Navigating the Rugged Data Landscape: The Impact of Data-Extrapolation Technologies on Knowledge Production

Soomi Kim[†]

April 25, 2025

Abstract

Data-extrapolation technologies allow innovators to import and apply empirical patterns from one domain to another. Yet, because they depend on prior data, these tools may steer innovation towards data-rich areas. The impact of this shift depends on whether the initial data was produced in areas prioritized for importance or feasibility. I present a simple conceptual framework in which data is not randomly generated but shaped by scientists' past priorities and constraints and show how data-extrapolation tools can alter the composition of viable ideas by lowering costs only in data-rich domains. I empirically examine these dynamics in the setting of structural biology, where I investigate the introduction of a data-extrapolation technology. I exploit the variation in prior data availability to compare areas where the tool was usable to those where it was not. I find that while the technology increased the quantity of research in data-rich areas, it had limited impact on yielding new biological insights.

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

1. Introduction

Data-extrapolation technologies—tools that import and apply insights from past data—are becoming increasingly ubiquitous. These tools mine empirical patterns from one context and extend them to another, enabling progress even when the underlying mechanisms are not well understood. Data-extrapolation technologies can serve as shortcuts, allowing knowledge workers to make progress without first developing full theoretical models or causal explanations.

Yet, despite their powerful capabilities, data-extrapolation technologies face a critical limitation: they cannot function without a stock of prior, digitized data (Cockburn, Henderson, and Stern 2018). As these tools become central to innovation, the availability of data may thus shape the trajectory of knowledge production, steering the line of inquiry towards data-rich areas over data-poor ones. Consider Large Language Models, which have flourished thanks to the vast corpus of digitized text. In contrast, progress in intelligent robotics—once central to early visions of AI—has lagged, partly due to the lack of video data (Gibney 2024). AI's trajectory has been shaped as much by data availability as by societal value.

Understanding where data accumulates is thus essential. While large-scale mapping efforts aim for comprehensive coverage (Williams 2013; Nagaraj 2022; Kao 2024), most datasets are unevenly generated, shaped by past priorities and constraints (Nagaraj and Stern 2020). If data accumulates in domains of high scientific promise, data-extrapolation tools can unearth valuable discoveries. The Human Genome Project, for example, was deliberately created to map the genome, catalyzing advancements in both biology and bioinformatics. However, in some cases, tractability—not underlying importance—drives data availability. For instance, Atari 2600 games from the late 1970s have become a benchmark environment for reinforcement learning not due to the games' intrinsic value but because they provide readily available action-reward data.

I develop a simple conceptual framework to understand where data tends to accumulate and how data-extrapolation tools impact subsequent innovation. Knowledge workers choose projects based on the expected benefit relative to difficulty, a process that both generates new data and is shaped by pre-existing data. Data-extrapolation technologies lower the cost of doing research and expand the set of viable projects—but only in domains with existing data. Whether these tools unlock important discoveries that were previously too costly or merely low-value projects that were formerly not worth pursuing depends on the composition of prior data and the extent to which earlier frictions left socially valuable projects unexplored. Data-extrapolation technologies may also introduce new frictions, such as free-riding, that discourage research efforts in data-poor areas.

Empirically studying this framework poses several challenges. To assess whether a data-extrapolation technology shifts the direction of innovation, one must observe not only realized research projects but also those that could have been pursued in the absence of the technology—an often unobservable counterfactual. There must also be a credible way to measure the availability of past data, as well as the similarity between potential projects, since extrapolation technologies work by applying insights from one past project to a new but related project. Finally, this 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 nearly all biological processes. 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 target the spike proteins of SARS-CoV-2. Crucially, structural biology offers several empirical features that allow me to identify how a data-extrapolation technology may have shaped the rate and direction of innovation.

First, unlike many settings where only realized projects are observable, structural biology provides a window into the entire project landscape. Using a database of all known proteins, I observe which proteins structural biologists chose to explore and which they left unexplored. In addition, similarity between projects can be quantified in structural biology: proteins are composed of amino acid sequences, so they can be grouped based on their sequence similarity. This makes it possible to map out the 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 do not have an accumulation of prior structural data (which I term "dark" clusters).

Second, structural biology is well-suited for studying data-extrapolation 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). Rather than solving a structure from scratch, MR extrapolates from previously solved proteins with similar amino acid sequences to help determine new structures.

Finally, this data-extrapolation technology differentially treated only parts of the project landscape. Since MR relies on prior structure data, MR only works for bright clusters of proteins (i.e., clusters with previously solved structures), and does not work for dark clusters. This enables a difference-in-differences design where bright and dark clusters serve as the treatment and control groups. By linking Swiss-Prot (a database of all known proteins) to the Protein Data Bank (a database of protein structures), I examine changes in the quantity and quality of solved structures in bright clusters after the arrival of MR, relative to dark clusters.

My first set of results focuses on the volume of innovation. MR reduced the cost of solving proteins in bright clusters, and this reduction in cost—perhaps unsurprisingly—increased the total number of structures solved in those clusters. This effect was sustained throughout the entire sample period: bright clusters got brighter and brighter.

The key question is whether MR facilitated the solving of important proteins that were previously inaccessible or low-importance proteins that only became worth solving after costs fell. To answer this, one must consider how scientists choose which proteins to solve—a decision process that helps explain both why bright clusters became bright in the first place and how MR impacts subsequent research activities. Scientists select proteins by weighing expected benefit (scientific importance of the protein) against cost (technical difficulty). If scientists were maximizing $\Delta =$ Importance – Cost, then bright clusters must have been initially targeted because they contained proteins whose importance outweighed their cost. In the absence of frictions, most high-value, tractable proteins would likely have been prioritized before MR. The remaining proteins in bright clusters would then fall into two categories: (i) important but too difficult, or (ii) unimportant and not easy enough to justify solving. If the former, MR—by lowering cost—may enable access to valuable targets that had been too challenging to pursue. If the latter, MR would lead to the solving of easy but less important proteins.

My results support the latter case. After MR, bright clusters disproportionately received structures that were scientifically less meaningful: they yielded fewer functional annotations, were cited less frequently by patents, and had lower publication impact. These patterns also hold after conditioning on difficulty, which isolates variation in importance from variation in cost. Even in hard clusters, MR did not unlock structures that yielded deeper insights.

While MR may not have led to the elucidation of biologically significant proteins, the increase in research quantity can still be valuable. I observe that proteins in bright clusters were frequently mentioned in the text of downstream articles, even if not formerly cited. This suggests that these structures may not have had broad citation impact but still served as useful technical inputs to follow-on research. MR also increased the number of well-executed structures (e.g., those with better resolution), which are especially important in contexts that rely on structural precision, such as drug development or machine learning applications that require high-quality data.

In addition to assessing whether MR unlocked high- or low-importance proteins in bright clusters, it is also important to ask whether MR led to an overall expansion of research in structural biology or instead reallocated effort from under-explored dark clusters. I find that the increase in bright clusters was largely driven by scientists from non-elite institutions and those with limited prior experience in structural biology, suggesting that MR appeared to have lowered entry barriers and attracted new entrants rather than diverting effort from dark clusters. Furthermore, while MR could reduce incentives to explore dark clusters—since solving the first structure generates spillovers that lower the cost of solving similar structures and enable others to free-ride using MR—the extent of this effect appears limited. Suggestive evidence indicates that labs often reuse their own templates and rapidly make use of the structure data they generate.

A potential identification concern is that bright and dark clusters may have been evolving on different trends before MR. First, while bright clusters may have been initially targeted because they were more important or easier, such differences in levels do not threaten the difference-in-difference design, as long as pre-trends are parallel, which I verify in the event study figures. Second, I control for predicted brightness based on ex-ante traits related to biological importance and technical feasibility. The idea is to compare clusters that are similar ex-ante but happened to differ in whether they were bright or dark when MR arrived. Even after controlling for predicted brightness, the effect of actual brightness remains significant.

Taken together, these results underscore both the benefits and the limitations of data-extrapolation technologies. These tools reduce the cost of exploiting data-rich areas by borrowing insights from past data and can generate socially valuable outputs. In the case of MR, this led to a sustained increase in solved structures—many of which were technically well-executed and frequently referenced—but they yielded limited biological insights. More broadly, these findings illustrate that the impact of data-extrapolation technologies depends on the availability and composition of prior data, and may channel research toward already well-explored areas where returns are incremental. While I find limited evidence of reallocation in this setting, the fact that research choices both depend on and contribute to the data landscape suggests such technologies can reinforce existing patterns of data accumulation.

This paper contributes to several lines of research. I first build on a body of evidence that examines how technologies can both advance and constrain knowledge production (Barbosu and Teodoridis 2024). Some studies caution that standardized tools may limit exploration by anchoring researchers to established approaches (Mannucci 2017; Anthony 2021; Miric, Ozalp, and Yilmaz 2023), while others emphasize that research tools and infrastructure can broaden participation and open new directions (Furman and Stern 2011; Teodoridis 2018). Recent literature on AI, in particular, often frames it as a tool of exploration and recombinant innovation (Agrawal, Gans, and Goldfarb 2018; Agrawal, McHale, and Oettl 2018). I add to this work by showing that data-extrapolation technologies, though celebrated for their ability to uncover novel connections, is constrained by their reliance on prior templates.

AI's dependence on data underscores a broader concern studied in a second stream of literature to which I contribute: how data shapes the development and use of AI. While much of this work has focused on algorithmic bias arising from poor-quality training data (Cowgill et al. 2020; Choudhury, Starr, and Agarwal 2020), I join a smaller set of studies that emphasize how the very absence or presence of data can influence where innovations occur (Cockburn, Henderson, and Stern 2018). A particularly relevant contribution is Hoelzemann et al. (2024), who illustrate the "streetlight effect" of data through a bandit model in which data on moderately promising options can deter exploration. While their model treats the initial data as given, I offer a complementary view in which data is the product of past choices and additionally examine how data-extrapolation tools shape innovation by lowering costs in data-rich domains.

This paper also joins the nascent literature on AI as a correlation tool (Mullainathan and Rambachan 2024; Tranchero 2023a; 2023b). Most discussions of AI's limitations focus on where AI is *feasible* given data quality or availability. I instead highlight where data-extrapolation tools like AI are *useful* by noting that they are especially valuable in domains where theory is underdeveloped. Domains with well-understood causal mechanisms have less need for correlation-driven tools.

Lastly, I leverage the setting of structural biology, which was first brought to the attention of social scientists by Hill and Stein (2025b; 2025a) in their analyses of the priority reward system in science. As a field with rich scientific achievements and unusually detailed empirical data, structural biology has emerged as an attractive setting for innovation research (e.g., Zhuo 2023).

The rest of the paper proceeds as follows. Section 2 provides a conceptual framework on data generation and the impact of data-extrapolation technologies on innovation. Section 3 introduces the context of structural biology, and Section 4 describes the main data sources. Section 5 outlines the difference-and-differences design that underpins this study's empirical strategy. Section 6 presents the results, while Section 7 discusses the broader implications and concludes.

2. Conceptual Framework

Data-extrapolation technologies import and apply empirical patterns from one domain to another. They enable recombinant innovation by uncovering connections across disparate areas (Agrawal, Gans, and Goldfarb 2018; Agrawal, McHale, and Oettl 2018). But beyond recombination, these tools' strength lies in detecting correlations without requiring causal understanding (Mullainathan and Rambachan 2024; Tranchero 2023a).

I argue that data-extrapolation technologies offer a form of shortcut—facilitating progress without requiring a deep understanding of causal mechanisms—and are especially useful in domains with limited theoretical foundations. In well-theorized domains, researchers have less need for pattern-driven tools and can instead rely on tools derived from causal models; mechanical engineers, for instance, often use simulation tools grounded in physical laws. In contrast, where theory is sparse, AI's ability to substitute empirical patterns for theoretical insights become particularly valuable.

However, data-extrapolation technologies face a critical limitation: the availability of data. Because data is unevenly generated and the product of past priorities and constraints, the impact of data-extrapolation technologies depends on where data exists and whether they amplify innovations in important or merely tractable domains. To explore these dynamics, I develop a simple conceptual framework in which scientists choose projects based on expected benefits minus costs—a decision that both generates new data and is shaped by existing data. Data-extrapolation technologies influence this process by lowering the cost of research in domains where data already exists. I illustrate this framework using the setting of structural biology.

2.1 Scientists' Objective

As described in greater detail in Section 3, the goal of structural biology is to "solve" the 3D atomic structure of proteins, as visualizing a protein's structure can offer critical insights into its biological function. I assume that scientists choose which proteins to solve by maximizing the difference between expected scientific importance and the cost of solving them:

$$\Delta = \text{Importance} - \text{Cost}$$

Importance refers to whether the scientist believes a protein structure will yield an important biological insight, while the cost refers to the technical difficulty of solving the protein due to its physiochemical properties.

A protein is socially beneficial to solve if its true scientific importance exceeds its true costs:

$$\Delta_{\text{social}} = \text{Social Importance} - \text{Social Cost}$$

Social importance reflects the structure's broader contribution to downstream innovation, beyond the private benefits to the scientist. Social cost includes both the direct cost of solving the protein and the opportunity cost of forgone research the scientist could have pursued instead. In the absence of distortions or information frictions, scientists' private decisions align with the socially optimal objectives.

2.2 Impact of MR

Data-extrapolation technologies depend on the availability of existing data. One example of a data-extrapolation technology is molecular replacement (MR), a method used in structural biology to solve the structure of proteins. Instead of relying on time-consuming experimental procedures, MR uses previously solved protein structures as templates to help determine the structure of new, similar proteins. Because MR relies on similarity to existing structures, it only works in parts of the protein landscape where structural data already exists. I refer to these as "bright clusters"—groups of proteins with at least one previously solved member.[1] MR lowers the cost of solving structures in these bright clusters and can turn formerly unattractive proteins ($\Delta < 0$) into viable candidates ($\Delta > 0$).

What remains an empirical question is where in the Δ distribution MR has its greatest effect: does it unlock high-importance proteins whose cost was too prohibitive? Or does it primarily facilitate solving low-importance proteins that only became worthwhile once costs fell? While MR is socially valuable in either case as long as it enables the solving of additional proteins with $\Delta_{\text{social}} > 0$, the extent of MR's benefits depends on the type of proteins it makes possible to solve.

MR's impact therefore hinges on two factors: (i) the availability and composition of prior data—specifically, which proteins in bright clusters had already been solved and which remained unsolved at the time of MR's introduction and (ii) whether the scientists' private Δ values diverge from the socially optimal values.

