substitute slow, iterative experimentation with analogical reasoning to speed up problem solving (Dunbar 2000). In the context of business strategy, using a simulation and a rich set of case studies, Gavetti, Levinthal, and Rivkin (2005) demonstrate that managers often face strategic problems that are best suited for analogical reasoning: problems that are neither too modular (where rational, deductive reasoning can instead be employed) nor too complex (where only trial-and-error can work).[4]

(ii) Analogies are not simple recombinations. Analogical reasoning's reliance on pattern recognition distinguishes analogies from the traditional concept of recombinant innovation (Schumpeter 1934; Weitzman 1998; Uzzi et al. 2013). The key insight in the recombinant literature is that new ideas can be generated from existing, well-understood ideas if they are combined in a novel way.[5] In contrast, rather than mixing well-understood ideas, analogical reasoning involves the borrowing of less-understood patterns from another domain. In the bird-airplane analogy, for example, inventors of flight did not understand the physics of aerodynamics and therefore did not know exactly how birds can fly. But these early inventors speculated that the motion of wings is

[4] In economics, Gilboa, Samuelson, and Schmeidler (2015) develop a formal model of reasoning, in which economic agents use analogical reasoning when the underlying data generating process is unknown and use rule-based reasoning when the data structure is known.

[5] For instance, Brynjolfsson and McAfee (2014) cite Waze, the navigation app that uses real-time, crowdsourced traffic data, as a classic example of recombination. The individual components of Waze (location sensors, social networks, and smartphones) were all widely known and used, but no one had thought to combine them together to optimize driving routes before Waze.

important and applied this pattern to human-powered flight. Analogical reasoning is the importing of relational patterns—correlations—not causal logic.

(iii) As shown in Figure 1, analogical reasoning has evolved from solely being conceptual in the mind of an individual to being automated and outsourced to machines. Conceptual analogies have been fundamental throughout the history of innovation by helping individuals understand a new domain through identifying patterns across domains.[6] The modern field of biomimetics is a prime instance of conceptual analogies as shortcuts; engineers often take inspiration from properties found in the natural world and reverse-engineer them, instead of beginning with first principles or replicating the millennia of trial-and-error experiments that nature conducted through natural selection (Pollack 2014). In managerial practice, conceptual analogical reasoning has also served as critical, if underappreciated, sources of strategic visions and entrepreneurial ventures (Hill and Levenhagen 1995; Gavetti and Rivkin 2005; Gavetti, Levinthal, and Rivkin 2005; Kaplan and Orlikowski 2013; Martins, Rindova, and Greenbaum 2015; Glaser, Fiss, and Kennedy 2016).[7]

With the rise of data-reliant, pattern-recognition algorithms, conceptual analogical reasoning has become increasingly automated (Kittur et al. 2019). This has several consequences. First, this automation has allowed innovators to apply the borrowed patterns from another domain, without necessarily understanding them. For example, when relying on machine learning software libraries like TensorFlow, knowledge workers often do not know why the algorithm made the prediction it did. Despite not necessarily understanding the correlations that connect the training and test domains, knowledge workers can simply apply those patterns with TensorFlow. Second, unlike conceptual analogical reasoning which is unconstrained in the types of templates it requires,

[6] Hesse (1966) and Holyoak and Thagard (1996) provide numerous examples. One of the first accounts of (conceptual) analogical reasoning was by the ancient Roman architect Vitruvius who proposed the wave-sound model. Since then, analogies have served as the genesis behind many discoveries, from Charles Darwin's theory of natural selection, which was based on an analogy to artificial selection by farm breeders, to the Nobel laureate Salvador Luria's analogy between slot machines and bacterial mutations. Analogies can even be found in mathematics, a field built on causal logic and thus may seem less amenable to analogical reasoning. In fact, many difficult problems in algebra have been solved by turning them into geometric problems—that is, by finding analogies between algebra and geometry (Hacking 2014; du Sautoy 2021).

[7] In addition to these studies on how analogical reasoning can serve as a source of innovation (i.e., create new strategies and ventures), another strand of management literature underscores a different aspect of analogies: a dissemination tool once an innovation has been produced. When introducing novel products or services, analogies can be used to increase legitimacy (Hargadon and Douglas 2001; Bingham and Kahl 2012; Etzion and Ferraro 2010; Cornelissen and Clarke 2010). Apple, for instance, popularized the term "desktop" when launching personal computers. By analogizing between the physical and the digital desks, Apple intuitively drew in customers who were not used to working in the virtual world (Bingham and Kahl 2012).

automated analogical reasoning needs specific templates: digitized training data. I discuss the implications below in Section 2.2.2.

2.2.2 Tradeoffs in Relying on Analogies

While analogies—and by extension, analogy-based technologies like machine learning—can help knowledge workers innovate in domains where there are no existing ideas yet to build on, analogies can also have several costs. First, because analogies require the availability of templates, they may constrain the direction of innovation towards areas of research with templates, even if those areas are less fruitful. Second, these templates may highlight just superficial similarities between the target and adjacent domains, leading to misleading conclusions (Gentner 1982). Third, because analogies do not build foundational knowledge from scratch, this leads to a weak foundation: knowledge workers may not fully understand the underlying mechanisms of how the target domain works; they can only translate the target in terms of the template.[8]

This paper focuses on the first cost—that analogies may constrain innovation towards areas with templates. This cost of analogies demonstrates an inherent limitation of analogies in balancing exploration and exploitation when knowledge workers search through the idea landscape (Kauffman 1993; Levinthal 1997; Nelson and Winter 1982; Fleming and Sorenson 2004; Kaplan and Vakili 2015). For example, Kneeland, Schilling, and Aharonson (2020) propose that inventors often make "long jumps" across disparate domains; analogical reasoning can be one mechanism that inventors rely on to make such jumps to uncharted domains. At the same time, its very need for templates from which to import patterns may result in anchoring that leads analogical reasoning astray and bound to local search (Holyoak and Thagard 1996; Gavetti and Rivkin 2005).

The automation of analogical reasoning may exacerbate this fixation cost. Although hailed as a tool of exploration (Agrawal, Gans, and Goldfarb 2018; Agrawal, McHale, and Oettl 2018), supervised machine learning has one critical weakness: it cannot work without digitized training data. While the prior literature on AI has focused on algorithmic bias that arises from poor-quality training data (Cowgill and Tucker 2020; Cowgill et al. 2020; Choudhury, Starr, and Agarwal 2020)—for example, how unrepresentative training data can lead to superficial analogies that

[8] The importing of correlations via analogies brings to the forefront an aspect of knowledge production that is sometimes overlooked: knowledge domains can vary in how "strong" or "shaky" their foundation can be. Prior studies have noted how retractions (Azoulay et al. 2015), institutions (Furman and Stern 2011), or certification (Greenblatt 2021) can impact the strength of foundational knowledge in a given domain. The use of analogical reasoning may also lead to a weaker foundation due to the accumulation of correlations (in lieu of causal theories).

identify misleading patterns—there has been less attention on how the very availability of training data can restrict the direction of innovation (Cockburn, Henderson, and Stern 2018; Hoelzemann et al. 2022). Given the growing importance of analogy-based technologies, understanding this cost is important as the locus of digitized data can have long-term implications for what gets innovated.

3. Empirical Setting

An ideal empirical setting needs three ingredients: (i) an observable idea landscape, where I can track which ideas get explored versus unexplored, as well as a measure of distance between ideas, (ii) the arrival of an analogy-based technology, and (iii) specifically, the differential arrival of the technology, such that it only arrives in some parts (treated) but not other parts of the setting (control). In this section, I first introduce the setting of structural biology and its scientific importance. I then describe the empirical features of structural biology that make it an attractive setting to study analogies.

3.1 Structural Biology: The Study of Proteins

Structural biology is a field that studies the 3D structures of proteins and aims to uncover the functional roles of proteins by elucidating their structures. As Francis Crick (who discovered the helical structure of DNA) remarked, "If you want to understand function, study structure" (Crick 1990). Since proteins are responsible for carrying out most functions in cells, insights from structural biology have helped with a broad range of applications, from identifying targets for new drugs to understanding disease progression. As one evidence of its wide-reaching impact, structural biology has been recognized with more than a dozen Nobel prizes.

Structural biology has also played an important role in the fight against the coronavirus pandemic. As shown in Figure 2A, researchers solved the structure of the spike proteins that stud the surface of SARS-CoV-2—that is, determined the 3D coordinates of individual atoms in the protein. Through this direct visualization, researchers learned how these proteins latch onto receptors on human cells like "a key to a lock" (Patel, Lucet, and Roy 2020), enabling the development of vaccines that are designed to block these proteins.

3.2 Structural Biology as an Empirical Setting

3.2.1 Observable Idea Landscape

In order to investigate whether an analogy-based technology changes the direction of innovation, it is important to be able to observe the entire idea landscape. In most settings, however, researchers can only observe ideas that were realized, while alternative ideas that could have been pursued (but were not) remain invisible. For example, in scientific research, not all papers that can be written ultimately get written. Only papers that actually became published can be observed, while other papers that may have been in consideration (but were not written) cannot be observed. An attractive feature of structural biology is that it provides a unique window into the idea landscape. As detailed in Section 4, I leverage a database of all known proteins, and I can observe which proteins structural biologists chose to explore versus could have explored but neglected.

Furthermore, in most settings, it is difficult to quantify the similarity between each potential idea. For instance, in the case of scientific publishing, measuring the distance between each paper is challenging; text similarity is often used, but this is an imperfect metric for measuring intellectual distance. In contrast, structural biology provides an objective measure (Hill and Stein 2020; 2021): proteins are composed of sequences of amino acids—which are given by nature—and therefore they can be grouped based on their sequence similarity.[9] Since analogies work by identifying similarities, this measure of distance enables me to track the use of analogies.

3.2.2. Arrival of an Analogy-Based Technology

Structural biologists developed various experimental techniques to reveal the atomic structure of proteins—or "solve" the structure. Solving a protein structure involves deep knowledge of biology, physics, and statistics, and this used to be—and remains—extremely challenging. A complex structure could take months, even years, to solve. For instance, determining the structure of the ribosome (a macromolecular machine responsible for translating DNA code to produce

[9] I build on the works of Hill and Stein (2020; 2021), who study the effect of competition in structural biology and cluster structures based on their sequence similarity to identify scientists engaged in "priority races" (i.e., competing teams that worked on structures in the same cluster, unbeknownst to each other). In this paper, rather than focusing on just proteins whose structures have been characterized, I look at instead the entire universe of proteins—both structurally characterized and uncharacterized—and cluster this universe of proteins based on their sequence similarity.

proteins) took over two decades, culminating in the 2009 Nobel Prize in Chemistry (Ramakrishnan 2018).

The dominant method of solving a structure is called X-ray crystallography,[10] which proceeds in three main steps (Figure 2B). First, the protein sample must be produced in a very specific way, which is to crystallize it—packing multiple copies of the protein in a well-ordered crystal lattice. Second, once the crystal is obtained, X-ray beams are shot at the crystal, which produces diffraction patterns, as electrons in the crystal diffract the X-ray. Third, using a combination of physical laws, statistics, and intuition, structural biologists construct a density map of electrons from the diffraction patterns and build up a 3D atomic model of the protein structure. I focus on this third step of interpreting the diffraction data. Unlike the days of Max Perutz (co-winner of the 1962 Nobel Prize in Chemistry) who solved the first protein structure (hemoglobin) through painstaking hand-calculations, many of the steps of interpreting the diffraction data have become automated.

The specific technology I examine is a software program called Phaser. Phaser was released in September 2003, which automates a method called molecular replacement, or MR. Figure 3 shows the rise in the number of structures solved by MR at the Protein Data Bank, a global repository of all solved structures.[11] One of the biggest challenges in interpreting the diffraction data is called the "phase problem," a problem difficult enough that one method to solve it resulted in a Nobel prize.[12] Prior to MR, structural biologists had to resort to time-consuming experimental methods to solve the phase problem, but MR allowed structural biologists to bypass experimental phasing. Instead of solving the phase problem from scratch, MR uses previously solved structures that share close sequence similarity to the unknown structure and use them as templates to solve the phase problem of the unknown structure. One structural biologist I interviewed noted that MR could be up to 100 times faster than experimental phasing methods,[13] potentially saving months of work.

---

[10] In addition to X-ray crystallography, two other methods can be used to solve a structure: nuclear magnetic resonance spectroscopy and cryo-EM. However, crystallography is by far the most common method, as over 95% of all protein structures are solved using this method.

[11] The method of MR was first proposed in 1962, but MR was not put into wide practice until decades later due to lack of available structures, as well as lack of ready-made software programs (Doerr 2014). While Phaser was not the first software program to implement MR, it is user-friendly and more efficient (Scapin 2013), and the most widely-used program.

[12] X-ray reflections have both amplitudes and phases, but the phase cannot be measured from the diffraction patterns. Without knowing the phase, a model of the protein structure cannot be constructed.

[13] There are two experimental methods that solve the phase problem from scratch. The first method is called isomorphous replacement, which involves producing a "native" target crystal and a "derivative" crystal with a heavy metal ion introduced. By measuring the difference in diffraction patterns between the native

In other words, MR can be viewed as an analogy-based technology. The key insight behind MR is that sequence similarity has been observed to be highly correlated with structural similarity (although little is understood regarding the causal mechanism of why sequences of amino acids cause proteins to fold into their particular 3D shapes). By taking advantage of this pattern, structural biologists use MR to import phase information from neighboring proteins that share sequence similarity, rather than solving the phase problem de novo.[14]

3.2.3 Differential Arrival of an Analogy-Based Technology

Finally, MR arrived in some parts of structural biology but not others. As mentioned earlier, I observe the entire map of known proteins and the distance between each protein in terms of their sequence similarity. While some clusters of proteins received attention from structural biologists before the arrival of MR (which I term "bright" clusters of proteins), other clusters of proteins did not get any attention ("dark" clusters). Since MR needs data on previously solved structures, MR can be applied in bright clusters but is not useful for dark clusters, so bright and dark clusters serve as my treatment and control groups, respectively. This paves the way for a difference-in-differences design, as described in Section 5.

4. Data

In order to construct my map of proteins, I use two main datasets: UniProtKB/Swiss-Prot, a database of all known proteins, and the Protein Data Bank, a database of all protein structures. I then cluster proteins based on their sequence similarity to construct my final sample.

---

and the derivative crystals, structural biologists can recover the phase information. The second method is called anomalous dispersion, where structural biologists vary the X-ray wavelength to induce atoms of specific elements to produce anomalous scattering. By locating these anomalous scattering atoms, the missing phase information of the rest of the protein can be backed out. While isomorphous replacement and anomalous dispersion do not rely on the availability of prior solved structures, they can require arduous experimental efforts.

[14] In November 2020, a technology that supersedes MR was introduced: the AI program AlphaFold, created by Google's DeepMind team. AlphaFold can predict the structure of a protein based on purely its sequence of amino acids. While MR helps with specifically the phase problem of experimental structure solving, AlphaFold bypasses the need to conduct experiments at all. While AlphaFold's success falls outside of the time period studied in this paper, I discuss potential implications in Section 7.

4.1 UniProt Knowledgebase/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProtKB) is a comprehensive database of proteins. A protein is composed of sequence of organic compounds called amino acids. Information for making a protein is stored in a gene's DNA, and by translating the DNA sequence of a gene, scientists can determine the protein's existence and the sequence of amino acids that will appear in the protein. Protein sequences in UniProtKB are thus sourced by translating genes from major genome sequence databases.

To define the complete set of proteins at risk of being structurally characterized, I focus on the Swiss-Prot section of UniProtKB.[15] Created in 1986, the Swiss-Prot database is extensively reviewed, maintained, and annotated by experts based on experimental results and literature review. I also use data constructed by Perdigão et al. (2015), which provides additional characteristics on each protein in Swiss-Prot. As of October 2020, Swiss-Prot contains 563,552 protein entries.

4.2 Protein Data Bank

Established in 1971, the Protein Data Bank (PDB) is a repository of protein structures and contains over 170,000 structures as of October 2020. Since 1989, most journals have required authors to deposit their structures at the PDB as a requirement for publication, and therefore the PDB contains the universe of all publicly available structures. The PDB provides detailed descriptions about each structure, as well as crosswalks to Swiss-Prot.

4.3 Sample Construction: Clustering Proteins

After identifying which proteins in Swiss-Prot were found to be structurally characterized in the PDB, the final step is to measure the distance between each protein and cluster proteins that share sequence similarity.

I rely on MMseqs2,[16] an algorithm used by both Swiss-Prot and the PDB to cluster similar proteins (Steinegger and Söding 2018; Hauser, Steinegger, and Söding 2016). Given that molecular replacement will likely be successful if the template and the target proteins share at least 30%

[15] In addition to Swiss-Prot, UniProtKB has a database called TrEMBL, which is larger but contains computationally annotated proteins whose existence are largely not proven. More details are provided in the Data Appendix.

[16] MMseqs2 can be downloaded from https://github.com/soedinglab/MMseqs2. More details on MMseqs2 are provided in the Data Appendix.

sequence identity (Schmidberger et al. 2010; Phenix), I chose a threshold of 30% sequence identity to group all proteins in Swiss-Prot into mutually exclusive clusters. I then restricted the sample to clusters with at least one human protein and clusters that had at least one protein discovered by 1998, the year before my panel begins. More details on sample construction can be found in the Data Appendix.

5. Empirical Strategy

5.1 Main Specification

As described in Section 3.2, since MR relies on having similar, previously solved structures as templates, MR only works in clusters of proteins with previously solved structures (i.e., bright clusters) and does not work for clusters of proteins that have not yet been structurally characterized (i.e., dark clusters).

This enables me to employ a difference-in-differences approach and estimate the following regression equation to examine the impact of MR:

$$Y_{ct} = \beta_0 + \beta_1 PostMR_t \times Bright_c + \delta_t + \gamma_c + \varepsilon_{ct} \quad (1)$$

Y_{ct} is the total number of structures that gets solved in cluster c in year t. $PostMR_t$ is an indicator variable that turns one after the arrival of MR in 2003, and $Bright_c$ is an indicator variable for bright clusters, defined as whether the cluster had at least one structure by 1998.[17] δ_t are calendar-year fixed effects, and γ_c are cluster fixed effects. β_1 is the coefficient of interest and can be interpreted as the impact of MR on the number of solved structures. Standard errors are clustered at the cluster level.

In order for the coefficient β_1 to capture the causal impact of MR, parallel trends assumption must hold: in the absence of MR, trends in outcomes between bright and dark clusters must have

[17] The treatment variable, $Bright_c$, is defined as whether the cluster had a structure by 1998 (the year before my panel begins) instead of 2003 (when MR arrived). If $Bright_c$ is defined using the year 2003, then the treatment is mechanically correlated with the outcome variable (the number of structures being solved each year) in the pre-period from 1999 to 2003 since the treatment is a lagged outcome of the pre-period. The panel was chosen to begin in 1999 because this is (i) early enough to yield at least five years of pre-period before the introduction of MR, but (ii) late enough that there has been some accumulation of prior solved structures in the PDB (6% of structures that will eventually be deposited at the PDB by 2019 had accumulated by 1998).

been the same, conditional on cluster fixed effects and year fixed effects (as well as time-varying controls that I use in some specifications). I discuss this concern in detail in Sections 6.1 and 6.4.

5.2 Descriptive Statistics

As shown in Table 1, my sample consists of 6,944 clusters, with 9% of the clusters classified as bright. Not surprisingly, in terms of levels, bright and dark clusters differ on several characteristics when MR arrived. First, bright clusters are on average older and bigger. Second, a higher share of the proteins in the bright clusters have characteristics that make them more amenable to crystallization and hence easier to solve: bright clusters contain proteins that are less likely to be membrane, disordered, or have compositional bias.[18] Proteins in the bright clusters are also on average shorter in sequence length. Third, while bright clusters had more publications and drugs related to their proteins, dark clusters are higher in one measure of biological significance: the share of human proteins in the cluster (human proteins are of high interest to drug developers). This reassuringly suggests that dark clusters are not devoid of biological importance.

While these differences in levels do not threaten my difference-in-difference strategy (as long as there are no differences in trends in outcomes in the pre-period), in Section 6.4, I revisit these characteristics in a robustness analysis to develop a "predicted brightness" measure, where I control for pre-period traits related to crystallization feasibility and biological significance.

6. Main Results

6.1 Impact of MR on the Number of Solved Structures

I begin by examining how MR impacted the number of solved structures. As shown in Table 2, bright clusters got brighter (i.e., received more structures) after MR, relative to dark clusters. The outcome is the total number of solved structures in a cluster each year. Columns 1-2 report

[18] Membrane proteins are proteins that are found in (or interact with) cell membranes; these proteins tend to be flexible and partially hydrophobic, which make crystallization challenging. Proteins with intrinsically disordered regions (i.e., regions that do not adopt a well-defined structure) or extreme sequence length (very short or long) can also impede crystallization (Slabinski et al. 2007). Finally, compositional bias refers to whether the protein contains regions with overrepresented subsets of amino acids. Proteins are typically composed of twenty amino acids, but not all amino acids may show up equally. For example, QHQQQGQHHQHHHQQQQHH has a bias for the amino acids Q (glutamine) and H (histidine) (Harrison 2017). Compositional bias is associated with decreased crystallization potential.

the outcome after Log(+1) transformation,[19] while Columns 3-4 report the results in levels (scaled by the standard deviation). As reported in Column 1, bright clusters experienced a 7% increase in the number of solved structures after the arrival of MR, relative to dark clusters. Results in levels also indicate that bright clusters got brighter. Bright clusters received an increase of 0.744 annual number of structures after the arrival of MR, which translates to a 30.1% increase relative to the baseline standard deviation of 2.47.

I conduct several analyses to ensure that these results are being driven by MR. First, one concern is that bright clusters may be getting more structures not necessarily due to MR but because it is also getting increasingly bigger (i.e., more protein sequences are being discovered) relative to dark clusters. In Columns 2 and 4, I additionally control for time-varying cluster size while estimating Equation 1; the magnitude of the impact of MR remains similar and significant.