[1] In Section 3, I provide a more complete explanation of MR and my empirical design of bright clusters.

2.3 Benchmark Case without Distortions

In a world without distortions, scientists' objective function implies that they prioritize proteins in descending order of Δ until all proteins with $\Delta > 0$ have been solved.

This has key implications for the composition of bright clusters. Prior to MR, bright clusters must have been targeted because they contained at least one protein with $\Delta > 0$—either a highly important protein with a manageable cost, or a low-importance protein with a very low cost. Since proteins are solved in descending order of Δ, most of those with $\Delta > 0$ were likely already targeted by the time MR arrives. The remaining proteins in bright clusters would be primarily those with $\Delta < 0$: proteins that are either important but too costly or unimportant and not easy enough to justify solving.

To understand which type of protein is more likely to remain, it is useful to consider how clusters are constructed. Because MR relies on sequence similarity to previously solved structures, I define clusters based on shared sequence similarity to identify proteins eligible for MR. Since sequence is a strong predictor of technical difficulty, and clusters are grouped based on sequence similarity, clusters are (i) relatively homogeneous in difficulty but (ii) could be heterogeneous in importance.[2] Under this framework, bright clusters can be divided into two types:

1. **High-cost bright clusters**, composed of difficult proteins with varying importance. Before MR, only the most important proteins in these clusters would have had $\Delta > 0$ and thus been solved. After MR lowers the cost of structure determination, additional proteins become viable—that is, their Δs turn positive. These include both previously inaccessible high-importance targets and less important ones that now clear the cost-benefit threshold.

2. **Low-cost bright clusters,** composed of easy proteins with varying importance. Before MR, high-importance proteins in these clusters would have been prioritized and solved, as cost was not a constraint. MR further reduces already-low costs, enabling the solving of remaining low-importance proteins that had not been worth the effort before.

This distinction yields two predictions that can be tested empirically. If bright clusters were primarily high-cost—meaning they still contained important proteins left unsolved due to technical difficulty—then the reduction of costs from MR should unlock these important proteins,

[2] This assumption is empirically verified in the data. Note that while some challenging proteins are known to play key biological roles (e.g., membrane proteins), importance and technical difficulty are not always correlated. This motivates treating the two dimensions separately both conceptually and empirically. In Section 7, I discuss the generalizability of this assumption.

leading to increases in both the quantity of structures and scientific insights. Alternatively, if bright clusters were mostly low-cost, MR will increase quantity but yield limited scientific insights, as the most valuable proteins were likely already solved.

2.4 Case with Distortions

2.4.1 Mitigating Distortions

In the benchmark world without distortions, proteins remaining at the time of MR's introduction are those with $\Delta < 0$. However, when scientists' private assessments of Δ diverge from Δ_{social}, some socially valuable proteins ($\Delta_{\text{social}} > 0$) may be left unsolved. MR helps mitigate such inefficiencies by reducing cost, making (privately) unattractive proteins newly viable.

There may be several potential sources of inefficiency. First is a bias towards novelty: in reward systems that prize being first (Merton 1957), scientists may undervalue important proteins that are too similar to already solved ones, even if these proteins are socially valuable. Second is a bias towards quantity: in reward systems that value output volume, scientists may favor easy structures, regardless of importance (e.g., grants may require "stamp collections" of structures).[3] As a result, high-importance, high-cost proteins may be undervalued because they do not align with short-term publication or funding incentives. Finally, due to imperfect information, scientists may misjudge the expected importance and cost of solving a protein. There may be "hidden gems"—important proteins with $\Delta_{\text{social}} > 0$—that were mistakenly overlooked.

This yields a third empirically testable prediction. In the benchmark world without distortions, low-cost clusters would have already exhausted their most valuable proteins by the time MR was introduced. However, in the presence of frictions, even low-cost bright clusters may have neglected high-importance proteins. By lowering cost, MR may encourage scientists to revisit socially valuable proteins that were ignored due to novelty or quantity bias and to uncover hidden gems, generating scientific insights.

2.4.2 Introducing Distortions

While MR can help correct existing inefficiencies, it may also introduce new distortions. In addition to understanding whether MR increased the solving of high- or low-importance proteins in bright clusters, it is important to consider whether MR shifted the allocation of research efforts.

[3] For example, Azoulay, Graff Zivin, and Manso (2011) show that life science grants that emphasize short-term evaluation cycles with predefined deliverables can lead to scientists favoring safer, incremental research.

Did MR lead to an overall expansion of structural biology via bright clusters? Or did it instead cause a reallocation of efforts away from clusters without prior structure data (i.e., “dark clusters”) to already-explored bright clusters?

As Hoelzemann et al. (2024) argue, when data shines light on satisfactory—but not the best—option, data can discourage workers from exploring further. Generating data is costly, and knowledge workers may be disincentivized from doing so if others capitalize on the data without bearing the cost of the data creation.

Building on this insight, I further argue that data-extrapolation technologies like MR may amplify this free-riding problem. MR generates spillovers that increases the value of solving structures in dark clusters because it reduces the cost of solving all other structures in the cluster. But paradoxically, this can create free-riding problems: if a scientist undertakes the cost of solving the first structure in a dark cluster, others can free-ride on that effort to easily solve the remaining structures in the cluster using MR. This can result in the under-exploration of dark clusters, even when they contain socially valuable targets.[4] Although harder to test empirically, this mechanism suggests a final prediction: if MR leads to a reallocation of effort rather than a net expansion, this may reflect free-riding dynamics that deter exploration of dark clusters.

2.5 Summary

MR's impact depends on (i) the availability and composition of prior data (bright clusters) and (ii) whether it corrects prior inefficiencies or introduces new distortions. While MR is socially valuable when it enables the solving of proteins with $\Delta_{\text{social}} > 0$, the extent of this value depends on which proteins are unlocked. Appendix Table 1 summarizes the four predictions derived from this framework, guided by a simple logic: MR yields meaningful insights if it unlocks overlooked high-importance proteins, and more incremental contributions if it facilitates low-importance proteins—outcomes that hinge on the composition of prior data, specifically which proteins have been solved versus remain unsolved. While I focus on structural biology, this framework can be applied broadly to other settings where uneven data availability—shaped by past scientific effort and feasibility—guides the direction of innovation. I further discuss the generalizability of this framework beyond structural biology in Section 7.

[4] Not all reallocation of scientific effort is inefficient. The concern arises when MR shifts effort toward proteins with high private Δ but low Δ_{social}. This can occur under free-riding dynamics since scientists underestimates the importance of proteins in dark clusters because they cannot internalize the full downstream spillovers of being the first to explore dark clusters.

3. Empirical Setting

An ideal empirical setting needs three ingredients: (i) an observable project landscape, where one can track which projects get explored versus unexplored, as well as a measure of similarity between potential projects, (ii) the arrival of a data-extrapolation 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 for this paper.

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 Appendix Figure 1A, researchers solved the structure of the spike proteins on the surface of SARS-CoV-2—that is, they determined the 3D coordinates of individual atoms of 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 Project Landscape

In order to investigate whether a data-extrapolation technology changes the direction of innovation, it is important to be able to observe the entire project landscape. In most settings, however, researchers can only observe projects that were realized, while alternative projects that could have been pursued (but were not) remain invisible. An attractive feature of structural biology is that it provides a unique window into the project 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 project. 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 2025a; 2025b): proteins are composed of sequences of amino acids—which are given by nature—and proteins can be grouped based on their sequence similarity.[5] Since data-extrapolation technologies work by identifying similarities from one project to another, this measure of distance enables me to track the use of such extrapolation.

3.2.2. Arrival of a Data-Extrapolation 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—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 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,[6] which proceeds in three main steps (Appendix Figure 1B). First, the protein sample must be produced in a 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

[5] I build on the works of Hill and Stein (2025a; 2025b), 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 the entire universe of proteins—both structurally characterized and uncharacterized—and cluster this universe of proteins based on their sequence similarity.

[6] 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 90% of protein structures are solved using this method during my sample period.

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 (MR). Figure 1 shows the rise in the number of structures solved by MR at the Protein Data Bank, a global repository of all solved structures.[7] 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.[8] Prior to MR, structural biologists had to resort to time-consuming experimental methods to solve the phase problem,[9] 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.

MR therefore can be viewed as a data-extrapolation method. MR operates on the empirical observation that sequence similarity is highly correlated with structural similarity, even though the causal mechanisms through which amino acid sequences determine a protein's 3D structure (a process called "protein folding") remains poorly understood. This underscores the value of data-extrapolation technologies in advancing scientific progress in domains where theoretical foundations are limited. 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.[10]

[7] 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 the most widely-used program, as it is user-friendly and considered to be the most efficient (Scapin 2013).

[8] X-ray reflections have both amplitudes and phases, but the phase cannot be inferred from the diffraction patterns. Without knowing the phase, a model of the protein structure cannot be constructed. Max Perutz and John Kendrew were awarded the 1962 Nobel Prize in Chemistry, in part for their pioneering work in overcoming this phase problem.

[9] There are two experimental methods that solve the phase problem from scratch. These methods do not rely on the availability of prior solved structures, but they can require arduous experimental efforts. See Appendix B for more details.

[10] 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.

3.2.3 Differential Arrival of a Data-Extrapolation 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 ("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 the treatment and control groups, respectively. This paves the way for a difference-in-differences design, as described in Section 5.

4. Data

To map the landscape of potential proteins that structural biologists can target, I rely on two main datasets: UniProt/Swiss-Prot, a database of all known proteins, and the Protein Data Bank, a database of all publicly available protein structures. I then cluster proteins based on their sequence similarity to construct the final sample.

4.1 UniProt Knowledgebase/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProt) 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 appear in the protein. Protein sequences in UniProt 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 UniProt.[11] 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.

[11] In addition to Swiss-Prot, UniProt has a database called TrEMBL, which is larger but contains computationally annotated proteins whose existence are largely not proven. More details are provided in Appendix. A.1.

4.2 Protein Data Bank

Established in 1971, the Protein Data Bank (PDB) is a repository of protein structures and contains approximately 170,000 structures as of October 2020. Since the early 1990s, most journals have required authors to deposit their structures in the PDB as a requirement for publication (Berman et al. 2000); the PDB therefore 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,[12] 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 MR will likely be successful if the template and the target proteins share at least 30% sequence identity (Schmidberger et al. 2010; Phenix 2022), 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 the panel begins. More details on sample construction can be found in Appendix A.

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 can only be applied 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 a difference-in-differences approach, where I estimate the following regression equation to examine the impact of MR:

$$\mathrm{Y_{ct}} = \beta_0 + \beta_1 \mathit{PostMR}_t \times \mathit{Bright}_c + \delta_\mathrm{t} + \gamma_\mathrm{c} + \varepsilon_\mathrm{ct} \qquad (1)$$

[12] MMseqs2 can be downloaded from https://github.com/soedinglab/MMseqs2. More details on MMseqs2 are provided in Appendix A.3.

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.[13] 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.

Identification relies on the parallel trends assumption: in the absence of MR, outcomes for bright and dark clusters would have followed similar trends, conditional on cluster and year fixed effects, as well as additional time-varying controls included in some specifications. In Section 6.1, I discuss this assumption in greater detail.

5.2 Descriptive Statistics

As shown in Table 1, my sample consists of 6,942 clusters, with 9% of the clusters classified as bright. The table summarizes several features related to the technical feasibility and biological importance of the proteins in the clusters at the time MR was introduced. The difficulty of solving a protein is determined by the protein's physiochemical properties, including whether the protein has membrane regions, intrinsic disorder, and compositional bias, as well as its sequence length.[14] To assess biological importance, I use several proxies: the number of publications and drugs associated with a protein, whether the protein has a known biological function or disease relevance, and whether it is of human origin.

In terms of levels, bright clusters were composed of proteins that were, on average, both easier to solve and more biologically important. For example, the average bright cluster had a

[13] 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).

[14] Membrane proteins, which are embedded in or associated with cell membranes, tend to be flexible and partially hydrophobic—features that make them especially challenging to characterize structurally. Proteins with intrinsically disordered regions, which lack a stable three-dimensional conformation, are similarly difficult to solve (Slabinski et al. 2007). Compositional bias—regions of the protein with overrepresented subsets of amino acids—is also linked to increased difficulty (Harrison 2017). Data on these features is primarily drawn from Perdigão et al. (2015), with additional updates based on my own data collection. Further details are provided in Appendix A.5.

lower proportion of disordered and membrane proteins—both known to hinder structure determination—compared to dark clusters. Bright clusters also exhibited a higher share of proteins with known biological function, as well as a greater number of associated publications and drugs.

To investigate this further, I construct composite difficulty and importance scores that range from 0 to 1 for each protein by averaging over the relevant features listed in Table 1. I then aggregated these protein-level scores to the cluster level by calculating the mean and standard deviation across proteins in each cluster. Appendix A.5 describes this process in more detail, and the resulting distributions are shown in Appendix Figure 2.

Panel A displays the distribution of mean difficulty and importance across the 6,942 clusters. As in Table 1, it shows that bright clusters tend to include proteins that are both easier and more important. Panel B shows the distribution of the within-cluster standard deviation. Clusters show relatively little variation in difficulty (average SD = 0.01) but greater variation in importance (average SD = 0.17). This pattern follows from how clusters were constructed. As described in Section 4.3, clusters were constructed using sequence similarity because MR operates across proteins with similar amino acid sequences. Since sequence determines a protein's physicochemical properties, they are correlated with technical feasibility. As a result, proteins within a cluster are likely to share difficulty but may still vary in importance.[15,16]

These descriptive patterns underscore several points. First, the fact that bright clusters are easier and more important is consistent with the framework in which scientists prioritize proteins with the highest Δ. Second, while some important proteins are also known to be challenging to solve, importance and feasibility are not always correlated,[17] which will allow me to examine how scientists respond to cost reductions across proteins that differ in importance but are similar in difficulty. Finally, while these differences in levels do not threaten the difference-in-differences strategy—so long as pre-trends are parallel—I revisit these characteristics in a robustness analysis.

[15] While sequence similarity can correlate with both structural and functional features, it more directly reflects a protein's physicochemical properties, which influence technical feasibility. In contrast, biological importance is shaped by broader functional context and can vary substantially even among proteins with similar sequences. (Hegyi and Gerstein 1999; Pearson 2013). As a result, clusters formed based on sequence similarity are likely to be more homogeneous in difficulty, but more heterogeneous in importance.