Second, the impact of MR should only show up in structures that were actually solved by MR. If I observe that bright clusters experienced an increase in both structures that were solved by MR and non-MR methods, this raises the concern that factors unrelated to MR may be causing bright clusters to get brighter. In Appendix Table 2, I confirm that MR only impacts structures that were solved by MR and does not impact structures that were not solved by MR.

Third, since MR needs just one previously solved structure in order to work, the impact of MR should be stronger when comparing dark clusters versus bright clusters with a single previously solved structure, and weaker when comparing bright clusters with a single structure versus bright clusters with multiple structures. Appendix Table 3 shows this exact result. I split the bright clusters into whether they had just a single or multiple previously solved structures. I then compare the impact of MR, comparing dark versus bright clusters with just a single structure (Column 2) and comparing bright clusters with just a single structure versus multiple structures (Column 3). The impact of MR is stronger in Column 2 relative to Column 3.

Fourth, and most importantly, to asses pre-period trends, I show an event studies version of Equation 1, replacing the single $PostMR_t$ indicator with indicators for every year before and after the introduction of MR. Figure 4 plots the dynamic effects of MR on the number of solved structures. Reassuringly, in both Panels A (Log(+1) transformation) and B (levels), there appears to be no difference in pre-trends between bright and dark clusters. Moreover, the impact of MR is sustained over the entire sample period: bright clusters got brighter and brighter.

[19] In Appendix Table 1, I provide a robustness analysis using inverse hyperbolic sine transformation. Results remain similar.

6.2 Impact of MR on the Quality of Solved Structures

MR deceased the cost of solving structures in well-explored, bright clusters, and, not surprisingly, increased the volume of structures in those areas. But what about quality? In the next set of results, I investigate how MR impacted the quality of solved structures.

I distinguish between two dimensions of quality: execution (how meticulously a project was completed) versus importance (whether a project led to a novel insight). While I provide below measures of execution and importance in the specific setting of structural biology, these are general dimensions of quality that apply to any innovative activity: the innovation should be well-executed, but it should also solve an important problem.

The impact of analogical reasoning on quality is not immediately obvious. On one hand, analogical reasoning has the power to make "long jumps" (Kauffman 1993; Kneeland, Schilling, and Aharonson 2020) to explore a novel domain and discover creative opportunities. But it may be more challenging to rigorously execute innovations stemming from analogies because analogies are rooted in correlations, not precise causal logic. On the other hand, one of the pitfalls of analogies is that their need for templates could cause fixation and steer the direction of innovation towards areas with templates, even if they are less fruitful. In particular, analogy-based technologies like MR and supervised machine learning require digitized training data of past successful innovations. Analogies may thus lead to "short jumps," landing on unexplored but near crowded areas where it may be harder to unearth new insights.

6.2.1 Execution

The first dimension of quality I examine is execution, and I take advantage of measures provided in the PDB called the R-free and resolution.[20] These are objective metrics used by the structural biology community to assess the technical execution—specifically, the accuracy and precision—of the structures (Kleywegt and Jones 1997).

The R-free refers to accuracy or goodness-of-fit: how well the model of the protein structure fits the observed experimental data. As discussed in Section 3.2, structural biologists build the atomic model of their protein structure from experimentally observed diffraction data. They then simulate diffraction patterns based on the model and compare the simulated diffractions to the

[20] Hill and Stein (2021) use the R-free and resolution as their main quality measures in their study of how competition affects the quality of scientific research. I interpret the R-free and resolution as indicating specifically the execution level of the structure.

experimentally observed patterns. The R-free can be improved as researchers undergo iterative refinement process of their model to better fit the experimental data.

Resolution refers to precision or the level of detail that can be found in the structure. Figure 2C shows an example of a protein (tyrosine 103 from myoglobin) at different resolutions, from a poor resolution where only general contours are visible to a resolution where individual atoms can be plotted. Resolution depends on the degree of order in the crystallized protein. Researchers can improve the resolution by obtaining high-ordered crystals (proteins that are packed and aligned identically in the crystal), which produce diffraction patterns with fine details.

With these measures, I investigate the impact of MR on execution in Table 3. Column 1 reports the same result as Column 1 of Table 2 and shows the impact of MR on the total number of structures solved in a cluster. Columns 2-4 decompose this result into terciles based on the structure's R-free values and investigate the impact of MR on the number of structures in the bottom tercile (Column 2), middle tercile (Column 3), and top tercile (Column 4) of R-free values. Columns 5-7 similarly report results using the resolution of the structures.

A clear pattern emerges: bright clusters especially received more structures that were well-executed. For the R-free, there was no difference between bright and dark clusters in the number of structures that were solved in the bottom tercile. In contrast, bright clusters received 14% more structures that were solved in the top tercile, relative to dark clusters. Likewise, for resolution, bright clustered received just 3% more structures from the bottom tercile, but 10% more structures from the top tercile.

6.2.2 Importance

The second dimension of quality is the scientific importance of the structure: did the structure lead to a novel insight about a biological process? When the PDB was established in 1971 with only seven structures in its database, every new structure provided valuable information. However, as the PDB grew, it became no longer enough to just solve structures for the sake of solving them. As early as 1994, the editors of *Nature Structural Biology* advocated in their inaugural issue, "[T]he static image of the molecule is rarely an end in itself, but rather a beginning of comprehension" (Nature Structural Biology 1994). Through additional biochemistry or cell biology experiments, structural biologists try to explicitly link a protein's function to its potential function to understand the role the protein plays in various biological processes (Cassiday 2014). To evaluate whether a structure led to a new biological understanding, I present below three measures.

First, did the structure lead to a publication? Structures that were simply deposited at the PDB without a corresponding publication to explain their biological significance are structures that contributed very little, if at all, to revealing biological insights. As a prominent researcher at Yale once declared, "The fact is that protein structures come alive intellectually only when they are connected with [other data] indicating what they do" (Moore 2007). There could be several reasons why some structures do not have accompanying publications. One structural biologist I interviewed noted that a "stamp collection" of structures are sometimes needed to win grants from funding agencies. Some of these structures are also from structural genomics consortiums, whose goals are to catalogue as many types of structures as possible without necessarily explicating them (Petsko 2007; Hill and Stein 2021).

Second, was the structure cited by Swiss-Prot? The Swiss-Prot database contains extensive annotations about a protein's function and provides references behind each functional annotation. Importantly, the references are added manually by experts who follow well-defined curation protocols, undergo quality checks, and are updated as new data becomes available, ensuring that selection of these references are impartial.

Finally, I measure whether the structure got cited by a patent, with the assumption that the protein must have led to enough functional insights in order for an inventor to develop commercial applications. I leverage data from Marx and Fuegi (2020) on patent citations to scientific articles to identify structures with papers that were cited by at least one patent.

Bright clusters especially received more structures that did *not* reveal functional insights. As shown in Table 4, bright clusters received 9% more *unpublished* structures, which have no accompanying articles that describe their function (Column 2). In contrast, bright clusters experienced a relative decrease in the number of structures that were cited by a patent (Column 4), and there were no differences between bright and dark clusters in the number of structures that were cited by Swiss-Prot (Column 6)—which are the set of structures that are the most likely to have yielded functional insights.

6.3 How Did the Scientific Community Receive MR?

How did the scientific community receive this shift in research direction as a result of MR? Did the scientific community value the fact that MR led to more structures that are well-executed? Or did the community find these types of structures less valuable since they ultimately did not lead to new functional insights?

To examine this question, I use standard measures of publication impact (citations and journal impact factor).[21] A natural question is why I interpret these publication measures as different from my measures on functional insights. My earlier measures (whether the structure has an accompanying publication and whether it was cited by Swiss-Prot or a patent) assess a specific fact: did the structure lead to some kind of function insight? In contrast, journal impact factor and publication citation serve as proxies for how the scientific community typically rewards a piece of research and can be due to either the technical execution or functional insights of the research, which is hard to disentangle. For instance, it is unclear whether a structural biology paper was published in a prestigious journal because the structure was technically well-executed or because it led to a new biological understanding (or a combination).

The idea behind this analysis is that if the scientific community valued execution more, they would have rewarded the well-executed structures in bright clusters by publishing them in prestigious journals and highly citing them. In contrast, if the scientific community valued functional insights more, they would have punished the structures in bright clusters by publishing them in less prestigious journals and citing them less.

In Table 5, I investigate the effect of MR on the publication impact of the structures, in terms of the mean number citations the structure's publication received and the journal impact factor. In Columns 3-5, I decomposed the total number of solved structures in a cluster into terciles based on citations, while in Columns 8-10, I decomposed the total number of structures into terciles based on the journal impact factor.[22]

Bright clusters especially received more structures with *less* publication impact. After the arrival of MR, bright clusters received approximately 7% more structures that were either unpublished or published with very few citations, relative to dark clusters. In contrast, bright clusters had a 2% decline in the number of structures with the highest number of citations. In terms of journal impact factor, bright clusters received 7% more structures that were published in the least prestigious journals, but there was no difference between bright and dark clusters in the number of solved structures that were published in the most prestigious journals. This suggests that the scientific community appears to believe that there has been a decline in the quality of research conducted as a result of MR.

[21] I linked the primary paper associated with each structure in the PDB to the PubMed data and obtained citation data from the Web of Science (specifically, the mean annual number of citations received by each structure, within the first five years of paper publication).
[22] Due to data availability, I restricted the panel to end in 2012 for the citation analyses and 2017 for the journal impact factor analyses.

Taken together, these results indicate that bright clusters received more structures overall and particularly structures that were well-executed. However, these structures tended to not lead to new functional insights or have a high publication impact.

6.4 Predicting Brightness

6.4.1 Specification with Predicted Brightness

The identification underpinning my difference-in-differences framework hinges on parallel trends assumption. While there was no evidence of pre-trends in the event studies as well as in other robustness analyses, there may still be concerns over whether bright and dark clusters were evolving on different trends for factors unrelated to the introduction of MR. For example, a particular concern is that proteins in the dark clusters cannot be crystallized and thus cannot be structurally characterized.

This concern is mitigated by the evidence in the bioinformatics literature, where it has been documented that while there are certain traits (e.g., membrane or disordered proteins) that indeed make a protein challenging to crystallize, the crystallization process remains an unpredictable art, rather than a science. For instance, Perdigão et al. (2015) surveyed the Swiss-Prot data to understand the features of dark proteins; they find that most of the dark proteins cannot be explained by the "usual suspects," that a majority of the dark proteins are in fact not membrane or disordered proteins. Other studies also point to the difficulty in predicting which proteins will crystallize (Elbasir et al. 2019; Terwilliger, Stuart, and Yokoyama 2009).

Since there are still some characteristics that are known to confound crystallization, I develop a predicted brightness measure, $Predicted\_Bright_c$, where I measure whether a cluster was predicted to be bright in 1998,[23] using the pre-period characteristics of the proteins in the cluster. I then modify Equation 1 to estimate the following:

$$Y_{ct} = \beta_0 + \beta_1 PostMR_t \times Bright_c + \beta_2 PostMR_t \times Predicted\_Bright_c + \delta_t + \gamma_c + \varepsilon_{ct} \qquad (2)$$

The thought experiment in Equation 2 is that I compare clusters of proteins that are similarly predicted to have their structures characterized by 1998 because they are ex-ante similar in traits related to biological importance and crystallization feasibility, but some clusters just happened to

[23] Recall that (actually) bright clusters are defined as whether they had a structure by 1998.

actually have characterized structures before the arrival of MR while other clusters did not. If I observe that only being predicted bright has an impact on the number of structures solved after MR (i.e., β_1 is non-significant but β_2 is significant), then there would be a concern that unobserved characteristics are driving bright clusters to both have had their structures characterized in 1998 and subsequent structure characterization after MR. However, if there is an added effect of being actually bright in addition to being predicted bright (i.e., β_1 is significant), then this reduces the concern of omitted variable bias.

6.4.2 Constructing Predicted Brightness

I first restrict the sample to proteins that were discovered by 1998 (n = 41,781 proteins). For each protein, I focus on several sets of characteristics as of 1998. First, I have characteristics on how hard it is to crystallize the protein: whether the protein is a membrane protein, disordered protein, has compositional bias, and has long sequence length. Second, I have characteristics on the biological importance of the proteins: which species the protein is from (2,814 indictors), the number of publications written about the protein (10 indicators), and the number of approved drugs that target the protein (8 indicators).[24] Finally, I have indicators for the year of when the protein was discovered (29 indicators). After dropping collinear variables, this translates to a total of 817 predictors.

I use Lasso to predict whether a protein is predicted to be bright (i.e., structurally characterized by 1998). Appendix Figure 1A shows the receiver operating characteristic (ROC) curve of the resulting prediction. The ROC curve plots the True Positive Rate (what share of actually bright proteins were correctly predicted to be bright?) against the False Positive Rate (what share of actually dark proteins were incorrectly predicted to be bright?). The area under the curve (AUC) of the ROC curve evaluates the performance of the prediction and can be interpreted as the probability that a random actually bright protein will have a higher predicted brightness than a random actually dark protein. The AUC can range from 0 to 1, and a general rule of thumb considers an AUC above 0.8 to be indicating high performance; the AUC of my prediction exercise is 0.91.

[24] Information on drugs is provided by DrugBank. This dataset provides comprehensive information on drugs at various development phases and their targets (i.e., proteins) and is freely available for academic use. A limitation of the free version of the data is that it only provides marketing dates for approved drugs, and there are no dates on when a drug entered pre-clinical or clinical trial phases.

I then use the fitted values to predict each protein's brightness in 1998. Appendix Figure 1B shows the distribution of this predicted brightness, by whether the protein was actually bright by 1998. There is variation and overlap in the distributions of predicted brightness between actually bright and dark proteins, suggesting that while there are some characteristics that may make some proteins more likely to be structurally determined, some proteins just happened to have structures actually determined by 1998, while other proteins with similarly predicted brightness did not. This supports the findings in the bioinformatics literature which also notes that many of the dark proteins cannot be explained by the usual factors that defy crystallization and that it is in fact difficult to predict crystallization. From this protein-level prediction, I aggregate up to the cluster-level by taking the sum of the predicted brightness of all proteins in each cluster to construct $Predicted\_Bright_c$.

6.4.3 Results with Predicted Brightness

I then estimate Equation 2 that additionally controls for $Post\text{-}MR_t \times Predicted\_Bright_c$, modifying my baseline difference-in-differences framework. Appendix Table 4 shows the results from estimating Equation 2. The coefficients on $Post\text{-}MR_t \times Bright_c$ remain positive and significant; among clusters predicted to be similarly bright, there is still an effect of being actually bright. While it may seem surprising that the coefficients on $Post\text{-}MR_t \times Predicted\_Bright_c$ are not significant, MR only works when there are actually solved prior structures in the cluster and should not work if the cluster is only predicted to be bright. Appendix Table 4 therefore supports the evidence that the arrival of MR indeed caused the number of solved structures in (actually) bright clusters to increase, relative to dark clusters, even among clusters that were ex-ante similarly predicted to be bright due to their traits related to biological importance or crystallization feasibility.

7. Discussion & Conclusion

This paper provides, to my knowledge, the first empirical study of how the automation of analogical reasoning may shape the direction of knowledge production. Using the setting of structural biology, which provides a unique window into the entire idea landscape, I study the introduction of an analogy-based technology, MR, which solves protein structures by relying on data of prior structures.

I find that MR increased the number of solved structures, specifically in bright clusters with already solved structure templates. This result from structural biology highlights the power of analogies in reducing the cost of producing innovation in certain areas, thereby shifting the direction of innovation. In particular, rather than building knowledge from scratch, analogies can provide shortcuts that enable knowledge workers to innovate in domains where there is no existing knowledge yet (e.g., structurally uncharacterized proteins) by importing knowledge from neighboring domains (e.g., borrowing structure information from similar, already solved proteins). Yet, because of their very need for templates, analogies can restrict knowledge workers into focusing on domains with neighbors (e.g., bright, already well-explored clusters).

One may argue that this shift in research direction due to analogies does not necessarily imply a decline in the quality of innovation, as analogies allow knowledge workers to quickly make progress in previously unexplored domains. An important question then is what quality of innovation is ultimately produced. My results suggest that structural biologists used the extra time gained from taking a shortcut with MR to improve the technical execution—the resolution and the goodness-of-fit—of the structures. These structures, however, had low scientific importance and publication impact. That is, at least in this specific setting, knowledge workers appear to use the imported knowledge from analogies to focus on incremental execution, rather than attempting to discover a fundamental insight.

Understanding this tradeoff of analogies can have crucial implications for the management of knowledge production, given the growing automation of analogical reasoning and, more broadly, data-driven exploration. A recent work by Hoelzemann et al. (2022) is close in spirit to this paper: using a laboratory experiment, the authors document the "streetlight" effect of data,[25] that when data reveals a satisfactory—but not the best—option, data can discourage workers from exploring further to reach the best. Extending this idea of the streetlight effect, I suggest a "snowballing" effect: analogies may steer the direction of innovation towards areas with templates, which in turn, gain more templates and thus become more amenable to analogies, while neglected (and potentially fruitful) areas without templates may never get attention.

This snowballing effect can manifest in several ways. In the case of structural biology, bright clusters got brighter and brighter after the arrival of MR; among clusters that were predicted to have similar crystallization potential and biological importance, clusters that happened to have structures before MR took off, while clusters without structures remained dark. This increasing

[25] The authors draw from the aphorism of the drunk looking for his keys under the streetlight, despite dropping the keys on the other side of the street, because "this is where the light is."

returns to training data can have profound strategic consequences for firms as well. For firms that produce products and services based on data, early entrants may use their control over data to crowd out competitors (Cockburn, Henderson, and Stern 2018; Bessen et al. 2022). At the frontier of AI, there are now even machine learning models (trained on real data) that generate synthetic data, which will be fed into other machine learning models, raising the possibility of amplifying the influence of the original training data (Zewe 2022).

A question still remains on what is the net value of increased innovative activities in the bright clusters. This is difficult to assess. First, my difference-in-differences framework is limited to reporting only the relative increase in solved structures between bright and dark clusters. This relative increase can mean either an overall increase in bright clusters or a reallocation of innovative efforts from dark to bright clusters. Second, while structures in the bright clusters may not have led to novel biological insights, the increased number of (well-executed) structures may still be valuable, especially for drug development (which requires precisely solved structures) as well as for serving as training data for machine learning algorithms.

Finally, while this paper focuses on how analogies can potentially constrain the direction of innovation, future work can explore other costs of analogies. In particular, because analogies do not build foundational knowledge from scratch, innovators may not fully understand the underlying mechanisms of how the target domain works. Overreliance on analogies may lead to knowledge domains with weak foundation, where only correlations accumulate and causal theories are neglected (Zittrain 2019; Tranchero 2023). Popular rhetoric often warns the danger of this "black box" nature of AI (and by extension, analogies). For instance, in drug discovery, analogical reasoning and pattern recognition can be employed to identify promising drug candidates based on prior approved drugs. A downside to this approach is that the drugs' mechanisms of action remain unknown, preventing drugmakers from anticipating side effects or applying the drugs for other diseases that may share the same mechanisms.

The setting of structural biology is currently facing its own AI revolution. In November 2020, Google's DeepMind team cracked a 50-year-old grand challenge in biology: to predict how a protein folds into its 3D structure from purely its sequence of amino acids. While MR helps with only one part of experimental structure solving, AlphaFold bypasses the need to conduct experiments at all. Celebrated as one of the most important applications of AI in science as of date, the source code of AlphaFold became publicly available in July 2021, followed by the release of a database of predicted structures a year later.

Despite its breakthrough, however, AlphaFold also serves as a reminder of the limitations of analogies. DeepMind's claim that AlphaFold has produced enough structures to cover the "entire protein universe" (Walsh 2022) must be qualified with important caveats. First, these are only *predicted* structures. When comparing AlphaFold's structures to experimentally determined structures, researchers have found that while many of these predicted structures can be remarkably accurate, some are still too inaccurate to be useful (Mullard 2021). Second, and most importantly, AlphaFold is limited by its training data, the PDB, and can only populate the protein structural space based on analogies to known PDB structures. In particular, protein structures can change in the presence of small molecule drugs. Because there could be up to one novemdecillion[26] small molecules (Reymond and Awale 2012), the PDB does not contain enough information on structures bound to small molecules for AlphaFold to predict how proteins might interact with drugs (Callaway 2022). Furthermore, diseases are often caused by mutations to proteins. Since these mutated proteins have no evolutionarily-related sequences, it is difficult for AlphaFold to predict their structures (Buel and Walters 2022; Callaway 2022). This is why, at least for now, many remain skeptical that AlphaFold will dramatically impact drug development.[27]

Finally, AlphaFold illustrates that analogies work by identifying patterns, not causal theories. While AlphaFold has advanced the ability to predict structures, scientists still have little understanding of the physics of *why* proteins fold into their shapes (Lowe 2022). The rise of AlphaFold may signal the trend of hypothesis-driven science turning into "data science." As one of the scientists in the field lamented, "We've focused too much on data and not enough on understanding . . [we may be] going away from human-conceived theories and models of natural phenomena to more data-driven methods and models" (Samuel 2019).

[26] A novemdecillion is equivalent to million billion billion billion billion billion billion billion (American Chemical Society 2012).

[27] While not about AlphaFold, Lou and Wu (2022) demonstrates the limits of AI in drug development; AI is less useful for developing drugs that are radically novel and have no known mechanisms of actions. In addition, a recent paper by Cavalli (2022) investigates how AlphaFold changed the organizational structure of academic labs in computational biology.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press.

Agrawal, Ajay, John McHale, and Alex Oettl. 2018. "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth." *NBER Working Paper* #24541.

American Chemical Society. 2012. "1 Million Billion Billion Billion Billion Billion Billion: Number of Undiscovered Drugs." *Phys.Org*, June 6, 2012. https://phys.org/news/2012-06-million-billion-undiscovered-drugs.html.

Anthony, Callen. 2021. "When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies." *Administrative Science Quarterly* 66 (4): 1173–1212.

Azoulay, Pierre, Jeffrey L Furman, Joshua L Krieger, and Fiona Murray. 2015. "Retractions." *The Review of Economics and Statistics* 97 (5): 1118–36.