[16] Part of the higher within-cluster variation in importance may also reflect measurement limitations. While the importance features are conceptually meaningful, they likely contain more measurement error than the difficulty features, since observed importance partly reflects accumulated research attention in addition to intrinsic scientific value. This measurement noise could inflate within-cluster variance.

[17] For example, membrane proteins are known to be both important and challenging to solve. In contrast, lysozyme plays a key role in the immune system but is relatively easy to solve—factors that contributed to it being one of the earliest structures deposited in the PDB (Chayen and Saridakis 2001).

In Appendix A.6, I construct and control for a “predicted brightness” measure based on pre-period traits related to both feasibility and importance.

6. Results

6.1 Impact of MR on the Quantity of Innovation

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 the outcome after Log(+1) transformation,[18] 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. These clusters received an increase of 0.73 annual number of structures after the arrival of MR, which translates to a 29.8% increase relative to the baseline standard deviation of 2.45.

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 larger (i.e., more protein sequences are being discovered) or older. In Columns 2 and 4, I additionally control for time-varying cluster size and cluster age while estimating Equation 1.

Second, the impact of MR should be limited to structures that were actually solved using the method. If bright clusters exhibit increases in both MR-solved and non-MR-solved structures, this raises concerns that factors unrelated to MR may be causing bright clusters to get brighter. In Appendix Table 3, I confirm that MR only increased the number of MR-solved structures.

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 4 shows this 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

[18] In Appendix Table 2, I provide a robustness analysis using inverse hyperbolic sine transformation. Results remain similar.

(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, 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 2 plots the dynamic effects of MR on the number of solved structures. 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.

In the Appendix, I provide additional robustness analyses. While the event studies show no evidence of pre-trends, one might still worry that bright and dark clusters followed different trajectories for reasons unrelated to MR. To address this, in Appendix A.6, I construct a predicted brightness measure to directly control for ex-ante traits. Finally, Appendix A.7 presents alternative specifications, where I vary the cluster construction.

6.2 Impact of MR on Functional Insight

MR decreased the cost of solving structures in bright clusters, and, perhaps unsurprisingly, increased the volume of structures in those areas. The key question is whether this increase in structures led to novel biological understanding. Structural biologists do not solve structures for simply the sake of solving them; they elucidate structures with the hope of learning a new functional insight. 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 link a protein's structure to its potential function to understand the role the protein plays in various biological processes (Cassiday 2014).

To evaluate whether MR contributed to new biological insights, I examine three outcomes. First, did the structure lead to a publication? Structures deposited in the PDB without an accompanying publication to explain their significance likely contributed only modestly to advancing biological understanding. 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).[19] Second, was the structure cited by Swiss-Prot? The

[19] There could be several reasons why some structures do not have accompanying publications. A "stamp collection" of structures may be required to win grants from funding agencies. Some of these structures are

Swiss-Prot database contains extensive annotations about a protein's function and provides references behind each 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. I track whether the publication associated with the structure was included as a reference under Swiss-Prot's functional annotation section. Finally, I assess whether the structure was cited by a patent, under the assumption that it must have generated sufficient functional insight to enable downstream commercial application. 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 within five years of publication.

I find that bright clusters disproportionately received more structures that did *not* reveal functional insights after MR. As shown in Table 3, bright clusters received 8% more *unpublished* structures, which have no accompanying articles that describe their function (Column 1). In contrast, bright clusters experienced a marginal 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.

Complementing these analyses, I explored how the scientific community valued MR-enabled structures by analyzing their publication impact. Specifically, for each structure's publication, I tracked their citation impact (Table 4) and journal prestige (Appendix Table 6), [20] and decomposed the number of solved structures into terciles based on each measure. I find that bright clusters especially received more structures with *less* publication impact. These patterns suggest that while MR increased structure output, it did not lead to work that the broader community deemed especially impactful.

Ease vs. Importance

To interpret these results, I return to the conceptual framework, where scientists choose proteins that maximize the difference between importance and cost (Δ). Absent frictions, scientists would prioritize proteins in descending order of Δ until only those with negative Δs remain. This implies that bright clusters became bright because they contained at least one protein with a

also from structural genomics consortiums, whose goals are to catalogue as many types of structures as possible without necessarily explicating them (Petsko 2007).

[20] I linked each structure's primary publication in the PDB to PubMed records and obtained citation data from OpenAlex (measured as the mean annual citations received within the first five years after publication), along with the journal's impact factor.

positive Δ, and by the time MR was introduced, the remaining proteins in those clusters were either (i) important but prohibitively costly or (ii) unimportant and not low-cost enough to justify solving. If the former, MR might unlock valuable proteins by reducing cost; if the latter, it would simply enable expansion into lower-value proteins.

To assess which force dominates, I classify clusters by technical difficulty using the composite score from Section 5.2, defining "easy" clusters as those below the median and "hard" clusters as those above the median. I then re-estimate the main regression separately by difficulty group. Because clusters are formed by sequence similarity, which strongly predicts difficulty, this classification allows me to condition on technical difficulty and examine whether MR's effects vary. Specifically, the type of proteins remaining when MR arrived likely differed by cluster difficulty. In hard clusters, some important proteins may have remained unsolved due to high cost. In easy clusters, high-value targets were likely already solved since cost is not a constraint, leaving mainly low-importance proteins.

Appendix Table 5 reports the results. First, easy bright clusters were disproportionately represented: among bright clusters, 72% were easy, low-cost clusters, while 28% were difficult, high-cost clusters. This prevalence of easy bright clusters helps explain why MR did not yield substantial functional insights in the main results. Second, even when splitting clusters into hard and easy clusters, the pattern of MR leading to limited insight largely persists. In hard clusters, MR appears to have unlocked some previously inaccessible proteins that were at least publishable—but these structures offered limited downstream value in terms of functional annotations or patent citations.[21] In easy clusters, the increase was driven mainly by proteins that never led to published findings—consistent with MR enabling the solving of low-cost, low-importance proteins, after more valuable ones were already exhausted.

Taken together, these results suggest that MR lowered technical barriers to solving proteins (including challenging ones in hard clusters), but the resulting insights from these newly solved proteins were modest. Furthermore, there is limited evidence that scientists systematically overlooked valuable proteins prior to MR due to frictions such as bias towards novelty or quantity. If such distortions had shaped early solving decisions, we would expect MR—by lowering costs—to reveal important-but-overlooked proteins, leading to gains in functional insight or citation impact. Such patterns, however, do not appear in the data.

[21] The fact that hard clusters did not yield dramatically important insights suggests that MR may have lowered costs just enough to make moderate-importance proteins viable, while truly high-importance, high-cost proteins remained out of reach (Δ stayed negative).

6.3 Impact of MR on Technical Utility and Execution Quality

The conceptual framework highlights that MR can be socially valuable if it expands the set of proteins that are worth solving by shifting proteins with previously negative Δ into the viable range. The extent of the benefits depends on whether these proteins are high-importance or low-importance proteins. While MR appears to have primarily enabled the solving of low-importance proteins in bright clusters—and did not generate novel biological insights—the resulting increase in incremental structures can still be valuable. To assess this possibility, I examine two sets of measures: technical utility and execution quality.

Technical Utility

The PDB provides a unique measure called "mentions," which captures whether a structure is referenced in the text of downstream articles, even if not formally cited.[22] Mentions likely capture technical use cases, such as when structures are incorporated into methods, figures, or analyses. As shown in Table 4, MR did not significantly increase the number of highly cited structures, but it did disproportionately raise the number of structures that were frequently mentioned in downstream research. This pattern suggests that MR helped increase the supply of structures, though not broadly influential in terms of citations, were still actively referenced and used by downstream researchers as technical inputs. Similarly, Appendix Table 6 shows that MR especially increased the number of structures published in lower impact journals and specialist structural biology journals, further suggesting that these contributions were technically valuable within the structural biology community, even if they were not recognized as high-impact by broader audiences.

Execution Quality

Another measure of quality to consider is execution. The quality of a scientific contribution reflects not only the significance of the problem it addresses, but also the rigor with which it is executed. In the context of structural biology, this refers to the degree to which a structure is carefully and meticulously constructed. In particular, I take advantage of measures provided in the PDB called the R-free and resolution. These are objective metrics used by the structural

[22] These measures are available in Protein Data Bank in Europe (PDBe). The PDB consists of a consortium of global partners, such as PDBe and RCSB PDB (the US branch). While each partner maintains its own website, they share the same underlying global data.

biology community to assess the technical execution—specifically, the precision and accuracy—of the structures (Kleywegt and Jones 1997).

Resolution captures precision, or the level of structural detail. Appendix Figure 1C 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 how well-ordered the protein crystals are, which researchers can influence by optimizing crystallization conditions. Highly ordered crystals—where proteins are tightly packed and uniformly aligned—produce diffraction patterns with finer details.

The R-free refers to accuracy or goodness-of-fit: how well the model of the protein structure matches the observed experimental data. As discussed in Section 3.2, structural biologists build atomic models of their protein structure, simulate diffraction patterns based on the models, and compare the simulated diffractions to the experimentally observed patterns. The R-free improves as researchers refine their models to better align with the experimental data.

With these measures, I investigate the impact of MR on execution in Table 5. Columns 1-3 decompose the total number of structures solved in a cluster into terciles based on the structure's resolution values, while Columns 4-6 similarly report results using the R-free of the structures.[23] A clear pattern emerges: bright clusters especially received well-executed structures. For resolution, 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 9% more structures that were solved in the top tercile, relative to dark clusters. Likewise, for R-free, bright clustered received just 4% more structures from the bottom tercile, but 9% more structures from the top tercile.

One interpretation of these results is that, in the absence of strong biological insights, scientists may have emphasized technical rigor to make their work more publishable. Indeed, well-executed structures are particularly useful for drug development, which require precision, as well as for serving as high-quality training data for machine learning applications. These findings suggest that MR expanded the set of solvable structures in ways that supported downstream research, even if its impact was more incremental than transformative.

[23] For both resolution and R-free, lower values indicate higher quality. I construct terciles such that that higher terciles correspond to higher quality.

6.4 Did MR Crowd-Out the Exploration of Dark Clusters?

In addition to investigating whether MR unlocked primarily high- or low-importance among bright clusters, a central question is whether this increase in bright clusters came at the expense of exploring dark clusters. The difference-in-differences approach is limited to estimating only the *relative* increase in solved structures between bright and dark clusters. This relative increase can mean either an overall increase in structural biology research via bright clusters or a reallocation of innovative efforts from dark to bright clusters. While I cannot separate this out perfectly, I conduct several suggestive analyses.

Principal Investigators

One way to assess whether MR caused an overall increase or a reallocation is to investigate whether MR attracted new entrants. This requires identifying the principal investigator (PI) responsible for each structure. However, the PDB does not provide author identifiers and only reports the last names and the initial of first names of authors, which makes it difficult to distinguish between PIs, especially those with common last names. To address this challenge, I relied on Author-ity (Torvik and Smalheiser 2021) and MapAffil datasets (Torvik 2021), which provide disambiguated author IDs and standardized affiliations for articles indexed in PubMed, the leading database of life science publications. I used a combination of matching heuristics and extensive manual cleaning to link PIs to PDB structures, as detailed in Appendix A.8.

After obtaining the author IDs, I constructed three measures. First, the PI was classified as being from an elite institution in a given year if their affiliation appeared in the QS World University Rankings for Life Sciences and Medicine.[24] Second, as an alternative measure of pedigree, I tracked the cumulative number of citations the PI had accrued each year. Third, I assessed the PI's prior experience in structural biology each year by calculating the cumulative share of structural biology–related keywords in their publication history.[25] This distinguishes PIs who are specialists in structural biology (those whose publications are primarily related to structural biology) versus PIs whose involvement in structural biology is only peripheral.

[24] I applied the 2023 QS World University Rankings, as the ranking system was first introduced in 2004 and is not available for the earlier part of my sample period. One limitation of this measure is that it excludes elite non-university research institutions.

[25] PubMed's curators annotate every publication with standardized keywords known as MeSH (Medical Subject Headings) terms. A MeSH term was classified as structural biology–related if it fell within the top 5% of MeSH terms that appear the most frequently in PDB publications. For each year in a PI's publication history, I compiled all MeSH terms appearing across their publications and calculated the share corresponding to structural biology, providing a year-specific measure of structural biology specialization.

Table 6 reports the results. While there is no significant difference between bright and dark clusters in the number of structures solved by PIs from elite institutions, bright clusters saw a greater increase in structures contributed by non-elite PIs. Similar patterns hold when pedigree is measured using citations. Furthermore, bright clusters experienced a disproportionate increase in structures from PIs with less specialization in structural biology.

These patterns point to the possibility that MR may have had a democratizing effect. In a scientific system that rewards priority, elite PIs would have been the most likely to pursue novel dark clusters prior to MR. However, there is no evidence that their activity in bright clusters increased after MR, making a reallocation of effort from dark to bright clusters less likely. Similarly, PIs whose research agendas were not focused on structural biology had limited involvement in the field beforehand, suggesting that their post-MR activity reflects entry rather than reallocation. Overall, MR appears to have lowered the entry cost into bright clusters and attracted new participants to structural biology, rather than diverting effort away from dark clusters.

Free-Riding

Finally, as outlined by the conceptual framework, MR has the potential to introduce new distortions that could discourage research in dark clusters. MR creates positive spillovers by lowering the cost of solving additional structures within a cluster, but this very feature can discourage initial exploration. Solving the first structure in a dark cluster enables others to more easily solve related proteins using MR, potentially creating free-riding concerns. As a result, scientists may underinvest in dark clusters—even when they contain socially valuable targets—because the initial costs are high while the private returns are limited.

While I cannot measure free-riding directly, suggestive evidence indicates that such concerns may be limited. For 75% of MR-solved structures, I can identify the prior template structure used and determine whether the template originated from within the same lab or from a different one. Notably, a substantial share (45%) of structures relied on templates from within the same lab. Furthermore, as shown in Appendix Table 9, using a template from within the same lab is associated with a five-year reduction in the time required to solve the subsequent structure, compared to relying on templates from a different lab. These patterns suggest that labs may be able to internalize the benefits of solving a structure and rapidly make use of the structure data they generate. One possible explanation is that, even when structure data is made publicly available, competing labs may face difficulties in using it immediately due to limited familiarity with the template protein. Importantly, the scientific community's emphasis on priority may provide additional incentives to initiate work in dark clusters, mitigating the risk of free-riding.