Baldwin, Carliss Y., and Kim B. Clark. 1997. "Managing in an Age of Modularity." *Harvard Business Review*, 1997.

Bessen, James, Stephen Michael Impink, Lydia Reichensperger, and Robert Seamans. 2022. "The Role of Data for AI Startup Growth." *Research Policy* 51 (5).

Bingham, Christopher B., and Steven J. Kahl. 2012. "How to Use Analogies to Introduce New Ideas." *MIT Sloan Management Review* 56 (2): 10–12.

———. 2013. "The Process of Schema Emergence: Assimilation, Deconstruction, Unitization and the Plurality of Analogies." *Academy of Management Journal* 56 (1): 14–34.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, NY: W. W. Norton & Company.

Buel, Gwen R., and Kylie J. Walters. 2022. "Can AlphaFold2 Predict the Impact of Missense Mutations on Structure?" *Nature Structural & Molecular Biology* 29 (1): 1–2.

Callaway, Ewen. 2022. "What's Next for AlphaFold and the AI Protein-Folding Revolution." *Nature*, April 13, 2022. https://www.nature.com/articles/d41586-022-00997-5.

Cassiday, Laura. 2014. "Structural Biology: More than a Crystallographer." *Nature* 505 (7485): 711–13.

Cavalli, Gabriel. 2022. "How Scientific Organizations React to Novel Methodological Advances: The Impact of AlphaFold V1." *Working Paper*.

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal. 2020. "Machine Learning and Human Capital Complementarities: Experimental Evidence on Bias Mitigation." *Strategic Management Journal* 41 (8): 1381–1411.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern. 2018. "The Impact of Artificial Untelligence on Innovation." *NBER Working Paper* #24449.

Cornelissen, Joep P., and Jean S. Clarke. 2010. "Imagining and Rationalizing Opportunities: Inductive Reasoning and the Creation and Justification of New Ventures." *Academy of Management Review* 35 (4): 539–57.

Cowgill, Bo, Fabrizio Dell'acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics." *Proceedings of the 21st ACM Conference on Economics and Computation*, 679–81.

Cowgill, Bo, and Catherine E. Tucker. 2020. "Algorithmic Fairness and Economics." *Columbia Business School Research Paper*. https://ssrn.com/abstract=3361280.

Crick, Francis. 1990. *What Mad Pursuit: A Personal View of Scientific Discovery*. London, UK: Penguin.

Doerr, Allison. 2014. "A Method Ahead of Its Time." *Nature* 511 (Suppl 7509): 13.

Dunbar, Kevin. 1999. "How Scientists Build Models In Vivo Science as a Window on the Scientific Mind."

Model-Based Reasoning in Scientific Discovery, 85–99.

———. 2000. "How Scientists Think in the Real World: Implications for Science Education." *Journal of Applied Developmental Psychology* 21 (1): 49–58.

Elbasir, Abdurrahman, Balasubramanian Moovarkumudalvan, Khalid Kunji, Prasanna R Kolatkar, Raghvendra Mall, Halima Bensmail, and John Hancock. 2019. "DeepCrystal: A Deep Learning Framework for Sequence-Based Protein Crystallization Prediction." *Bioinformatics* 35 (13): 2216–25.

Etzion, Dror, and Fabrizio Ferraro. 2010. "The Role of Analogy in the Institutionalization of Sustainability Reporting." *Organization Science* 21 (5): 1092–1107.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a Map in Technological Search." *Strategic Management Journal* 25 (8–9): 909–28.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research." *American Economic Review* 101 (5): 1933–63.

Furman, Jeffrey L., and Florenta Teodoridis. 2020. "Automation, Research Technology, and Researchers' Trajectories: Evidence from Computer Science and Electrical Engineering." *Organization Science* 31 (2): 330–54.

Gavetti, Giovanni, Daniel A. Levinthal, and Jan W. Rivkin. 2005. "Strategy Making in Novel and Complex Worlds: The Power of Analogy." *Strategic Management Journal* 26 (8): 691–712.

Gavetti, Giovanni, and Jan W. Rivkin. 2005. "How Strategists Really Think: Tapping the Power of Analogy." *Harvard Business Review* 83 (4): 54–63.

Gentner, Dedre. 1982. "Structure Mapping: A Theoretical Framework for Analogy." *Cognitive Science* 7 (2): 155–70.

Gentner, Dedre, Keith J. Holyoak, and Boicho N. Kokinov. 2001. *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gilboa, Itzhak, Larry Samuelson, and David Schmeidler. 2015. *Analogies and Theories: Formal Models of Reasoning*. New York, NY: Oxford University Press.

Glaser, Vern L., Peer C. Fiss, and Mark Thomas Kennedy. 2016. "Making Snowflakes like Stocks: Stretching, Bending, and Positioning to Make Financial Market Analogies Work in Online Advertising." *Organization Science* 27 (4): 1029–48.

Greenblatt, Wesley. 2021. "Building on Solid Ground: Foundational Knowledge and the Dynamics of Innovation." *SSRN Working Paper*. https://ssrn.com/abstract=3919866.

Hacking, Ian. 2014. *Why Is There Philosophy of Mathematics at All?* Cambridge, UK: Cambridge University Press.

Hargadon, Andrew B., and Yellowlees Douglas. 2001. "When Innovations Meet Institutions: Edison and the Design of the Electric Light." *Administrative Science Quarterly* 46 (3): 476–501.

Harrison, Paul M. 2017. "FLPS: Fast Discovery of Compositional Biases for the Protein Universe." *BMC Bioinformatics* 18 (1): 1–9.

Hauser, Maria, Martin Steinegger, and Johannes Söding. 2016. "MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets." *Bioinformatics* 32 (9): 1323–30.

Henderson, Rebecca. 1994. "The Evolution of Integrative Capability: Innovation in Cardiovascular Drug Discovery." *Industrial and Corporate Change* 3 (3): 607–30.

Hesse, Mary B. 1966. *Models and Analogies in Science*. Notre Dame, IN: Univ Notre Dame Press.

Hill, Robert C., and Michael Levenhagen. 1995. "Metaphors and Mental Models: Sensemaking and Sensegiving in Innovative and Entrepreneurial Activities." *Journal of Management* 21 (6): 1057–74.

Hill, Ryan, and Carolyn Stein. 2020. "Scooped! Estimating Rewards for Priority in Science." *Working Paper*.

———. 2021. "Race to the Bottom: Competition and Quality in Science." *Working Paper*.

Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero. 2022. "The Streetlight

Effect in Data-Driven Exploration." *Working Paper*.
Hofstadter, Douglas R, and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York, NY: Basic Books.
Holyoak, Keith J., and Paul Thagard. 1996. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.
Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76 (1): 283–317.
Kaplan, Sarah, and Wanda J Orlikowski. 2013. "Temporal Work in Strategy Making" 24 (4): 965–95.
Kaplan, Sarah, and Keyvan Vakili. 2015. "The Double-Edged Sword of Recombination in Breakthrough Innovation." *Strategic Management Journal* 36 (10): 1435–57.
Kauffman, Stuart A. 1993. *The Origins of Order: Self-Organization and Selection in Evolution*. New York, NY: Oxford University Press.
Kittur, Aniket, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E. Kraut, and Dafna Shahaf. 2019. "Scaling up Analogical Innovation with Crowds and AI." *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 1870–77.
Kleywegt, G. J., and T. A. Jones. 1997. "Model Building and Refinement Practice." *Methods in Enzymology* 277: 208–30.
Kneeland, Madeline K., Melissa A. Schilling, and Barak S. Aharonson. 2020. "Exploring Uncharted Territory: Knowledge Search Processes in the Origination of Outlier Innovation." *Organization Science* 31 (3): 535–57.
Koonin, Eugene V., Yuri I. Wolf, and Georgy P. Karev. 2002. "The Structure of the Protein Universe and Genome Evolution." *Nature* 420 (6912): 218–23.
Levinthal, Daniel A. 1997. "Adaptation on Rugged Landscapes." *Management Science* 43 (7): 934–50.
Lou, Bowen, and Lynn Wu. 2022. "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms." *Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3524985.
Lowe, Derek. 2022. "The Law of Conservation of Data." *Chemistry World*, January 11, 2022. https://www.chemistryworld.com/opinion/the-law-of-conservation-of-data/4014927.article.
Madrigal, Alexis C. 2019. "The Servant Economy." *The Atlantic*, March 6, 2019. https://www.theatlantic.com/technology/archive/2019/03/what-happened-uber-x-companies/584236/.
Mannucci, Pier Vittorio. 2017. "Drawing Snow White and Animating Buzz Lightyear: Technological Toolkit Characteristics and Creativity in Cross-Disciplinary Teams." *Organization Science* 28 (4): 711–28.
Martins, Luis L., Violina P. Rindova, and Bruce E. Greenbaum. 2015. "Unlocking the Hidden Value of Concepts: A Cognitive Approach to Business Model Innovation." *Strategic Entrepreneurship Journal* 9: 99–117.
Marx, Matt, and Aaron Fuegi. 2020. "Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles." *Strategic Management Journal* 41 (9): 1572–94.
Mill, John Stuart. 1974. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Toronto, Canada: University of Toronto Press.
Mirdita, Milot, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. 2017. "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments." *Nucleic Acids Research* 45 (D1): D170–76.
Miric, Milan, Hakan Ozalp, and Dogukan Yilmaz. 2021. "Tradeoffs to Using Standardized Tools: An Innovation Enabler or Creativity Constraint?" *USC Marshall School of Business Research Paper*.