7. Discussion

Using the setting of structural biology—which offers a unique window into the full project landscape—I study the introduction of a data-extrapolation tool, MR, which facilitates structure determination by leveraging data on prior structures. MR reduced the cost of solving proteins in data-rich (bright) clusters, sparking an increase in the quantity of structures. While this did not translate into a corresponding increase in functional insights, the additional structures served as technical inputs to downstream research, were often well-executed, and attracted new entrants to the field.

Broadly, these findings highlight that data-extrapolation technologies like MR operate by lowering costs in nearby areas where data is available. Whether this yields transformative or incremental impact depends on where and why the data exists—that is, which problems had already been explored and what remains tractable once extrapolation becomes possible. Much of the prior literature on data-driven tools assumes that data simply exists and focuses on how algorithms perform given those data. In contrast, I emphasize that data is not randomly distributed. The idea landscape is often conceptualized as "rugged," with peaks and valleys of opportunities (Kauffman 1993; Stuart and Podolny 1996; Levinthal 1997; Fleming and Sorenson 2004). The data used to navigate this terrain is unevenly generated, accumulating in areas that were important or historically tractable. Extrapolation technologies inherit this uneven coverage and may reinforce existing research trajectories, accelerating discovery where data is abundant, while offering limited traction in data-scare regions.

To assess whether these findings extend beyond structural biology, it is useful to consider three field-specific features that shaped MR's impact. First is the joint distribution of scientific importance and technical difficulty. In structural biology, these two attributes are imperfectly correlated: important proteins are not necessarily hard to solve, and vice versa. This implies the presence of high-importance, low-cost proteins that would have been solved before MR arrived (e.g., lysozyme), leaving fewer important targets for MR to unlock. In contrast, in settings where importance and difficulty are more strongly correlated—such that the most important projects are also the most technically challenging—many high-value targets would remain unsolved due to prohibitive costs. In such cases, a cost-reducing tool like MR could have the potential to unlock important discoveries, making its contributions more substantial.

Second, this setting exhibits limited evidence of the frictions outlined in the conceptual framework. My results do not suggest that MR led to the solving of important-but-overlooked "hidden gems." This suggests that scientists appear to have selected proteins in ways that were broadly consistent with their underlying Δ, and the institutional context did not seem to have induced distortions such as free-riding. These dynamics may not generalize to fields where informational frictions or incentive structures play a more distortive role.

Third is the scope of the extrapolation enabled by MR. Like other data-extrapolation technologies, MR facilitated progress in a domain with underdeveloped theory (i.e., the theory of protein folding remains poorly understood) by borrowing insights from existing data. However, MR enables only limited extrapolation across closely related proteins, operating within narrowly defined technical neighborhoods. Data-extrapolation tools that can span more disparate domains may yield larger cost reductions and catalyze higher-impact breakthroughs, from increasing workplace productivity (Brynjolfsson, Li, and Raymond 2025) to spurring novel materials discovery (Toner-Rodgers 2024).

While MR operates within narrower technical bounds, the framework developed in this paper can help interpret the effects of more expansive data-extrapolation technologies. A timely example from structural biology is AlphaFold. In 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, DeepMind's AlphaFold algorithm bypasses the need for experiments at all. Celebrated as one of the most significant applications of AI, the DeepMind team was awarded the Nobel Prize in 2024.

Despite its breakthrough, however, AlphaFold also serves as a reminder of the limitations of data-extrapolation technologies. DeepMind's claim that AlphaFold has produced enough structures to cover the "entire protein universe" (Walsh 2022) must be qualified with an important caveat: 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 often change in the presence of small-molecule drugs, yet the PDB contains limited data on drug-bound complexes, making it difficult for AlphaFold to model such interaction (Callaway 2022).[26]

[26] Lou and Wu (2021) have found that AI is less useful for developing drugs that are radically novel and have no known mechanisms of actions. In addition, a recent paper investigates how AlphaFold changed the organizational structure of academic labs in computational biology (Cavalli 2022).

Facing these data limitations, a consortium of pharmaceutical firms announced in 2025 that they plan to build their own version of AlphaFold, trained on their internal database of proprietary structures (Callaway 2025). This announcement highlights the free-riding dynamics discussed in this paper. The PDB, built over decades by academic scientists, made the data freely available to generate spillovers in exchange for priority. This enabled the remarkable development of AlphaFold, whose own source code has also been made public. In contrast, the pharmaceutical consortium reveals an alternative response to free-riding concerns arising from data-extrapolation tools: keeping data private altogether. As a result, one of the most critical applications of AlphaFold—drug discovery—will likely occur within private firms.

AlphaFold therefore highlights that when assessing the impact of data-extrapolation tools, it is essential to consider not only where data is generated (e.g., in high- or low-value areas), but also who owns it. This concern is particularly salient as the frontier of AI increasingly resides in industry, where firms can leverage proprietary datasets for strategic advantage (Ahmed, Wahed, and Thompson 2023). For example, for firms that produce products and services based on data, early entrants may use their control over data to outcompete rivals (Bessen et al. 2022).

While this paper focuses on how data-extrapolation technologies may shift research towards data-rich regions, AlphaFold illustrates another consideration for future work: underinvestment in theory. These tools can serve as shortcuts, allowing scientists to rely on empirical patterns over causal explanations, and are especially useful in domains with limited theoretical foundations. Overreliance on such tools, however, risks accumulating correlational knowledge without advances in causal understanding (Zittrain 2019; Tranchero 2023a). The physics of protein folding remains poorly understood, and as one scientist remarked on AlphaFold, science may be "going away from human-conceived theories . . . to more data-driven methods" (Samuel 2019).

Finally, AlphaFold underscores the importance of comprehensive, exploratory data generation—"mapping for the sake of mapping" (Nagaraj and Stern 2020). MR may have enabled structures that were initially dismissed as incremental, and initiatives whose sole goals were to solve as many structures as possible were often criticized for lacking hypothesis-driven agenda (Petsko 2007; Zhuo 2023). Yet, this serves as an example of scientists' imperfect ability to predict the importance of a project; data generated without immediate scientific value may later prove pivotal. MR's role in increasing the number of solved structures helped lay the groundwork for AlphaFold by expanding the pool of training data. As data-extrapolation tools become increasingly central to innovation, navigating the rugged data landscape requires understanding how prior data generation efforts enable or constrain future discoveries.

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.

Ahmed, Nur, Muntasir Wahed, and Neil C Thompson. 2023. "The Growing Influence of Industry in AI Research." *Science* 379 (6635): 884–86.

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, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* 42 (3): 527–54.

Barbosu, Sandra, and Florenta Teodoridis. 2024. "The Role of Technology in Research and Innovation: A Taxonomy of Implications." *Working Paper.* https://ssrn.com/abstract=4371451.

Berman, Helen M, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42.

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

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2025. "Generative AI at Work." *Quarterly Journal of Economics*.

Büssow, Konrad, Christoph Scheich, Volker Sievert, Ulrich Harttig, Jörg Schultz, Bernd Simon, Peer Bork, Hans Lehrach, and Udo Heineman. 2005. "Structural Genomics of Human Proteins - Target Selection and Generation of Public Catalog of Expression Clones." *Microbial Cell Factories* 4 (July).

Callaway, Ewen. 2022. "What's Next for AlphaFold and the AI Protein-Folding Revolution." *Nature* 604:234–38.

———. 2025. "AlphaFold Is Running out of Data — so Drug Firms Are Building Their Own Version." *Nature* 640:297–98.

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.*

Chayen, Naomi E, and Emmanuel Saridakis. 2001. "Is Lysozyme Really the Ideal Model Protein?" *Journal of Crystal Growth* 232:262–64.

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 Intelligence on Innovation." *NBER Working Paper* #24449.

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.

Cowtan, Kevin. 2003. "Phase Problem in X -ray Crystallography, and Its Solution." *Encyclopedia of Life Sciences*, 1–5.

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.

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

Foos, Nicolas, Mahmoud Rizk, and Max H. Nanao. 2022. "Single-Support Serial Isomorphous Replacement Phasing." *Acta Crystallographica Section D: Structural Biology* 78:716–24.

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.

Gibney, Elizabeth. 2024. "The AI Revolution Is Coming to Robots: How Will It Change Them?" *Nature* 630:22–24.

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.

Hegyi, Hedi, and Mark Gerstein. 1999. "The Relationship between Protein Structure and Function: A Comprehensive Survey with Application to the Yeast Genome." *Journal of Molecular Biology* 288 (1): 147–64.

Hill, Ryan, and Carolyn Stein. 2025a. "Race to the Bottom: Competition and Quality in Science." *Quarterly Journal of Economics*.

———. 2025b. "Scooped! Estimating Rewards for Priority in Science." *Journal of Political Economy* 113 (3): 793–845.

Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero. 2024. "The Streetlight Effect in Data-Driven Exploration." *NBER Working Paper #32401.*

Kao, Jennifer. 2024. "Charted Territory: Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry." *Working Paper.*

Kauffman, Stuart A. 1993. *The Origins of Order: Self-Organization and Selection in Evolution*. New York, NY: Oxford University Press.

Kleywegt, G. J., and T. A. Jones. 1997. "Model Building and Refinement Practice." *Methods in Enzymology* 277:208–30.

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.

Leonard, Gordon. 2018. "The Phase Problem - and How to Solve It." In *CCP4-DLS Workshop*.

Levinthal, Daniel A. 1997. "Adaptation on Rugged Landscapes." *Management Science* 43 (7): 934–50.

Lou, Bowen, and Lynn Wu. 2021. "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms." *MIS Quarterly* 45 (3): 1451–82.

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.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles." *Strategic Management Journal* 41 (9): 1572–94.

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22 (6): 635–59.

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. 2023. "Tradeoffs to Using Standardized Tools: Innovation Enablers or Creativity Constraints?" *Strategic Management Journal* 44 (4): 909–42.

Moore, Peter B. 2007. "Let's Call the Whole Thing Off: Some Thoughts on the Protein Structure Initiative." *Structure* 15 (11): 1350–52.

Mullainathan, Sendhil, and Ashesh Rambachan. 2024. "From Predictive Algorithms to Automatic Generation of Anomalies." *NBER Working Paper #32422.*

Nagaraj, Abhishek. 2022. "The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry." *Management Science* 68 (1): 564–82.

Nagaraj, Abhishek, and Scott Stern. 2020. "The Economics of Maps." *Journal of Economic Perspectives* 34 (1): 196–221.

Nature Structural Biology. 1994. “The Changing Structure of Biology.” *Nature Structural Biology* 1 (1).

Oksanen, Esko, and Adrian Goldman. 2010. “Introduction to Macromolecular X-Ray Crystallography.” In *Comprehensive Natural Products II: Chemistry and Biology*, edited by Hung-Wen (Ben) Liu and Tadhg P. Begley, 7:312–43. Elsevier.

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.

Pearson, William R. 2013. “An Introduction to Sequence Similarity (‘Homology’) Searching.” *Current Protocols in Bioinformatics*, no. SUPPL.42.

Perdigão, Nelson, Julian Heinrich, Christian Stolte, Kenneth S. Sabir, Michael J. Buckley, Bruce Tabor, Beth Signal, et al. 2015a. “Unexpected Features of the Dark Proteome.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (52): 15898–903.

———. 2015b. “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. 2022. “Overview of Molecular Replacement in Phenix.” 2022. https://phenix-online.org/documentation/reference/mr_overview.html.

Ramakrishnan, Venki. 2018. *Gene Machine: The Race to Decipher the Secrets of the Ribosome*. New York, NY: Basic Books.

Read, Randy J. 2005. “The Phase Problem: Introduction to Phasing Methods.” Protein Crystallography Course. 2005. https://www-structmed.cimr.cam.ac.uk/Course/Basic_phasing/Phasing.html.

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.

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.

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* 16 (11): 2472–82.

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

Stuart, Toby E, and Joel M Podolny. 1996. “Local Search and the Evolution of Technological Capabilities.” *Strategic Management Journal* 17 (S1): 21–38.

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

Terwilliger, Thomas C., Gábor Bunkóczi, Li Wei Hung, Peter H. Zwart, Janet L. Smith, David L. Akey, and Paul D. Adams. 2016. “Can I Solve My Structure by SAD Phasing? Anomalous Signal in SAD Phasing.” *Acta Crystallographica. Section D, Structural Biology* 72 (3): 346–58.

Toner-Rodgers, Aidan. 2024. “Artificial Intelligence, Scientific Discovery, and Product Innovation.” *Working Paper*. http://arxiv.org/abs/2412.17866.

Torvik, Vetle. 2021. “MapAffil 2018 Dataset -- PubMed Author Affiliations Mapped to Cities and Their Geocodes Worldwide with Extracted Disciplines, Inferred GRIDs, and Assigned ORCIDs.” https://databank.illinois.edu/datasets/IDB-2556310.

Torvik, Vetle, and Neil Smalheiser. 2021. “Author-Ity 2018 - PubMed Author Name Disambiguated Dataset.” https://databank.illinois.edu/datasets/IDB-2273402.

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

———. 2023b. “Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation.” *Working Paper*.

Walsh, Bryan. 2022. “Finally, an Answer to the Question: AI — What Is It Good For?” *Vox*, August 3, 2022.

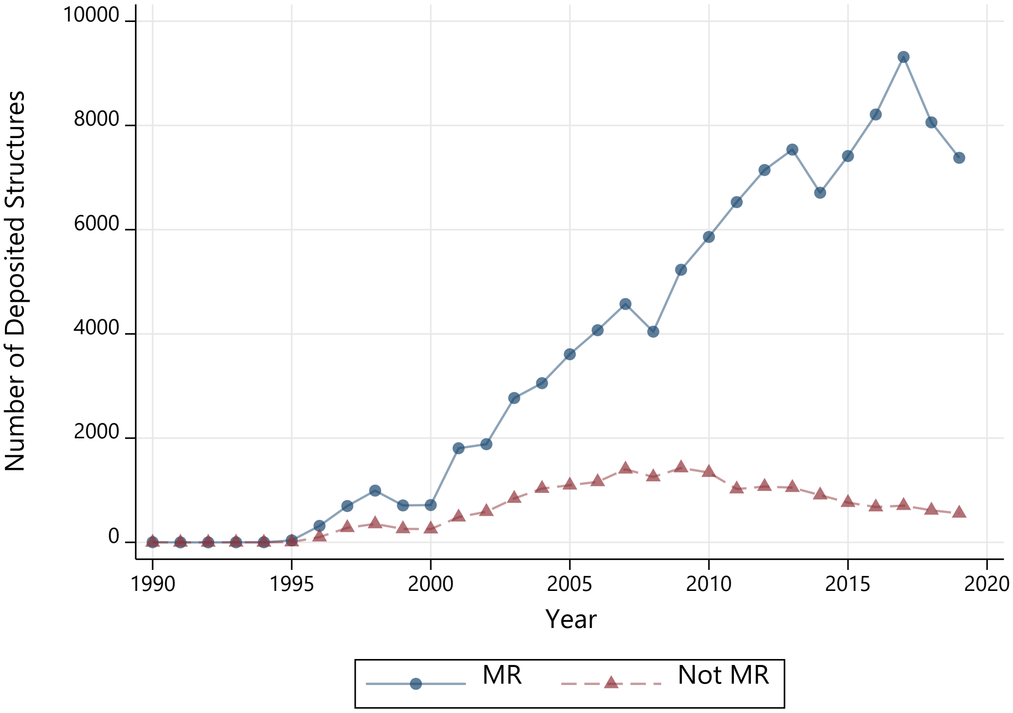
Williams, Heidi L. 2013. “Intellectual Property Rights and Innovation: Evidence from the Human Genome.” *Journal of Political Economy* 121 (1): 1–27.

Zhuo, Ran. 2023. “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.

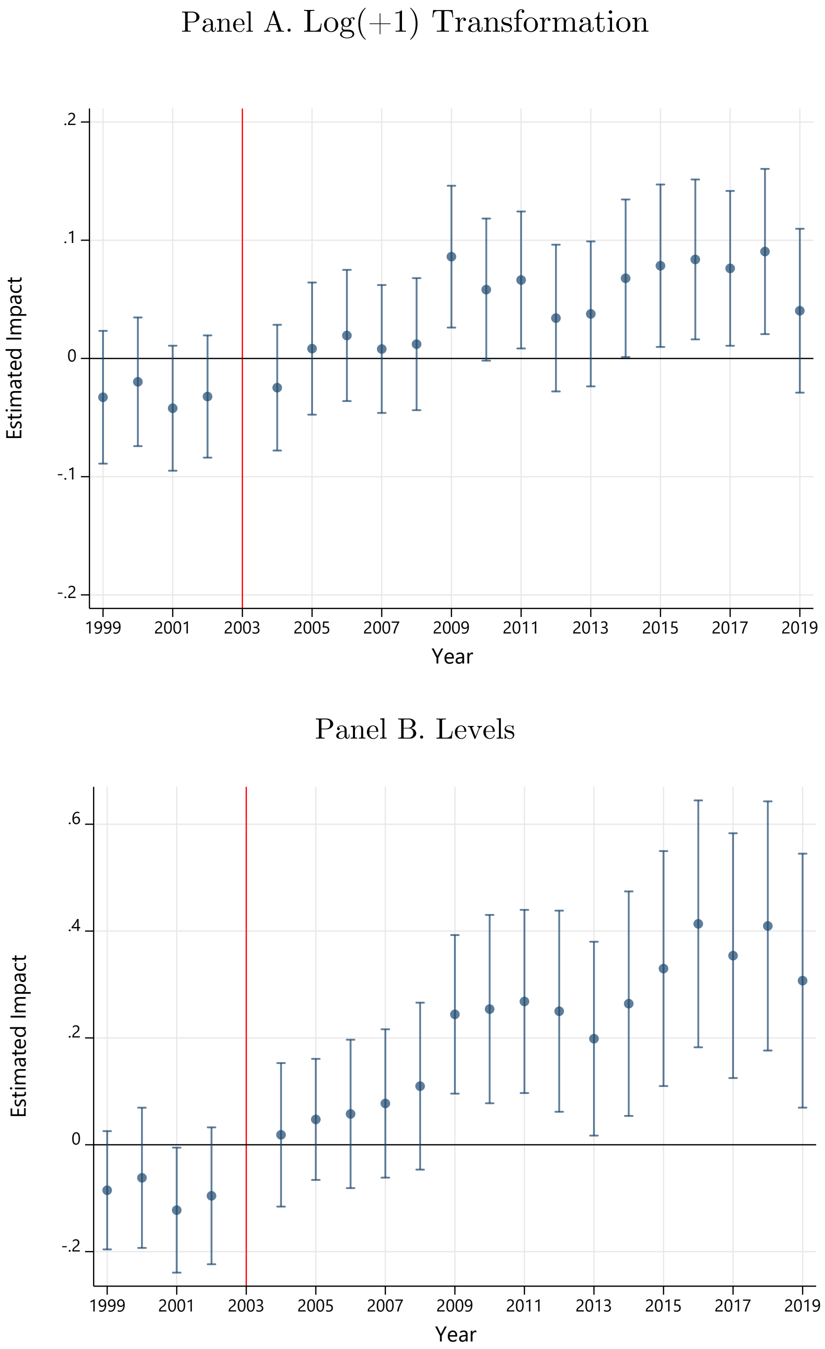
Figures & Tables

Figure 1. 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 2. 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 an 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 6,942 clusters, which translates to 145,782 cluster-years.

Table 1. Summary Statistics

	Bright Clusters		Dark Clusters	
	Mean	SD	Mean	SD
Cluster Discovery Year	1983.5	7.7	1993.1	5.1
Cluster Size	39.1	80.5	7.6	14.7
Features Related to Structure Solving Difficulty				
Share of Disordered Proteins	20%	36%	38%	46%
Share of Membrane Proteins	2%	14%	7%	25%
Share of Proteins with Compositional Bias	1%	8%	4%	18%
Sequence Length	512.0	1,434.9	634.8	676.0
Features Related to Biological Importance				
Share of Human Proteins	18%	15%	36%	24%
Share of Proteins with Known Function	67%	32%	42%	40%
Share of Proteins with Disease Relevance	5%	12%	4%	13%
N of Publications per Protein	6.9	10.2	4.2	8.6
N of Approved Drugs per Protein	0.2	0.7	0.1	0.5
N of Structures Solved per Year	1.3	3.7	0.1	0.5
N of Clusters	653		6,289	

Notes: This table provides the summary characteristics of clusters when MR was introduced in 2003. The sample consists of 6,942 clusters, of which 653 are classified as "bright" (i.e., had at least one structure by 1998) and 6,289 are classified as "dark." Cluster discovery year is the earliest recorded existence of a protein in the cluster, based on either its initial entry in major sequence databases (such as UniProt, the European Molecular Biology Laboratory database, and the DNA Data Bank of Japan) or its earliest mention in a publication, as documented by UniProt. Proteins are categorized as disordered, membrane, or compositionally biased when at least a quarter of their amino acid sequence is associated with these properties. Information on publications, functional annotation, and disease relevance as of 2003 was parsed from UniProt, while drug-related information was obtained from DrugBank.

Table 2. Impact of MR on Number of Solved Structures

VARIABLES	(1) Log(+1) N of Structures	(2) Log(+1) N of Structures	(3) Levels N of Structures	(4) Levels N of Structures
Post-MR × Bright	0.072***	0.042**	0.298***	0.223***
	(0.018)	(0.018)	(0.051)	(0.042)
R-squared	0.471	0.474	0.390	0.394
Calendar year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
Cluster size FE	NO	YES	NO	YES
Cluster age FE	NO	YES	NO	YES
N of structures (mean)	0.400	0.400	0.400	0.400
N of structures (SD)	2.450	2.450	2.450	2.450
N of clusters	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782

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). Means and standard deviations of the outcomes are reported in levels. 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 cluster size and cluster age. 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 FUNCTIONAL INSIGHTS

VARIABLES	(1) Unpublished Structures	(2) Published 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	0.082***	0.041**	0.069***	-0.018*	0.041**	0.005
	(0.008)	(0.017)	(0.017)	(0.009)	(0.017)	(0.004)
R-squared	0.213	0.465	0.416	0.318	0.473	0.123
N of structures (mean)	0.0500	0.340	0.260	0.080	0.310	0.030
N of structures (SD)	1.010	2.090	1.760	0.810	2.040	0.340
N of clusters	6,942	6,942	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782	145,782	145,782

NOTES: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures by whether the structure elucidated functional insights. 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. Means and standard deviations of the outcomes are reported in levels. Columns 1 and 2 parallel the outcome in Column 1 of Table 2 but decompose the number of solved structures by whether they were cited in a scientific article. Columns 3 and 4 decompose the number of solved structures by whether they were cited by a patent within 5 years of publication. Columns 5 and 6 decompose the number of solved structures by whether they were cited by the functional summary section of Swiss-Prot as of January 2024. 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 Citations and Mentions

	(1)	(2)	(3)	(4)	(5)
	Published Structures *Citations*			Published Structures *Mentions*	
VARIABLES	Bottom Tercile	Middle Tercile	Top Tercile	Below Median	Above Median
Post-MR × Bright	0.046***	0.033***	0.016	0.025*	0.059***
	(0.012)	(0.012)	(0.010)	(0.015)	(0.012)
R-squared	0.363	0.309	0.332	0.427	0.385
N of structures (mean)	0.110	0.110	0.120	0.190	0.150
N of structures (SD)	0.900	0.870	1.190	1.640	0.800
N of clusters	6,942	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782	145,782

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 citations and mentions. A structure's citation impact is measured as the mean annual number of citations within 5 years of publication. In addition to being formerly cited, a structure can be "mentioned" in the text of downstream articles; a structure's mention impact is measured as the mean annual number of mentions within 5 years of publication. The unit of analysis is a cluster × year. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Means and standard deviations of the outcomes are reported in levels. Columns 1-3 parallel Column 1 in Table 2 but decompose the outcome into the number of solved structures by terciles of citation impact. Columns 4 and 5 decompose the number of solved structures by whether they fall below or above the median level of mentions. 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 Execution

VARIABLES	(1) *Resolution* Bottom Tercile	(2) *Resolution* Middle Tercile	(3) *Resolution* Top Tercile	(4) *R-Free* Bottom Tercile	(5) *R-Free* Middle Tercile	(6) *R-Free* Top Tercile
Post-MR × Bright	0.009	0.072***	0.092***	0.043***	0.067***	0.089***
	(0.011)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
R-squared	0.366	0.404	0.423	0.376	0.394	0.420
N of structures (mean)	0.130	0.130	0.110	0.120	0.120	0.110
N of structures (SD)	0.700	0.940	1.460	0.800	1.060	1.060
N of clusters	6,942	6,942	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782	145,782	145,782

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 resolution and R-free values). 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. Means and standard deviations of the outcomes are reported in levels. Columns 1-3 parallel Column 1 in Table 2 but decompose the outcome into the number of solved structures by terciles based on their resolution. Columns 4-6 similarly decompose the number of solved structures by terciles based on their R-free values. For both resolution and R-free, lower values indicate higher quality; I construct terciles such that that higher terciles correspond to higher quality. 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: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

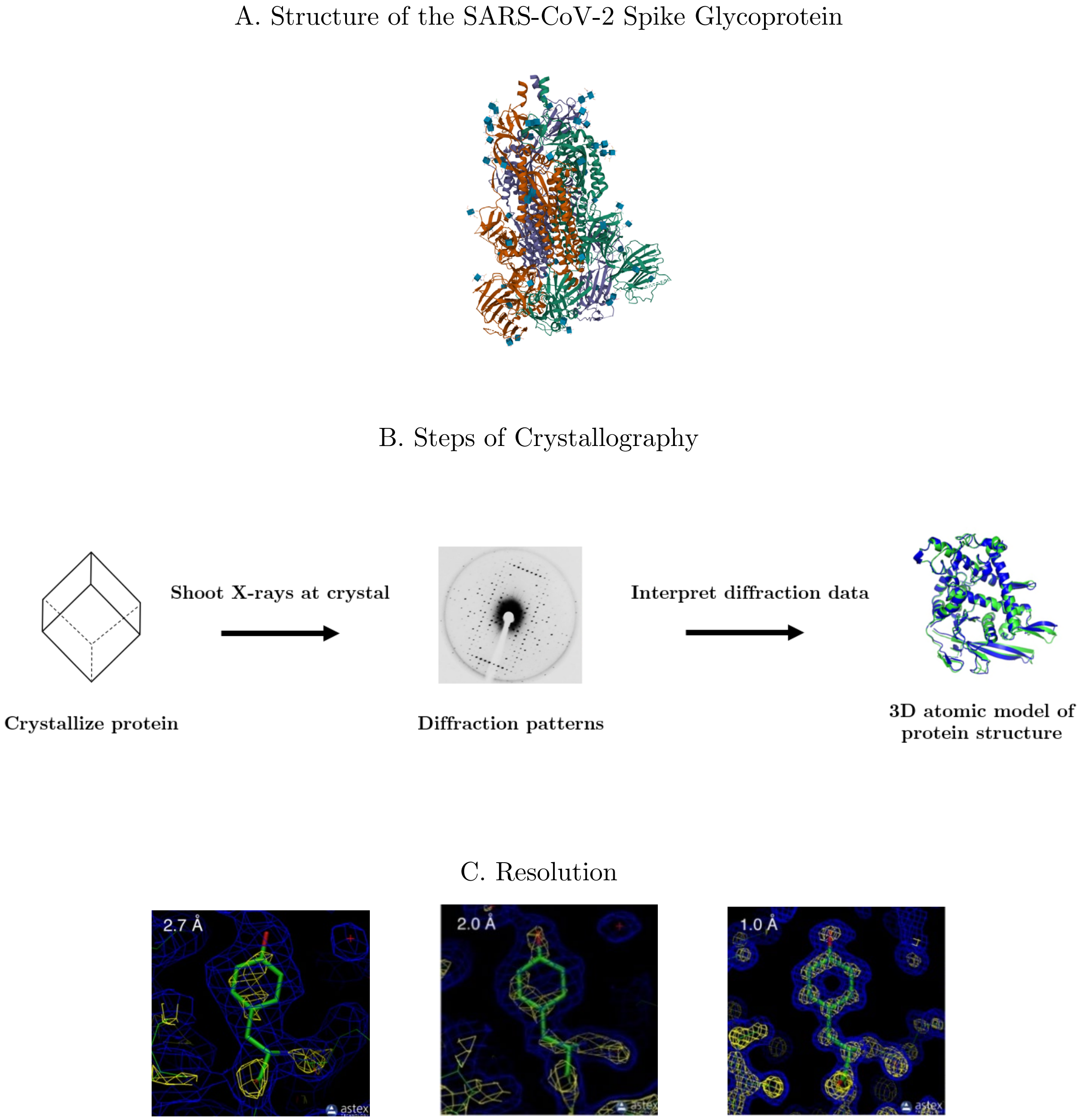
Table 6. Impact of MR By the Characteristics of the Principal Investigator