https://ssrn.com/abstract=3358801.
Moore, Peter B. 2007. "Let's Call the Whole Thing Off: Some Thoughts on the Protein Structure Initiative." *Structure* 15 (11): 1350–52.
Mullard, Asher. 2021. "What Does AlphaFold Mean for Drug Discovery?" *Nature Reviews Drug Discovery* 20 (10): 725–27.
Nagle, Frank, and Florenta Teodoridis. 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.
Nature Structural Biology. 1994. "The Changing Structure of Biology." *Nature Structural Biology* 1 (1).
Nelson, Richard R, and Sidney G Winter. 1982. "The Schumpeterian Tradeoff Revisited." *The American Economic Review* 72 (1): 114–32.
Parnas, D. L. 1972. "On the Criteria to Be Used in Decomposing Systems into Modules." *Communications of the ACM* 15 (12): 1053–58.
Patel, Onisha, Isabelle Lucet, and Michael Roy. 2020. "'Like a Key to a Lock': How Seeing the Molecular Machinery of the Coronavirus Will Help Scientists Design a Treatment." *The Conversation*, March 24, 2020. https://theconversation.com/like-a-key-to-a-lock-how-seeing-the-molecular-machinery-of-the-coronavirus-will-help-scientists-design-a-treatment-134135.
Perdigão, Nelson, Julian Heinrich, Christian Stolte, Kenneth S. Sabir, Michael J. Buckley, Bruce Tabor, Beth Signal, et al. 2015. "Unexpected Features of the Dark Proteome." *Proceedings of the National Academy of Sciences of the United States of America* 112 (52): 15898–903.
Petsko, Gregory A. 2007. "An Idea Whose Time Has Gone." *Genome Biology* 8 (6): 1–3.
Phenix. n.d. "Overview of Molecular Replacement in Phenix." Accessed September 1, 2022. https://phenix-online.org/documentation/reference/mr_overview.html.
Pollack, John. 2014. *Shortcut: How Analogies Reveal Connections, Spark Innovation, and Sell Our Greatest Ideas*. New York, NY: Gotham Books.
Ramakrishnan, Venki. 2018. *Gene Machine: The Race to Decipher the Secrets of the Ribosome*. New York, NY: Basic Books.
Reymond, Jean Louis, and Mahendra Awale. 2012. "Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database." *ACS Chemical Neuroscience* 3 (9): 649–57.
Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5).
Samuel, Sigal. 2019. "How One Scientist Coped When AI Beat Him at His Life's Work." *Vox*, February 15, 2019. https://www.vox.com/future-perfect/2019/2/15/18226493/deepmind-alphafold-artificial-intelligence-protein-folding.
Sanchez, Ron, and Joseph T. Mahoney. 1996. "Modularity, Flexibility, and Knowledge Management in Product and Organization Design." *Strategic Management Journal* 17 (Suppl Winter): 63–76.
Sautoy, Marcu du. 2021. *Thinking Better: The Art of the Shortcut in Math and Life*. New York: Basic Books.
Scapin, Giovanna. 2013. "Molecular Replacement Then and Now." *Acta Crystallographica Section D: Biological Crystallography* 69 (11): 2266.
Schmidberger, Jason W., Mark A. Bate, Cyril F. Reboul, Steve G. Androulakis, Jennifer M.N. Phan, James C. Whisstock, Wojtek J. Goscinski, David Abramson, and Ashley M. Buckle. 2010. "MrGrid: A Portable Grid Based Molecular Replacement Pipeline." *PLOS ONE* 5 (4): e10049.
Schumpeter, Joseph. 1934. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.
Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5 (1): 29–41.
Simcoe, Timothy. 2015. "Modularity and the Evolution of the Internet." In *Economic Analysis of the Digital Economy*, 21–47. University of Chicago Press.

Slabinski, Lukasz, Lukasz Jaroszewski, Ana P.C. Rodrigues, Leszek Rychlewski, Ian A. Wilson, Scott A. Lesley, and Adam Godzik. 2007. "The Challenge of Protein Structure Determination--Lessons from Structural Genomics." *Protein Science : A Publication of the Protein Society* 16 (11): 2472–82.

Spenser, Jay. 2008. *The Airplane: How Ideas Gave Us Wings*. New York, NY: HarperCollins.

Steinegger, Martin, and Johannes Söding. 2018. "Clustering Huge Protein Sequence Sets in Linear Time." *Nature Communications* 9 (1): 1–8.

Teodoridis, Florenta. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64 (8): 3625–48.

Terwilliger, Thomas C, David Stuart, and Shigeyuki Yokoyama. 2009. "Lessons from Structural Genomics." *Annual Review of Biophysics* 38: 371–83.

Thomke, Stefan H. 1998. "Managing Experimentation in the Design of New Products." *Management Science* 44 (6): 743–62.

Tranchero, Matteo. 2023. "Data-Driven Search and Innovation in Well-Defined Technological Spaces." *Working Paper*.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science* 342 (6157): 468–72.

Walsh, Bryan. 2022. "Finally, an Answer to the Question: AI — What Is It Good For?" *Vox*, August 3, 2022. https://www.vox.com/future-perfect/2022/8/3/23288843/deepmind-alphafold-artificial-intelligence-biology-drugs-medicine-demis-hassabis.

Weitzman, Martin L. 1998. "Recombinant Growth." *The Quarterly Journal of Economics* 113 (2): 331–60.

Zewe, Adam. 2022. "When It Comes to AI, Can We Ditch the Datasets?" *MIT News*, March 15, 2022. https://news.mit.edu/2022/synthetic-datasets-ai-image-classification-0315.

Zhuo, Ran. 2022. "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs." *Working Paper*.

Zittrain, Jonathan. 2019. "The Hidden Costs of an Automated Thinking." *The New Yorker*, July 23, 2019. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.

Figures & Tables

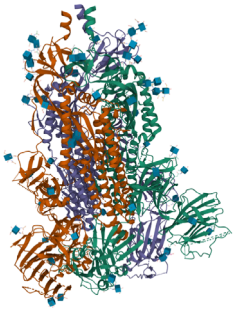
FIGURE 1. A TAXONOMY OF SHORTCUTS

	Vertical Shortcuts (Used in older domains with foundational knowledge)	**Horizontal Shortcuts (Used in newer domains without foundational knowledge)**
Understanding	**Summaries** Provide synopses of foundational knowledge underlying a domain e.g., Wikipedia	**Conceptual Analogies** Understand a new domain by importing patterns from a known domain e.g., biomimetics
Application	**Calculators** Execute instructions based on foundational knowledge underlying a domain e.g., Stata	**Automated Analogies** Apply patterns to a new domain from a known domain e.g., TensorFlow

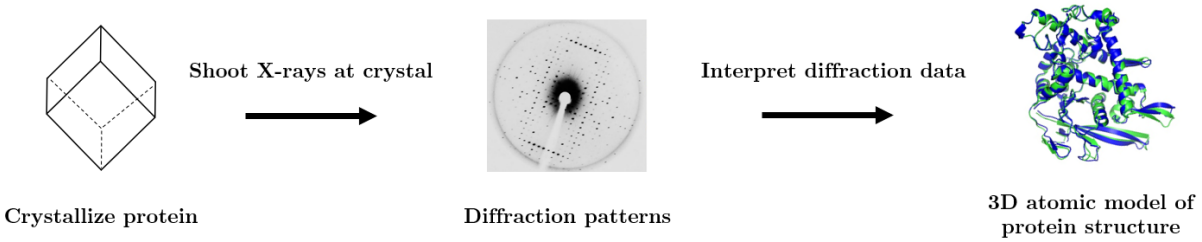
NOTES: This figure provides a taxonomy of shortcuts that can be used in cumulative knowledge production.

FIGURE 2. STRUCTURAL BIOLOGY

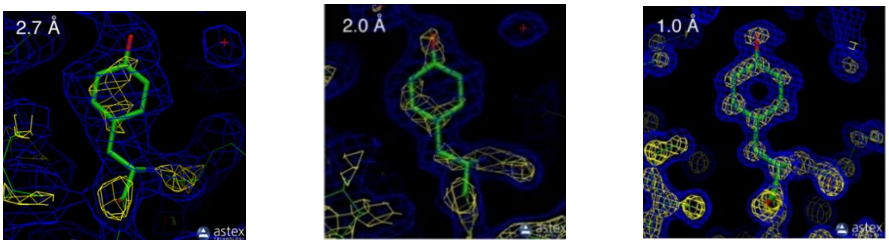
A. Structure of the SARS-CoV-2 Spike Glycoprotein



B. Steps of Crystallography

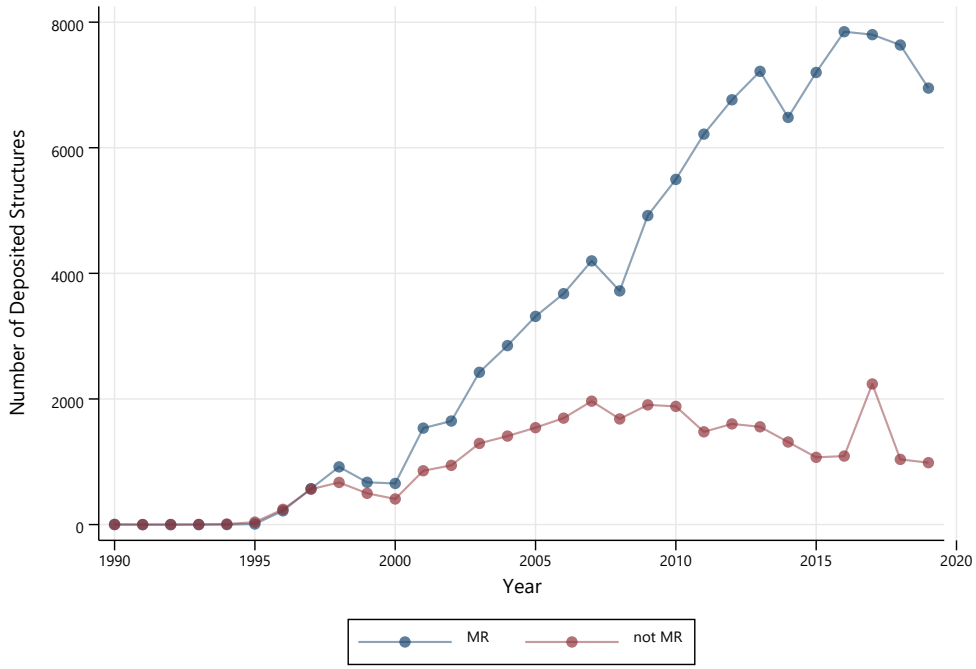


C. Resolution



NOTES: Panel A shows the structure of a spike protein on the surface of the coronavirus (PDB entry 6VYB; source: https://www.rcsb.org/structure/6VYB). Panel B shows the three main steps of crystallography; this paper focuses on the automation of solving the "phase problem" that occurs during the interpretation of the diffraction data. Panel C shows an example of the electron density map behind the structure of tyrosine 103 from myoglobin, at three different resolutions; lower resolution is better and shows finer details (source: https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/resolution).

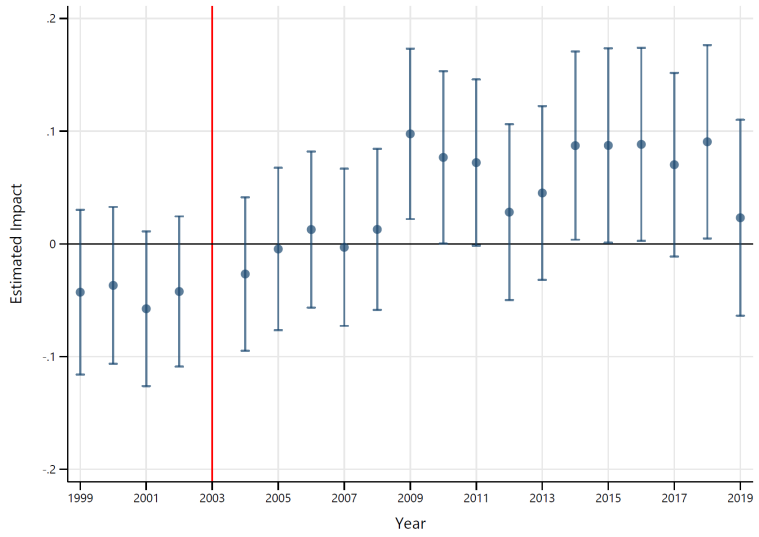
FIGURE 3. NUMBER OF STRUCTURES SOLVED BY MOLECULAR REPLACEMENT



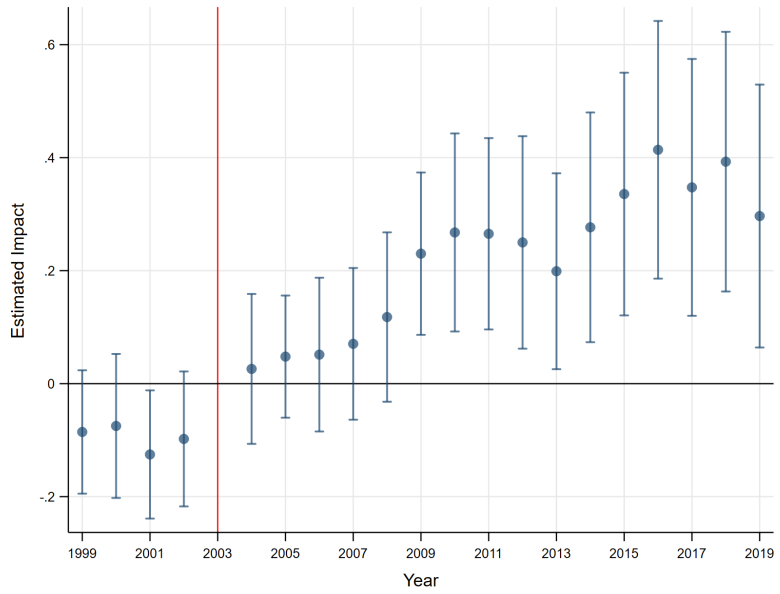
NOTES: This figure plots the number of X-ray crystallography structures in the Protein Data Bank that were solved by molecular replacement (MR) vs. non-MR methods.

FIGURE 4. EVENT STUDY: IMPACT OF MR ON NUMBER OF SOLVED STRUCTURES

Panel A. Log(+1) Transformation



Panel B. Levels



NOTES: This figure shows the impact of MR on the number of solved structures. The figure plots the coefficients and 95% confidence intervals from estimating a modified, event studies version of Equation 1 that replaces the pooled $PostMR_t$ indicator with separate indicators for every year before and after the arrival of MR. The outcome is the total annual number of solved structures in a cluster; Panel A reports the outcome after Log(+1) transformation, while Panel B reports the outcome in levels. The unit of analysis is a cluster × year, and the sample consists of 6,944 clusters, which translates to 145,824 cluster-years.

TABLE 1. SUMMARY STATISTICS

	Bright			Dark		
	Mean	Median	SD	Mean	Median	SD
Discovery Year	1983.47	1986.00	7.7	1993.11	1995.00	5.1
Cluster Size	39.73	15	81.6	7.66	4	15
Protein Production Feasibility						
Disorder	0.15	0.1	0.2	0.24	0.1	0.2
Membrane	0.02	0	0.1	0.04	0	0.1
Compositional Bias	0.01	0	0	0.03	0	0.1
Sequence Length	513.61	339.1	1,438.50	634.97	457.3	675.5
Biological Importance						
% of Cluster that is Human	0.18	0.1	0.2	0.36	0.3	0.2
N of Publications	142.47	75	199.2	26.79	12	57.4
N of Approved Drugs	3.63	0	14.5	1.23	0	21.8
N of Solved Structures per Year	1.99	0	5.64	0.23	0	1.78
N of Clusters		649			6,295	

NOTES: This table provides the summary characteristics of clusters when MR was introduced. The sample consists of 6,944 clusters, of which 649 are classified as "bright" (i.e., had at least one structure in 1998) and 6,295 are classified as "dark."

TABLE 2. IMPACT OF MR ON NUMBER OF SOLVED STRUCTURES

	(1) Log(+1)	(2) Log(+1)	(3) Levels	(4) Levels
VARIABLES	N Structures	N Structures	N Structures	N Structures
Post-MR × Bright	0.071***	0.065***	0.301***	0.292***
	(0.018)	(0.018)	(0.052)	(0.052)
Cluster Size		0.013**		0.019*
		(0.006)		(0.011)
R-squared	0.471	0.471	0.400	0.400
Calendar-year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster, reported after Log(+1) transformation (Columns 1-2) or in levels scaled by the standard deviation (Columns 3-4). The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects; Columns 2 and 4 additionally control for time-varying (standardized) cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

TABLE 3. IMPACT OF MR ON EXECUTION

VARIABLES	(1) All Structures	(2) *R-Free* Bottom Tercile	(3) *R-Free* Middle Tercile	(4) *R-Free* Top Tercile	(5) *Resolution* Bottom Tercile	(6) *Resolution* Middle Tercile	(7) *Resolution* Top Tercile
Post-MR × Bright	0.071***	-0.001	0.077***	0.133***	0.034***	0.047***	0.096***
	(0.018)	(0.011)	(0.012)	(0.013)	(0.011)	(0.012)	(0.013)
R-squared	0.471	0.381	0.397	0.403	0.365	0.414	0.427
Calendar-year FE	YES	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824	145,824	145,824	145,824

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different terciles of execution level (a structure's level of execution can be defined in terms of its R-free value and resolution). The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-4 decompose this result by examining the number of solved structures in the bottom (Column 2), middle (Column 3), and top terciles (Column 4) with respect to the structures' R-free values. Columns 5-6 similarly decompose the number of solved structures into terciles based on their resolution. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

TABLE 4. IMPACT OF MR ON FUNCTIONAL INSIGHTS

VARIABLES	(1) All Structures	(2) Unpublished Structures	(3) Published Structures *Not Cited by Patent*	(4) Published Structures *Cited by Patent*	(5) Published Structures *Not Fxn Annotated*	(6) Published Structures *Fxn Annotated*
Post-MR × Bright	0.071***	0.085***	0.117***	-0.064***	0.039**	0.006
	(0.018)	(0.009)	(0.016)	(0.011)	(0.017)	(0.004)
R-squared	0.471	0.221	0.389	0.350	0.473	0.117
Calendar-year FE	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824	145,824	145,824	145,824

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different levels of functional insights (a structure is considered to have contributed to new insights about a protein's function if it is published in a scientific article and additionally cited by a patent or by the functional summary section of Swiss-Prot). The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-4 decompose this result into the number of solved structures that do not get published in a scientific article (Column 2), the number of solved structures that are published but not cited by a patent (Column 3), and the number of solved structures that are both published and cited by a patent (Column 4). Columns 5 and 6 parallel Columns 3 and 4 but decompose the number of solved structures based on whether they were cited by the functional summary section of Swiss-Prot. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

TABLE 5. IMPACT OF MR ON PUBLICATION IMPACT

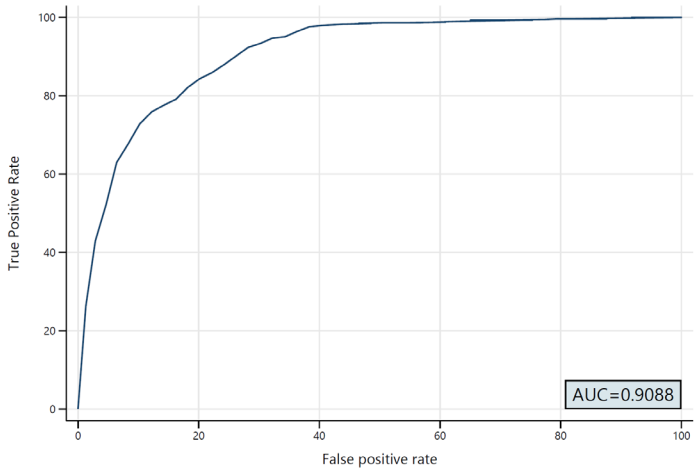
	Citations					*Journal Impact Factor*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	All Structures	Unpublished Structures	*Citations* Published Structures (Bottom Tercile)	*Citations* Published Structures (Middle Tercile)	*Citations* Published Structures (Top Tercile)	All Structures	Unpublished Structures	*JIF* Published Structures (Bottom Tercile)	*JIF* Published Structures (Middle Tercile)	*JIF* Published Structures (Top Tercile)
Post-MR × Bright	0.056***	0.075***	0.067***	0.012	-0.022**	0.070***	0.083***	0.070***	0.006	0.014
	(0.017)	(0.009)	(0.012)	(0.012)	(0.010)	(0.018)	(0.009)	(0.012)	(0.012)	(0.009)
R-squared	0.487	0.227	0.401	0.330	0.331	0.477	0.223	0.378	0.363	0.285
Calendar-year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
N of clusters	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944	6,944
N of cluster-years	97,216	97,216	97,216	97,216	97,216	131,936	131,936	131,936	131,936	131,936

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures at different terciles of publication impact. A structure's publication impact is measured as the mean number of citations and the journal impact factor (JIF). The unit of analysis is a cluster × year. Due to data availability, the panel ends in 2012 for the citation analyses and in 2017 for the JIF analyses. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Column 1 parallels Column 1 in Table 2 and reports the total number of solved structures. Columns 2-5 decompose this result into the number of solved structures that do not get published in a scientific article (Column 2) and the number of solved structures that are published and in bottom (Column 3), middle (Column 4), or top (Column 5) terciles in terms of citation impact. Columns 6-7 similarly decompose the number of solved structures into terciles based on JIF. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

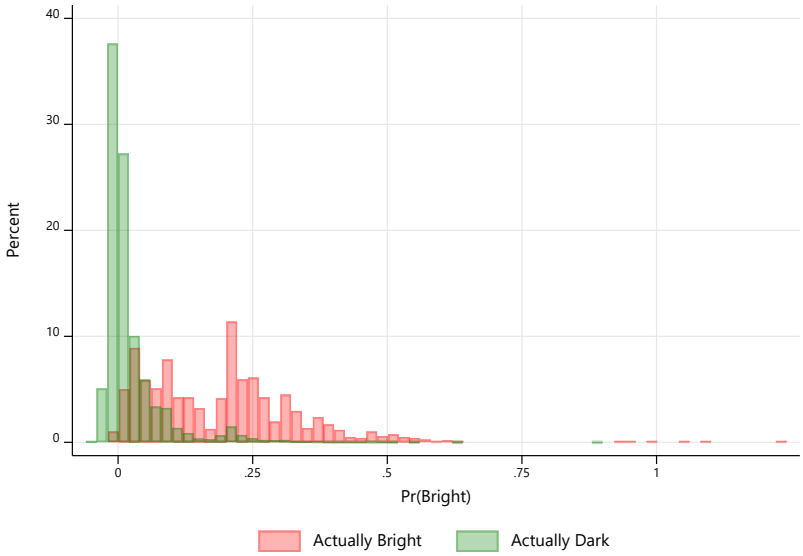
Appendix Figures & Tables

Appendix Figure 1. Predicting Brightness

A. ROC Curve



B. Distribution of Predicted Brightness



Notes: Panel A plots the ROC curve of a prediction exercise, where I predict whether a protein is bright by 1998. Panel B plots the distribution of the resulting predicted brightness by whether the protein was actually bright (i.e., had a structure by 1998) or dark (i.e., did not have a structure by 1998). The unit of analysis is a protein. The sample consists of 41,781 proteins that were discovered by 1998. Each protein's predicted brightness was constructed by using the fitted values from estimating a Lasso model that predicted whether the protein had a structure by 1998.