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pedigree					Prior Structural Biology Experience		
	Affiliation		Prior N of Citations			Share of Structural Biology Keywords in Prior Papers		
VARIABLES	Not Top 20	Top 20	Bottom Tercile	Middle Tercile	Top Tercile	Bottom Tercile	Middle Tercile	Top Tercile
Post-MR × Bright	0.073***	0.006	0.050***	0.053***	0.026**	0.058***	0.047***	0.018*
	(0.016)	(0.006)	(0.013)	(0.012)	(0.012)	(0.015)	(0.012)	(0.010)
R-squared	0.448	0.226	0.392	0.322	0.317	0.445	0.313	0.267
N of structures (mean)	0.250	0.050	0.120	0.120	0.110	0.140	0.120	0.090
N of structures (SD)	1.740	1.000	0.970	1.500	0.840	1.140	1.240	1.060
N of clusters	6,942	6,942	6,942	6,942	6,942	6,942	6,942	6,942
N of cluster-years	124,956	124,956	124,956	124,956	124,956	124,956	124,956	124,956

Notes: This table reports results from estimating Equation 1 and shows the impact of MR on the number of solved structures, decomposed by the characteristics of the Principal Investigator (PI). The unit of analysis is a cluster × year, and the panel spans from 1999-2016, ending in 2016 due to data availability. The outcomes of all columns are the annual number of solved structures in a cluster, with Log(+1) transformation. Means and standard deviations of the outcomes are reported in levels. Columns 1–2 parallel Column 1 of Table 2 but distinguish between number of structures solved by PIs affiliated with institutions ranked in the top 20 of the QS World University Rankings and those at institutions outside the top 20. Columns 3–5 decompose the number of solved structures by terciles of the PI's prior total citation count. Columns 6–8 decompose the number of solved structures by the PI's prior structural biology experience, using terciles based on the share of structural biology-related MeSH keywords in the PI's publication history in a given year. 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. Structural Biology



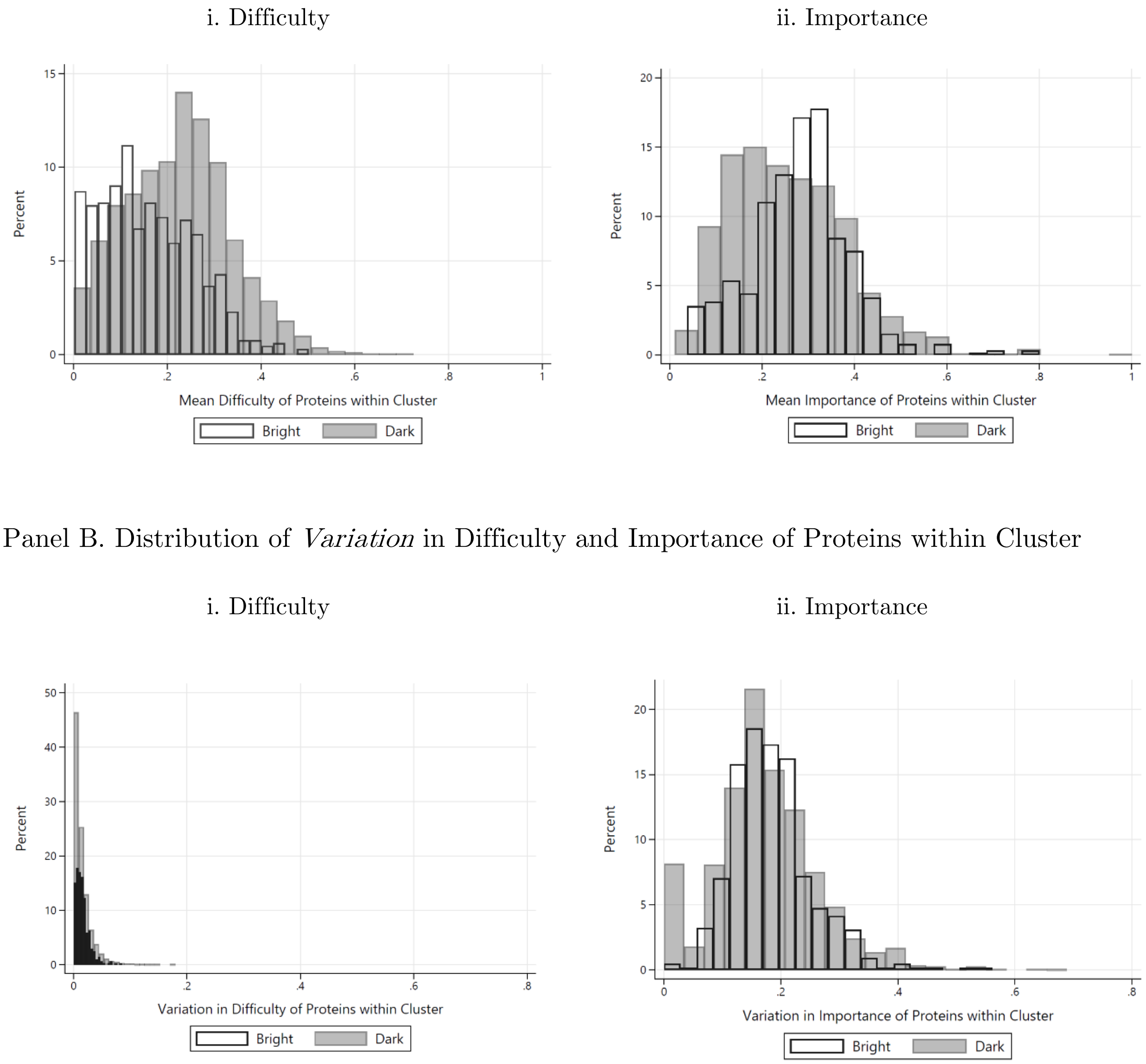
A. Structure of the SARS-CoV-2 Spike Glycoprotein

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).

Appendix Figure 2. Cluster-Level Distributions of Difficulty and Importance



Panel A. Distribution of *Mean* Difficulty and Importance of Proteins within Cluster

i. Difficulty

ii. Importance

Panel B. Distribution of *Variation* in Difficulty and Importance of Proteins within Cluster

i. Difficulty

ii. Importance

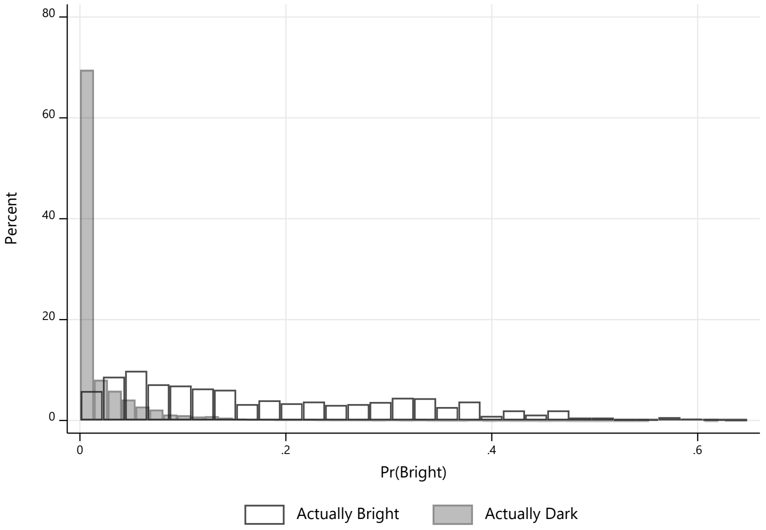
Notes: These figures plot the distribution of cluster-level difficulty and importance scores among 6,942 clusters as of 2003 when MR arrived. Panel A plots the mean scores of proteins within each cluster, while Panel B plots the variation in scores (measured by the standard deviation) of proteins within each cluster. The difficulty score of each protein is calculated as the average of four features: (i) the share of amino acids classified as membrane-associated, (ii) the share of amino acids classified as intrinsically disordered, (iii) the share of amino acids exhibiting compositional bias, and (iv) the percentile (scaled 0–1) of sequence length. The importance score of each protein is calculated as the average of five features: (i) whether the protein is associated with a disease, (ii) whether the protein has a known function, (iii) whether the protein is targeted by an approved drug, (iv) whether the protein is from human, and (v) the percentile (scaled 0–1) of number of publications written about the protein. Protein-level scores are then aggregated to the cluster level by taking the mean and standard deviation across proteins in each cluster.

Appendix Figure 3. Performance and Distribution of Predicted Brightness

A. Performance of Lasso Logit

Performance Metric	Out-of-Sample (Cross-Validated)	In-Sample (Fitted Model)
ROC AUC	0.91	0.93
PR AUC	0.24	0.33
Log Loss	0.09	0.08
N of Features Entered	143	
N of Features Selected	28	

B. Distribution of Predicted Brightness



Notes: This figure provides details on the performance and distribution of the predicted brightness measure. The predicted brightness measure was estimated using a Lasso Logit model, which predicts whether a protein was structurally characterized by 1998 based on pre-period biological and technical feasibility features (see Appendix A.6 for details). Panel A reports model performance based on both out-of-sample (cross-validated) and in-sample (fitted) predictions, including Receiver Operating Characteristic Area Under Curve (ROC AUC), Precision-Recall Area Under Curve (PR AUC), and log loss. The Lasso Logit model selected 28 features out of 143 candidates. 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 sample consists of 42,547 proteins that were discovered as of 1998.

Appendix Table 1. Summary of Conceptual Framework

Case	Remaining Proteins When MR is Introduced	MR's impact	Empirically Testable Predictions
Benchmark (no distortions or inefficiencies)	Only proteins with $\Delta < 0$ In high-cost clusters: this includes important proteins that are too difficult to solve In low-cost clusters: only low-importance proteins left	MR expands the set of feasible proteins: MR lowers cost and turns previously unattractive proteins $\Delta < 0$ into viable candidates ($\Delta > 0$). This includes both high-importance and low-importance proteins that now clear the threshold	In high-cost clusters: biological insights generated, as high-importance proteins are unlocked by MR In low-cost clusters: limited insights, as only low-importance proteins are remaining
Pre-MR distortions (e.g., novelty bias, quantity bias, imperfect information)	Both proteins with $\Delta < 0$ and some overlooked proteins with $\Delta_{\text{social}} > 0$	MR can correct past misallocation by unlocking important-but-neglected proteins	Increase in biological insights, even in low-cost clusters
Post-MR distortions (free-riding)	--	Free-riding discourages the exploration of dark clusters, even if they contain socially valuable targets	Reallocation of effort from dark to bright clusters, rather than an overall expansion

Notes: This table summarizes the four empirical predictions derived from the conceptual framework in Section 2. The impact of MR depends on (i) the availability and composition of prior data (bright clusters) and (ii) whether MR corrects prior inefficiencies or introduces new distortions. Scientists choose which proteins to solve by maximizing the difference between expected scientific importance and the cost of solving them: Δ = Importance – Cost. A protein is socially beneficial to solve if its true scientific importance exceeds its true costs: Δ_{social} = Social Importance − Social Cost. Social importance reflects the structure's broader contribution to downstream innovation, beyond the private benefits to the scientist. Social cost includes both the direct cost of solving the protein and the opportunity cost of forgone research the scientist could have pursued instead.

Appendix Table 2. Impact of MR on Number of Solved Structures, Inverse Hyperbolic Sine Transformation

VARIABLES	(1) IHS N of Structures	(2) IHS N of Structures
Post-MR × Bright	0.084*** (0.022)	0.047** (0.022)
R-squared	0.464	0.468
Calendar year FE	YES	YES
Cluster FE	YES	YES
Cluster size FE	NO	YES
Cluster age FE	NO	YES
N of structures (mean)	0.400	0.400
N of structures (SD)	2.450	2.450
N of clusters	6,942	6,942
N of cluster-years	145,782	145,782

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 inverse hyperbolic sine transformation. Means and standard deviations of the outcomes are reported in levels. 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 cluster size and cluster age. 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 on MR vs. Non-MR Structures

VARIABLES	(1) All Structures	(2) MR Structures	(3) Non-MR Structures
Post-MR × Bright	0.072***	0.143***	-0.007*
	(0.018)	(0.017)	(0.004)
R-squared	0.471	0.475	0.102
N of structures (mean)	0.400	0.290	0.0200
N of structures (SD)	2.450	2.170	0.190
N of clusters	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782

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 MR structures (Column 2) and non-MR structures (Column 3). All of the outcomes are Log(+1) transformed. Means and standard deviations of the outcomes are reported in levels. 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 4. Impact of MR, Splitting Bright Clusters

	(1) Dark vs. All Bright Clusters	(2) Dark vs. Bright Clusters with 1 Structure	(3) Bright Clusters with 1 Structure vs. Bright Clusters with >1 Structures
Post-MR × Bright (1 or more structure)	0.072***	0.068***	
	(0.018)	(0.022)	
Post-MR × Bright (more than 1 structure)			0.006
			(0.034)
R-squared	0.471	0.324	0.592
N of structures (mean)	0.400	0.250	1.980
N of structures (SD)	2.450	1.830	5.540
N of clusters	6,942	6,549	653
N of cluster-years	145,782	13,7529	13,713

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. Means and standard deviations of the outcomes are reported in levels. 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 5. Impact of MR on Functional Insights: Hard vs. Easy Clusters

	Hard Clusters						Easy Clusters					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES	Unpublished Structures	Published Structures	Published Structures *Not Cited by Patent*	Published Structures *Cited by Patent*	Published Structures *Not Fxn Annotated*	Published Structures *Fxn Annotated*	Unpublished Structures	Published Structures	Published Structures *Not Cited by Patent*	Published Structures *Cited by Patent*	Published Structures *Not Fxn Annotated*	Published Structures *Fxn Annotated*
Post-MR × Bright	0.057***	0.093***	0.117***	0.005	0.090***	0.006	0.092***	0.026	0.053***	-0.024**	0.027	0.004
	(0.014)	(0.032)	(0.030)	(0.019)	(0.032)	(0.006)	(0.010)	(0.02)	(0.020)	(0.011)	(0.020)	(0.006)
R-squared	0.158	0.401	0.328	0.318	0.406	0.112	0.246	0.504	0.464	0.318	0.513	0.130
N of structures (mean)	0.04	0.260	0.190	0.0700	0.240	0.0200	0.0700	0.42	0.330	0.0900	0.390	0.0400
N of structures (SD)	1.24	1.960	1.630	0.800	1.920	0.290	0.700	2.21	1.870	0.810	2.140	0.370
N of clusters	3,471	3,471	3,471	3,471	3,471	3,471	3,471	3,471	3,471	3,471	3,471	3,471
N of bright clusters	185	185	185	185	185	185	468	468	468	468	468	468
N of dark clusters	3,286	3,286	3,286	3,286	3,286	3,286	3,003	3,003	3,003	3,003	3,003	3,003
N of cluster-years	72,891	72,891	72,891	72,891	72,891	72,891	72,891	72,891	72,891	72,891	72,891	72,891

Notes: This table parallels Table 3 but splits the sample by cluster-level difficulty. The difficulty score of each protein is calculated as the average of four features: (i) the share of amino acids classified as membrane-associated, (ii) the share of amino acids classified as intrinsically disordered, (iii) the share of amino acids exhibiting compositional bias, and (iv) the percentile (scaled 0–1) of sequence length. Protein-level scores are then averaged to the cluster level. Clusters with above-median scores are classified as "hard," while those below the median are classified as "easy." See Appendix A.5 for more details.

Appendix Table 6. Impact of MR on Journal Impact and Specialization

	(1)	(2)	(3)	(4)	(5)	(6)
	Journal Impact Factor			Journal Specialization in Structural Biology		
VARIABLES	Bottom Tercile	Middle Tercile	Top Tercile	Low Specialization Tercile	Moderate Specialization Tercile	High Specialization Tercile
Post-MR × Bright	0.056***	0.021*	0.014	0.016	0.004	0.058***
	(0.012)	(0.011)	(0.010)	(0.011)	(0.012)	(0.010)
R-squared	0.406	0.302	0.287	0.345	0.326	0.269
N of structures (mean)	0.130	0.110	0.110	0.130	0.110	0.0900
N of structures (SD)	1.070	0.890	1.080	1.230	0.930	0.720
N of clusters	6,942	6,942	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782	145,782	145,782

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 journal impact and specialization. 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. Means and standard deviations of the outcomes are reported in levels. Columns 1-3 decompose the number of solved structures by terciles of journal impact factor. Columns 4-6 decompose the number of solved structures in journals that are high or low in structural biology specialization. Journal specialization is measured as the average share of structural biology–related keywords per article, calculated at the journal level over the sample period; journals are then classified into terciles based on this measure. 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 7. Impact of MR on Number of Solved Structures with Predicted Brightness

A. Inclusion of Post-MR × Predicted Bright

VARIABLES	(1) Log(+1) N of Structures	(2) Log(+1) N of Structures	(3) Levels N of Structures	(4) Levels N of Structures
Post-MR × Bright	0.065***	0.046***	0.276***	0.224***
	(0.019)	(0.018)	(0.052)	(0.043)
Post-MR × Predicted Bright	0.168	-0.173	0.542**	-0.040
	(0.127)	(0.132)	(0.251)	(0.271)
R-squared	0.471	0.474	0.390	0.394
Calendar year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
Cluster size FE	NO	YES	NO	YES
Cluster age FE	NO	YES	NO	YES
N of structures (mean)	0.400	0.400	0.400	0.400
N of structures (SD)	2.450	2.450	2.450	2.450
N of clusters	6,942	6,942	6,942	6,942
N of cluster-years	145,782	145,782	145,782	145,782

B. Use of Propensity Score Matching

VARIABLES	(1) Log(+1) N of Structures	(2) Log(+1) N of Structures	(3) Levels N of Structures	(4) Levels N of Structures
Post-MR X Bright	0.055***	0.047**	0.262***	0.199***
	(0.021)	(0.021)	(0.053)	(0.042)
R-squared	0.568	0.574	0.536	0.548
Calendar year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
Cluster size FE	NO	YES	NO	YES
Cluster age FE	NO	YES	NO	YES
N of structures (mean)	1.190	1.190	1.190	1.190
N of structures (SD)	4.220	4.220	4.220	4.220
N of clusters	1,306	1,306	1,306	1,306
N of cluster-years	27,426	27,426	27,426	27,426

Notes: These tables parallel Table 2 and show the impact of MR on the number of solved structures, controlling for predicted brightness. The predicted brightness measure was estimated using a Lasso Logit model, predicting whether a protein was structurally characterized by 1998 based on pre-period features on biological importance and technical feasibility. These scores were then averaged to the cluster level (see Appendix A.6 for details). In Panel A, the "Post-MR × Predicted Bright" term was added to Equation 1. In Panel B, one-to-one nearest-neighbor propensity score matching was implemented on the predicted brightness measure; unmatched dark clusters are dropped from the sample. 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). Means and standard deviations of the outcomes are reported in levels. 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 cluster size and cluster age. Standard errors are clustered at the cluster level. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix Table 8. Impact of MR on Number of Solved Structures: Alternative Sample Construction

A. Unbalanced Panel

VARIABLES	(1) Log(+1) N of Structures	(2) Log(+1) N of Structures	(3) Levels N of Structures	(4) Levels N of Structures
Post-MR × Bright	0.088*** (0.018)	0.046*** (0.017)	0.396*** (0.063)	0.279*** (0.051)
R-squared	0.464	0.468	0.379	0.383
Calendar year FE	YES	YES	YES	YES
Cluster FE	YES	YES	YES	YES
Cluster size FE	NO	YES	NO	YES
Cluster age FE	NO	YES	NO	YES
N of structures (mean)	0.250	0.250	0.250	0.250
N of structures (SD)	1.970	1.970	1.970	1.970
N of clusters	12,292	12,292	12,292	12,292
N of cluster-years	248,677	248,677	248,677	248,677

B. UniProt Similarity Families

VARIABLES	(1) Log(+1) N of Structures	(2) Log(+1) N of Structures	(3) Levels N of Structures	(4) Levels N of Structures
Post-MR × Bright	0.131*** (0.020)	0.069*** (0.020)	0.392*** (0.055)	0.282*** (0.046)
R-squared	0.542	0.545	0.453	0.457
Calendar year FE	YES	YES	YES	YES
Family FE	YES	YES	YES	YES
Family size FE	NO	YES	NO	YES
Family age FE	NO	YES	NO	YES
N of structures (mean)	0.550	0.550	0.550	0.550
N of structures (SD)	3.360	3.360	3.360	3.360
N of families	6,066	6,066	6,066	6,066
N of family-years	127,386	127,386	127,386	127,386

Notes: These tables parallel Table 2 and show the impact of MR on the number of solved structures under alternative sample constructions. Panel A presents results using an unbalanced panel that includes clusters discovered by 2003. Panel B reports results based on an alternative cluster definition using UniProt similarity families. See Appendix A.7 for more details.

Appendix Table 9. Time Lag Between Template and Downstream Structures

VARIABLES	(1) Time Lag in Structure Solving between Template and Downstream Structures	(2) Time Lag in Structure Solving between Template and Downstream Structures
Template from Own Lab	-5.044***	-4.967***
	(0.034)	(0.033)
R-squared	0.232	0.253
Deposition year FE	NO	YES
Time lag (mean)	5.230	5.230
N of structures	69,214	69,214

Notes: This table reports cross-sectional regression estimates where the outcome is the number of years between the deposition of a template structure and the deposition of a downstream structure that reused the template via MR. The template and the downstream structures were identified to be from the same lab if they shared the same PI (last author). The sample consists of 69,214 MR-solved structures deposited at the PDB between 1999-2018 for which a template structure and the PI can be identified. Robust standard errors in parentheses. Statistical significance is indicated as: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix A. Data and Additional Analyses

A.1 UniProt/Swiss-Prot

The Universal Protein Resource Knowledgebase (UniProt) 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 by translating the DNA sequence of a gene, scientists can determine the protein's existence and the sequence of amino acids that appear in the protein. Protein sequences in UniProt are thus sourced by translating genes from major genome sequence databases.

UniProt 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 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. 2015a); TrEMBL primarily contains proteins whose existence is only predicted. Second, since Swiss-Prot primarily includes well-characterized proteins, focusing on Swiss-Prot ensures that I examine proteins with a comparable baseline level of annotation and visibility to structural biologists, rather than unreviewed entries that may not correspond to 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 difficulty and importance scores for each protein, as well as a "predicted brightness" measure, as described in Appendix A.5 and A.6.

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

The PDB provides crosswalks to Swiss-Prot, which can be used 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 can be 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. If 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 identity, given that MR will likely be successful if the template and the target proteins share at least 30% sequence identity.[27] If the identity falls below 30%, MR will be usually challenging, if at all possible, to implement (Schmidberger et al. 2010; Phenix 2022). 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 identity
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

[27] I also restricted the search such that the focal protein and the candidate protein shares at 80% coverage in their alignment in terms of sequence length.

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 identity 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 73,956 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 = 13,145 clusters, which is equivalent to 158,623 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 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 clusters with human proteins are

[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).

getting more structures due to MR or because there are simply more human proteins being discovered. To additionally address the concern of changing cluster size, I have specifications that control for time-varying cluster size.

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.[31] Since my panel starts in 1999, I kept clusters that were born on or before 1998. This led to my final sample of 6,942 clusters (which is equivalent to 136,965 proteins).

A.5 Difficulty and Importance of Clusters

This section describes how cluster-level difficulty and importance scores were constructed. Difficulty features are time-invariant, as they reflect the physicochemical properties of the protein. Importance features are measured as of 2003, the year MR was introduced, and the analysis focuses on proteins that existed by that time. I first compute composite scores at the protein level and then aggregate them to the cluster level.

Below are the features related to technical difficulty of solving a protein:

- Membrane: what % of amino acids of the protein are in membrane regions?
- Disorder: what % of amino acids of the protein are in membrane regions?
- Compositional bias: what % of amino acids of the protein are compositionally biased?
- Sequence length: binned into percentiles, scaled from 0 to 1.[32]

The data is drawn from Perdigão et al. (2015), who constructed these measures using the 2014 release of Swiss-Prot. To account for discrepancies with the 2020 release used in this study, I updated the measures following the same procedures as Perdigão et al. (2015), where needed.

Below are features related to the importance of a protein:

- Number of publications about the protein: binned into percentiles, scaled from 0 to 1

[31] Protein discovery year is the earliest recorded year of its existence, based on either its initial entry in major sequence databases (such as UniProt, the European Molecular Biology Laboratory database, and the DNA Data Bank of Japan) or its first earliest mention in a publication, as documented by UniProt.

[32] Longer proteins are generally more difficult to solve structurally, as they often exhibit increased complexity in expression and crystallization (Büssow et al. 2005), However, some very short proteins (e.g., those under 100 amino acids) can also pose challenges for protein production (Perdigão et al. 2015; Slabinski et al. 2007). Since such sequences are rare in my dataset (less than 5%), I include sequence length as a positive component of the protein difficulty score (i.e., longer = more difficult). In my predicted brightness analysis—where I use Lasso Logit to predict whether a protein has a structure—multiple decile indicators of sequence length were selected, with the shortest deciles associated with a higher likelihood of being solved and the longest with a lower likelihood. This pattern supports the inclusion of sequence length as a proxy for difficulty.

- Is the protein targeted by a drug (binary)?[33]
- Is the function of the protein known (binary)?
- Disease relevance (binary)?
- Is the protein from human (binary)?

Data on importance was parsed from UniProt, except for drug information which was collected from DrugBank. While these importance features serve as conceptually relevant proxies, they likely contain more measurement error than the difficulty features, since observed importance partly reflects accumulated research effort in addition to intrinsic scientific value.

Since all of the features above are either naturally bounded or scaled to range from 0 to 1, I construct composite measures of difficulty and importance by taking the average of the respective features. This approach implicitly assumes equal weights and linearity across features—assumptions that may not fully capture the true relationship between these characteristics and the underlying constructs. One could, in principle, regress these features on whether a protein is bright or dark (as in the predicted brightness measure in Appendix A.6), but this conflates whether brightness reflects scientific importance or technical feasibility—which the regression cannot separately identify. Ideally, difficulty would be measured using data on structure attempts and failures, which is a direction that future research could pursue. For the purposes of this robustness analysis, however, the composite scores offer a transparent and interpretable approximation of technical difficulty and biological importance.

To aggregate to the cluster level, I take the average and standard deviation of the proteins' difficulty and importance scores for each cluster.

A.6 Predicting Brightness

The identification underpinning 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. Proteins in bright clusters may be inherently more important or easier to solve, which could affect their trajectories, regardless of MR.

[33] 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.

To address this concern, I develop a predicted brightness measure, where I measure whether a protein was predicted to be bright in 1998 (i.e. had a structure characterized),[34] using ex-ante traits related to both technical feasibility and biological importance.

Constructing Predicted Brightness

I restrict the sample to proteins that were discovered by 1998 (n = 42,547 proteins, of which 3% had a structure characterized by 1998). For each protein, I compiled features related to both technical feasibility and biological importance as of 1998, using more granular versions of those summarized in Table 1.

Specifically, technical feasibility measures include: (i) % of amino acids in membrane regions, (ii) % of amino acids in disordered regions, (iii) % of amino acids with compositional bias, and (iv) sequence length, binned into deciles.

Biological importance measures include: (i) the number of publications written about the protein (binned into 8 indicators), (ii) the number of drugs targeting the protein (binned into 6 indicators), (iii) whether the protein's function was known (binary), (iv) whether the protein has disease relevance (binary), and (v) species (85 indicators).[35]

I also control for discovery year using 29 indicators (with a single category absorbing all pre-1970 years). All continuous variables were standardized, and binning thresholds were selected to avoid extreme sparsity when creating indicator variables from categorical data. In total, the model includes 143 features.

To estimate the likelihood that a protein was structurally characterized by 1998, I fit a Lasso-penalized logistic regression model using Python's scikit-learn library. The model was trained with 5-fold cross-validation on the full sample, selecting the penalty level that minimized average log loss across folds. After identifying the optimal penalty, the final model was fitted on the entire dataset. Of the 143 input features, the model selected 28 with non-zero coefficients.

As shown in Appendix Figure 3A, the model performance was evaluated using both cross-validated (out-of-sample) predictions and in-sample predictions from the final model. The cross-validated Receiver Operating Characteristic Area Under Curve (ROC AUC)[36] is 0.91, indicating strong ability to distinguish between bright and dark proteins. Given that bright proteins were

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

[35] Species was defined at the genus level. If a genus had fewer than 100 observations, the phylum level was used instead. Species with fewer than 10 observations even at the phylum level were coded as "other."