Appendix Table 1. Impact of MR on Number of Solved Structures, Inverse Hyperbolic Sine Transformation

	(1)	(2)
	IHS	IHS
VARIABLES	N of Structures	N of Structures
Post-MR × Bright	0.083***	0.075***
	(0.022)	(0.022)
Cluster Size		0.017**
		(0.007)
R-squared	0.464	0.464
Calendar-year FE	YES	YES
Cluster FE	YES	YES
N of clusters	6,944	6,944
N of cluster-years	145,824	145,824

Notes: This table parallels Table 2 but presents a robustness analysis using inverse hyperbolic sine transformation of the outcome. The table shows the impact of MR on the number of solved structures. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster after inverse hyperbolic sine transformation. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects; Columns 2 additionally control for time-varying (standardized) cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 2. Impact of MR on MR vs. Non-MR Structures

VARIABLES	(1) All Structures	(2) MR Structures	(3) Non-MR Structures
Post-MR × Bright	0.071***	0.143***	-0.007*
	(0.018)	(0.017)	(0.004)
R-squared	0.471	0.475	0.101
Calendar-year FE	YES	YES	YES
Cluster FE	YES	YES	YES
N of clusters	6,944	6,944	6,944
N of cluster-years	145,824	145,824	145,824

Notes: This table parallels Column 1 from Table 2. The table reports results from estimating Equation 1 and shows the impact of MR on the total number of solved structures (Column 1) and decomposes this into number of solved MR structures (Column 2) and non-MR structures (Column 3). All of the outcomes are Log(+1) transformed. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 3. Impact of MR, Splitting Bright Clusters

	(1)	(2)	(3)
	Dark vs. All Bright Clusters	Dark vs. Bright Clusters with 1 Structure	Bright Clusters with 1 Structure vs. Bright Clusters with >1 Structures
Post-MR × Bright (1 or more structure)	0.071***	0.072***	
	(0.018)	(0.022)	
Post-MR × Bright (more than 1 structure)			-0.002
			(0.034)
R-squared	0.471	0.325	0.595
Calendar-year FE	YES	YES	YES
Cluster FE	YES	YES	YES
N of clusters	6,944	6,558	649
N of cluster-years	145,824	137,718	13,629

Notes: Column 1 of this table parallels Column 1 of Table 2 and shows the impact of MR on the total number of solved structures in the full sample. Column 2 investigates the impact of MR on the sample of dark clusters and bright clusters with just 1 structure solved by 1998; the treatment variable "Bright" is defined as clusters that had just one structure by 1998. Column 3 investigates the impact of MR on the sample of bright clusters with 1 or more structures solved by 1998; the treatment variable "Bright" is defined as clusters that had more than 1 structure by 1998. All of the outcomes are Log(+1) transformed. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. All columns include calendar-year and cluster fixed effects. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 4. Impact of MR on Number of Solved Structures with Predicted Brightness

VARIABLES	(1) Log(+1) N Structures	(2) Log(+1) N Structures	(3) Levels N Structures	(4) Levels N Structures
Post-MR × Bright	0.049**	0.044**	0.218***	0.213***
	(0.021)	(0.021)	(0.055)	(0.054)
Post-MR × Predicted Bright	0.024	0.023	0.089	0.087
	(0.017)	(0.016)	(0.056)	(0.055)
Cluster Size		0.012**		0.014
		(0.006)		(0.011)
R-squared	0.472	0.472	0.401	0.401
Calendar-year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
N of clusters	6,878	6,878	6,878	6,878
N of cluster-years	144,438	144,438	144,438	144,438

Notes: This table reports results from estimating Equation 2 and shows the impact of MR on the number of solved structures, controlling for predicted brightness. The unit of analysis is a cluster × year, and the panel spans from 1999-2019. The outcome variable is the total annual number of solved structures in a cluster, reported after Log(+1) transformation (Columns 1-2) or in levels scaled by the standard deviation (Columns 3-4). The treatment variable "Bright" is defined as clusters that had at least one structure by 1998, while "Post-MR" includes years 2004 and onwards. "Predicted Bright" was constructed after predicting whether a protein was structurally characterized by 1998, using its pre-period characteristics and aggregating to the cluster-level. All columns include calendar-year and cluster fixed effects; Columns 2 and 4 additionally control for time-varying standardized cluster size. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data Appendix

A.1 UniProtKB/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProtKB) is a comprehensive database of known proteins. A protein is composed of sequence of organic compounds called amino acids. Information for making a protein is stored in a gene's DNA, and, therefore, by translating the DNA sequence of a gene, scientists can determine the protein's existence and the sequence of amino acids that will appear in the protein. Protein sequences in UniProtKB are thus sourced by translating genes from major genome sequence databases.

UniProtKB is divided into two parts: Swiss-Prot (manually reviewed) and TrEMBL (computationally reviewed). Created in 1986, the Swiss-Prot database is extensively reviewed, maintained, and annotated by experts based on experimental results and literature review. As of October 2020, Swiss-Prot contains 563,552 protein entries. In contrast, TrEMBL was created in 1996 and houses computationally annotated protein entries. Once a protein from TrEMBL becomes manually reviewed, it is removed from TrEMBL and enters Swiss-Prot. TrEMBL was established in recognition that manual curation efforts cannot keep pace with the increased number of protein sequences resulting from genome sequence projects and contains nearly two hundred million entries.

To define the complete set of proteins at risk of being structurally characterized, I follow Perdigão et al. (2015) —a bioinformatics paper that descriptively mapped out which proteins' structures have been determined—and focus on the proteins in the Swiss-Prot database. While smaller than TrEMBL, using the Swiss-Prot database has several advantages. First, Swiss-Prot is one of the best datasets of proteins whose existence is experimentally proven (Perdigão et al. 2015); TrEMBL primarily contains proteins whose existence is only predicted. Second, since Swiss-Prot tends to include more well-described proteins, this allows me to ensure that I examine proteins that share a similar baseline level of documentation and thus similarly at risk of catching the attention of structural biologists, instead of looking at unreviewed proteins that may not even be real proteins. Third, Swiss-Prot's expertly curated annotation provides rich descriptions of each protein, including its function, clinical impact, and sequence features, which allows me to develop a "predicted brightness" measure, as described in Section 6.4.

A.2 Linking Swiss-Prot to the Protein Data Bank

The PDB provides crosswalks to Swiss-Prot, allowing me to observe which proteins in Swiss-Prot have had their structures characterized in the PDB. However, the level of the crosswalk between an entry in the PDB and an entry in Swiss-Prot is not a many-to-one crosswalk as one might expect (a many-to-one, since a protein in Swiss-Prot can have its structure solved multiple times), but rather a many-to-many crosswalk (i.e., a single protein structure in the PDB can also be linked to multiple Swiss-Prot entries). This is because in the PDB, large protein structures are composed of discrete regions called "entities"; the crosswalk between the PDB and Swiss-Prot is at this entity level. Approximately 80% of the structures in the PDB are composed of a single entity, while the remaining 20% have multiple entities and therefore linked to multiple Swiss-Prot entries. Whenever a single protein structure from the PDB links to multiple Swiss-Prot entries, I split the protein structure into fractions based on the percentage of amino acids each Swiss-Prot entry contributes to the protein structure.

A.3 MMseqs2

MMseqs2 is a software package that clusters databases of proteins and can be downloaded at https://github.com/soedinglab/MMseqs2 (Steinegger and Söding 2018; Hauser, Steinegger, and Söding 2016). MMseqs2 uses a greedy set cover algorithm and aims to create the fewest number of mutually exclusive clusters, given a set of proteins at a user-specified sequence similarity. In this paper, I chose the threshold of 30% sequence similarity, given that MR will likely be successful if the template and the target proteins share at least 30% sequence identity. If the sequence similarity falls below 30%, MR will be usually challenging, if at all possible, to implement (Schmidberger et al. 2010; Phenix). The algorithm takes the following steps:

1. MMseqs2 first computes all pairwise sequence identities between proteins in Swiss-Prot
2. MMseqs2 chooses a "representative" sequence, which is the protein with the highest number of neighbors that share at least 30% sequence similarity
3. MMseqs2 forms the first cluster with this representative sequence and all of its neighbors
4. MMseqs2 then looks at the remaining sequences and chooses the next representative sequence with the highest number of neighbors
5. MMseqs2 iterates through Steps 2-4 until all sequences belong in a cluster

This ensures that each member of a cluster shares at least 30% sequence similarity with the representative sequence of the cluster.[28] MMseqs2 is used by both Swiss-Prot and the PDB to cluster similar proteins.

A.4 Sample Construction

Using the MMseqs2 algorithm, I grouped all 563,552 proteins in Swiss-Prot into 74,017 mutually exclusive clusters, using 30% sequence identity threshold.

Restricting to clusters with at least one human protein: I restricted the sample to clusters with at least one human protein (n = 13,150 clusters, which is equivalent to 161,392 proteins). There are two reasons for this restriction. First, restricting to clusters with at least one human protein ensures that all of the clusters in the final sample have a minimum baseline level of biological importance; one of the main goals of structural biology is to understand human biological processes and thus structural biologists are especially interested in proteins from humans (and their similarity neighbors). Second, focusing on human proteins mitigates the concern of growing cluster size. The total number of possible proteins in the universe is essentially infinite,[29] and new protein sequences are continuously being discovered. However, all human proteins have been discovered by the early 2000s when the human genome project was completed; since one gene encodes one protein, and humans have approximately 20,000 genes, they also have 20,000 proteins.[30] Since the number of newly discovered human proteins have plateaued since the early 2000s when MR arrived, this alleviates the concern of whether human proteins are getting more structures due to MR or because there are simply more human proteins being discovered. To

[28] A caveat is that while it is likely that all possible pairs of sequences within the cluster also share at least 30% sequence similarity with each other (since they are all similar to the representative sequence), this is not guaranteed. Mirdita et al. (2017) performed a cluster quality check that mitigates this concern; the authors computed the mean sequence identity among all possible pairs of sequences in a cluster and found that MMseqs2 indeed yielded clusters where all possible pairs of sequences shared on average >30% sequence similarity.

[29] Given that there are 20 different amino acids and an average protein has a sequence length of 200 amino acids, this amounts to 20^{200} possible proteins, which is larger than the number of electrons in the universe (Koonin, Wolf, and Karev 2002).

[30] This is called the "one gene, one protein" rule, which contributed to the 1941 Nobel Prize in Medicine. As explained in Section 4.1, by translating the DNA sequence of a gene, scientists can determine the protein's existence, and the sequence of amino acids that will appear in the final protein. Recently, the "one gene, one protein" rule has been challenged, as one gene may produce multiple proteins through, for instance, alternative splicing. Nonetheless, this paper follows the "one gene, one protein" rule since Swiss-Prot provides a non-redundant set of proteins, in that all proteins that are encoded by one gene in a species is folded into a single entry (including alternative splicing isoforms).

additionally address the concern of changing cluster size, I also control for time-varying cluster size in some of my specifications.

Restricting to clusters born on or before 1998: For each cluster, I compute its discovery year by taking the earliest discovery year among the proteins in the cluster. (Discovery year is defined as the earliest known documentation of the protein's existence.) Since my panel starts in 1999, I only keep clusters that were born on or before 1998 in my sample.[31] This led to my final sample of 6,944 clusters of proteins.

[31] Alternatively, in a robustness analysis, I restricted the sample to clusters born on or before 2003 when MR arrived. This unbalanced sample consists of 12,294 clusters. Results remained similar.