[36] ROC AUC 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.

rare as of 1998, the Precision-Recall AUC (PR AUC) may provide a more appropriate measure of performance, as it focuses on the model's ability to correctly identify rare positive cases. The model achieves a PR AUC of 0.24, eight times larger than the base rate of 0.03, which is what would be expected under random guessing. The log loss, which evaluates how well-calibrated the predicted probabilities are by penalizing overconfident or incorrect predictions, is 0.09—30% improvement over the log loss of 0.13 that would be obtained under the base rate.[37] In-sample performance is slightly stronger, but the small gap between in-sample and out-of-sample metrics indicates that the model is not overfitting.

Appendix Figure 3B shows the distribution of this predicted brightness, by whether the protein was actually bright by 1998. The variation and overlap in predicted brightness between actually bright and dark proteins is useful: it allows for comparisons between clusters that were similar ex ante but differed in realized brightness. This makes predicted brightness a useful control, isolating the effect of MR from underlying differences in observable protein characteristics.

Finally, from this protein-level prediction, I aggregate up to the cluster-level by taking the average of the predicted brightness of all proteins in each cluster. I then use this cluster-level predicted brightness in two approaches: (i) including it as a control in the difference-in-differences specification and (ii) using it for matching.

Inclusion of Predicted Brightness in the Difference-in-Differences

I modify my baseline difference-in-differences framework to additionally control for predicted brightness. Specifically, I adapt Equation 1 to estimate the following:

$$Y_{ct} = \beta_0 + \beta_1 PostMR_t \times Bright_c + \beta_2 PostMR_t \times Predicted\_Bright_c + \delta_{\mathrm{t}} + \gamma_{\mathrm{c}} + \varepsilon_{\mathrm{ct}} \qquad (2)$$

Equation 2 compares 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 technically feasibility, but differ in whether they were actually structurally characterized.

The inclusion of predicted brightness helps isolate the added effect of actual brightness. If only predicted brightness is associated with post-MR solving activity (i.e., β_1 is non-significant but β_2 is significant), this would suggest that underlying characteristics are driving bright clusters to both have had their structures characterized in 1998 and subsequent structure characterization

[37] Random guessing log loss can be calculated as -(p·ln(p)+(1− p)·ln(1−p)), where p is the proportion of positive cases (bright proteins). Lower log loss is better.

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.

Appendix Table 7A reports the results from estimating Equation 2. The coefficients on *Post-MR$_t$ × Bright$_c$* remain positive and significant; among clusters predicted to be similarly bright, there is still an effect of being actually bright. With the exception of Column 3, the coefficient on *Post-MR × Predicted_Bright$_c$* is insignificant. This pattern reinforces the idea that while predicted brightness captures observed ex-ante traits, it cannot substitute for the presence of actual structural data, which MR requires to operate. These results support the interpretation that MR played a causal role in increasing structure solving in actually bright clusters, above what would be expected based on ex-ante observables.

Propensity Score Matching on Predicted Brightness

As an additional robustness check, I leveraged propensity score matching to construct a more comparable control group. I implemented a one-to-one nearest-neighbor matching on the predicted brightness score, matching without replacement. Unmatched dark clusters were excluded. The resulting matched sample includes all bright clusters and a subset of dark clusters with similar predicted brightness. I then re-estimated Equation 1 using this smaller, matched sample. As reported in Appendix Table 7B, results remain consistent with those from the full sample, suggesting that the findings are not driven by the imbalance in ex-ante observables.

A.7 Alternative Sample Construction

Unbalanced panel: I restrict the sample to include all clusters that were discovered by 2003 (the year MR was introduced), rather than 1998 (the year before my sample begins) as in the main specification. This yields a larger, unbalanced panel of 12,292 clusters. As shown in Appendix Table 8A, results remain similar.

Similarity families: I explore an alternative classification of protein groupings by using UniProt's curated protein families.[38] Families are defined using sequence similarity as well as structural and functional similarities when available. Unlike the main clustering approach, which is based on an algorithmic threshold for sequence similarity defined by the user, the UniProt family classification provides a more "stable" partition of the protein universe since they are

[38] The list of families can be accessed here: https://www.uniprot.org/help/family_membership.

predefined and curated by biologists. These families often capture deeper evolutionary and functional relationships; however, the disadvantage is that they may exhibit lower sequence similarity. Following the same inclusion criteria as in my main specification, I restrict the sample to UniProt protein families that existed as of 1998 and contain at least one human protein.[39] As shown in Appendix Table 8B, results remain similar.

A.8 Identifying Principal Investigators

Assigning Author IDs

Identifying the principal investigator (PI) behind each structure is a challenging task. The PDB does not provide author identifiers and only reports the last names and the initial of first names of authors, which makes it difficult to distinguish between PIs, especially those with common last names. To tackle this, I relied on Torvik (2021)'s Author-ity data, which provides disambiguated author IDs for all articles indexed in PubMed, the leading database of life science articles.[40] This allowed me to identify the author IDs for all PDB structures that had a publication. Following scientific norms, I defined PIs as those who appear as the last author.

22% of structures, however, do not have publications and therefore cannot be directly linked to the Author-ity data. For these unpublished structures, I relied on a combination of a set of heuristics and extensive manual cleaning to match an author ID to each structure. Specifically:

1. I first grouped all structures by the last authors' last names and first name initials.
2. If the last name and the first name initial were associated with a single author ID in the Author-ity data among structures with publications, I applied the author ID to the other unpublished structures that share the same last name and first name initial.
    - i. It is possible that a PI may deposit a few structures in the PDB but never write any PDB paper. In this case, I would mistakenly view two PIs as the same PI. To reduce this likelihood, I manually reviewed PIs whose last names are common and ensured that they appear to be the same PI by reviewing co-authors, structure

[39] Some proteins are not assigned to any UniProt family. For these cases, I explore two approaches: treating them as singleton families or dropping them from the sample. In both cases, the results remain similar. Appendix Table 8B reports the version that includes the singleton families. Additionally, approximately 1% of proteins are classified into multiple families. For these proteins, I randomly assign each to a single family to maintain mutually exclusive units of analysis.

[40] Due to data availability of Torvik (2021), the sample period for my analyses on PI runs from 1999-2016, instead of 2019.

titles, organisms they study, middle initial if available, and years of activity in their publication history. If it appeared that the unpublished structure's PI is in fact a different scientist, I returned to the entirety of the Author-ity data and looked up all authors that share the same last name and first name initial and manually identified the most likely match.

3. For last name and first name initial that were associated with multiple author IDs among published structures: I manually reviewed the unpublished structures to assign the appropriate author ID using similar procedures as 2b.
4. For PIs whose last name and first name initials appear on structures but not on any published PBD papers: these are likely PIs who peripherally engaged in structural biology by depositing a few structures in the PDB, but their research focus was outside of structural biology. To find author IDs for them, I followed the procedures below:
   i. I pulled all potential matches that share the same last name and first name initial in the Author-ity data.
   ii. I searched if any of these potential matches had the first author of the structure as a co-author to narrow the potential matches. I then identified the most likely match by reviewing years of publication activity and publication keywords (MeSH) that involves proteins.
   iii. For those with common last names with too many potential matches, no author ID was assigned to limit false matches.

Using these procedures, I identified author IDs for 98% of structures.

Identifying Affiliations

To identify affiliation, I leveraged Torvik (2021)'s MapAffil data, which standardized free-text affiliations available in PubMed articles.[41] PubMed started providing affiliation data of first authors in 1988 and affiliation for all authors in 2014. As a result, affiliation coverage is limited before 1988 (MapAffil merged on complimentary sources, where available), and only the affiliation data of first authors is available before 2104. I followed the procedures below to identify affiliations:

[41] The benefit of using the MapAffil data is that it provides Global Research Identifier Database (GRID) codes. This allows me to obtain standardized affiliations from which to match to the 2023 QS World University Rankings in Life Sciences and Medicine to identify PIs from elite vs. non-elite institutions. One limitation of relying on this ranking is that it does not rank elite research institutions that are not universities.

1. For structures with publications, I used the affiliation of the last author, where available, and the first author if not. Because the norm in science is such that the first author is usually the lead graduate student or the post-doc, it is reasonable to assume that the first and last authors would share the same affiliation.
2. For structures without publications: since I have the author ID, I identified other papers by the same author in the year to infer the author's affiliation.

Using these procedures, I identified affiliations for 87% of structures.

Appendix B. Solving the Phase Problem

As explained in Section 3.2.2, elucidating a protein structure requires crystallizing the protein and exposing the crystal to X-rays. This generates diffraction patterns as X-rays reflect off the crystal. Using a combination of statistical and physical principles, scientists analyze the diffraction data to construct a 3D model of the protein structure.

The “phase problem” is a critical challenge in this process. While X-ray reflections contain both amplitudes and phases, only the amplitudes can be directly measured from diffraction patterns. Without the phase information, the protein structure cannot be reconstructed.

This appendix details three methods used to solve this phase problem.[42] Two experimental phasing methods—isomorphous replacement and anomalous diffraction—solve the phase problem *de novo*, from scratch. These methods do not require templates of previously solved structures but are complex and labor-intensive. The third method, molecular replacement, is a computational approach that bypasses experimental phasing by leveraging existing templates.

B.1 Isomorphous Replacement

The oldest experimental method to solve the phase problem *de novo* is called isomorphous replacement. This method involves making a known change to the target protein and analyzing how the change affects diffraction patterns. Structural biologists must produce at least two different types of crystals: a “native” target crystal and a “derivative” crystal with a heavy metal ion introduced. By comparing the diffraction patterns between the native and the derivative crystals, structural biologists can locate the heavy atoms and deduce possible phases of the other atoms in the proteins. Isomorphous replacement typically necessitates multiple derivative crystals with different heavy metals to fully solve the phase problem.

Isomorphous replacement can be laborious. Crystallizing proteins is a difficult process, and isomorphous replacement requires multiple derivations of the crystal, increasing the challenge involved. Furthermore, the metal ion must be introduced in a way that does not disturb the protein structure (i.e., maintain isomorphism), which can be difficult to achieve (Foos, Rizk, and Nanao 2022) and necessitates many rounds of trial-and-error.

[42] To construct this appendix, I consulted and adapted various sources. In particular, Cowtan (2003), Read (2005), and Terwilliger et al. (2016) were especially helpful.

B.2 Anomalous Dispersion

The second experimental method to solve the phase problem *de novo* is called anomalous dispersion, where structural biologists vary the X-ray wavelength to induce atoms of specific elements in the protein to produce anomalous scattering. While the scattering of atoms is usually independent of the wavelength of the X-ray, at their respective "absorption edges," each atomic type produces anomalous scattering, which introduces differences in the intensities of certain diffraction spots. By identifying the positions of these anomalous scattering atoms, structural biologists gain clues to recover the missing phase information of the rest of the protein.

While anomalous dispersion has advantages over isomorphous replacement (in particular, anomalous dispersion only requires a single crystal), it comes with its own challenges. The experiment has to be conducted at high precision because signals from anomalous dispersion are weaker than those from isomorphous replacement (Leonard 2018). Because small differences in diffraction patterns must be detected, anomalous dispersion experiments require access to synchrotron facilities that can provide tunable X-ray beams at high intensities (Oksanen and Goldman 2010). Moreover, radiation damage to the crystal is a concern when structural biologists need to tune the X-ray wavelength multiple times, as many sets of data are collected from a single crystal.

B.3 Molecular Replacement

Molecular replacement (MR) allows structural biologists to bypass the time-intensive and challenging experimental phasing methods discussed above, eliminating the need for additional experiments. Rather than solving the phase problem *de novo*, MR imports phase information from existing structures that are similar to the unknown target structure.

MR proceeds in the two steps: (i) identifying previously solved structures that can be used as templates and (ii) orienting the template structure to match the position of the unknown target structure within the crystal. After these two steps, the phase information can be transferred from the template to the target protein.

First, the key insight behind molecular replacement rests on empirical patterns. Proteins are composed of sequences of amino acids, and sequence similarity has been observed to be strongly correlated with structural similarity. Although the exact mechanisms of how specific amino acid sequences determine a protein's 3D shape are not fully understood, this correlation allows structural biologists to leverage MR.

Specifically, because sequence similarity is correlated with structure similarity, structural biologists can reasonably assess whether a previously solved structure will serve as a suitable template for the unknown target protein based on sequence similarities before attempting MR. The closer the sequence similarity, the more likely that the previously solved structure and the unknown structure of the target protein are also analogous. As a general rule of thumb, MR will likely work if the template and target proteins share at least 30% sequence identity. If the sequence identity falls below 30%, MR will be usually challenging to implement (Schmidberger et al. 2010; Phenix 2022).

Second, to import the phase information from the known template to the unknown target protein, the template protein must be correctly oriented and positioned within the repeating unit of the target protein crystal. The template structure is first rotated in three dimensions, and the resulting amplitudes and phases are calculated for every orientation. The software program Phaser uses maximum likelihood to identify the orientation that best matches the observed experimental diffraction patterns of the unknown target protein. The oriented template structure is then placed at every possible position in the unit cell. Again, the position that best matches the observed experimental diffraction patterns is chosen.

At the end, Phaser outputs a list of possible solutions with a noise-to-signal metric called the "Z-score," which provides guidance on whether the solution has been identified. Once the correct orientation and position of the template structure are found, phase information can be directly imported into the unknown target protein.

Molecular Replacement vs. Machine Learning

One note is that while MR and machine learning methods are both data-extrapolation approaches that rely on the existence of prior data from which they import information, the principles behind MR and machine learning are distinct: MR is based on a pre-specified model grounded in the laws of physics, whereas as machine learning learns the model directly from the data without such pre-specified models.

Machine learning takes training data as input, infers a model from the statistical patterns in the training data, and then applies the model in unseen test data. In contrast, MR does not "learn" a model from the data like machine learning; it instead relies on the physics of X-ray diffractions to guide the process. MR involves solving a system of equations that describes the interaction of X-rays with the protein crystal lattice. Using numerical optimization, MR determines the correct orientation and position of a template protein within the unit cell to match the target protein.

While the specifics of implementing MR come from models of physics, the ability to use MR stems from empirical patterns. As described earlier, although the exact mechanisms of how amino acid sequences determine 3D structures remain unclear, structural biologists take advantage of the observed correlation between sequence similarity and structural similarity. When structural biologists decide what structure to solve, they have a reasonable assessment of whether they can use MR based on the availability and sequence of previously solved structures. This allows them to assess whether they will be able to bypass the labor-intensive experimental phasing methods and use MR. Finally, like machine learning, the key limitation of MR is that it can only work by analogy to known structures in the PDB. Just as supervised machine learning requires training data, MR cannot be applied if there is no data of similar, previously solved templates.