NBER WORKING PAPER SERIES

HOW ARTIFICIAL INTELLIGENCE SHAPES SCIENCE: EVIDENCE FROM ALPHAFOLD

Ryan R. Hill
Carolyn Stein

Working Paper 35143
http://www.nber.org/papers/w35143

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2026

We are especially grateful to Paul D. Adams, Susan Tsutakawa, Joanna Slusky, and Brian Shoichet for providing their expertise and insight into their respective fields of science, and to Sylvain Poux and Elisabeth Gasteiger for sharing their expertise on UniProt. We thank Daron Acemoglu, David Autor, Eric Budish, Florian Ederer, Daniel Gross, Anders Humlum, Benjamin Jones, Juan Mateos-Garcia, Jeremy Stein, and seminar participants at Berkeley/Stanford IOFest, Northwestern University, University of Kansas, Johns Hopkins University, Berkeley Haas, and University of Chicago for comments that substantially improved the paper. Thomas Barden and Alexia Witthaus Vine provided excellent research assistance. All errors and omissions are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Artificial Intelligence Shapes Science: Evidence from AlphaFold
Ryan R. Hill and Carolyn Stein
NBER Working Paper No. 35143
April 2026
JEL No. I23, J24, L65, O31, O33

ABSTRACT

We study how a frontier AI model affects scientific discovery by examining the release of the AlphaFold2 algorithm and its impact on structural biology and related fields of science. Structural biology is the field of science concerned with understanding the structure and function of proteins. Researchers in this field historically devoted substantial time and resources to experimentally solving three-dimensional protein structures. AlphaFold can predict these structures without running experiments. In July 2021, researchers gained access to hundreds of thousands of these AI-predicted structures virtually overnight. Yet, to date, we find that the rate of experimental structure determination has remained almost unchanged. Instead, researchers appear to use predicted structures to facilitate and complement experimental structure determination. Looking at downstream science that builds on protein structures, we find that basic research on proteins that had no structure information prior to AlphaFold increases by 15 to 40% relative to proteins that already had a structure, shifting the direction of research toward less-studied proteins. However, we find no evidence so far that more applied, early-stage drug development is targeting these proteins, though such activity may emerge in the future.

Ryan R. Hill
Northwestern University
Kellogg School of Management
Strategy
and NBER
ryan.hill@kellogg.northwestern.edu

Carolyn Stein
University of California, Berkeley
and NBER
carolyn_stein@berkeley.edu

1 Introduction

Artificial intelligence (AI) is rapidly reshaping industries in real time. A growing body of evidence suggests that AI is automating an increasingly broad set of economic tasks, ranging from routine work in call centers (Brynjolfsson, Li, and Raymond, 2025) to creative tasks in photography and digital design (Goldberg and Lam, 2025; Zhou and Lee, 2024), as well as high-skill occupations such as software engineering and radiology (Cui et al., 2025; Agarwal et al., 2023). In many cases, breakthroughs are improving productivity in the market for goods and services. At the same time, AI is also beginning to influence the production of ideas, which has tantalizing implications for the possibility of accelerated economic growth (Aghion, Jones, and Jones, 2017). The non-rivalrous and cumulative nature of ideas suggests that the automation of research tasks can have amplifying effects for the productivity of the broader economy (Romer, 1990; Jones, 1995; Weitzman, 1998; Jones, 2005). Practically, however, the R&D pipeline is full of bottlenecks. The potential complementarity between tasks in a long and uncertain research pipeline may act as a natural damper on the pace of progress, even when AI fully automates many steps along the chain (Jones, 2025). Despite many promising anecdotes about the changing pace of AI-enabled R&D, there has been scant empirical evidence about how AI is changing the research process in practice.

In this paper, we provide some of the first systematic evidence on how AI affects scientific discovery, using the introduction of protein structure prediction models as an important case study. Specifically, we study AlphaFold2, a Nobel-prize winning machine learning algorithm that rapidly increased the availability of predicted structures relevant for a broad range of basic and applied research fields related to proteins.[1] This was a massive technology shock to structural biology, the field of research that seeks to determine the three-dimensional structures of proteins.[2] These structural models help us understand how proteins function inside of cells and how they might be targeted in pharmaceutical therapies. Since the early 1970s, we have relied on slow and expensive experimental methods to generate the roughly 130,000 unique publicly available protein structures. At the same time, advances in DNA sequencing have made it inexpensive and routine to identify the genetic makeup of proteins, leading to databases containing hundreds of millions of protein sequences—of which less than 0.1 percent have known structures. This gap motivates the longstanding *protein folding problem*: can we predict a protein's three-dimensional structure from its genetic code, without

[1] There is an earlier version of the algorithm, released in 2018, known as AlphaFold. However, because AlphaFold2 represented the primary advance, throughout this paper we will use the phrase "AlphaFold" to refer to AlphaFold2, unless explicitly stated otherwise.

[2] In recent years, several economics of science papers have used this field of science as an empirical setting, including Hill and Stein (2025b), Hill and Stein (2025a), Kim (2025), and Zhuo (2025).

running any experiments?

For decades, computational biologists and computer scientists made slow progress on improving the capabilities of computational models. A breakthrough occurred when Google DeepMind debuted the AlphaFold2 algorithm in late 2020. For the first time, an AI model achieved near-experimental levels of accuracy at a vastly reduced cost in terms of time and resources, to the surprise of many in the research community. Figure 1 shows examples of two predicted protein structures overlaid with their experimental counterparts.

Many expert and casual observers characterized this advance as effectively solving the protein folding problem and anticipated far-reaching implications for both basic scientific research and medical innovation. One of the scientists tasked with assessing the algorithm's performance, evolutionary biologist Andrei Lupas, memorably said of AlphaFold2, "This will change medicine. It will change research. It will change bioengineering. It will change everything." Structure prediction tools may open up new paths for scientific research about protein function, molecular mechanisms, and cellular processes. They may also reshape early-stage drug discovery by enabling structure-based approaches to new molecule discovery. Downstream outcomes such as new approved drugs may take years to materialize, but the sudden expansion of high-quality structural information represents a major technological shock to early stages of this process. AlphaFold therefore provides a natural setting to study both the narrow and broader downstream impacts of artificial intelligence on scientific progress.

While AlphaFold is not the only major AI advance in science, it presents a particularly compelling case study for several reasons. First is its timing. AlphaFold2 was first developed in late 2020, with predicted structures made widely available in mid-2021. Thus, it arrived early compared to other major AI breakthroughs,[3] giving us more time to understand its impact. Second, the breakthrough was unexpected. While the first version of AlphaFold, launched in 2018, outperformed competitors, it was not considered good enough to replace experimental structure determination. The breakthrough improvement in AlphaFold2 was unexpected by the scientific community, and led experts to declare that the protein folding problem had been solved. Conference attendees who saw the initial reports of AlphaFold2's prediction accuracy claimed to be "in shock" (Ball, 2024). Third, the nature of the algorithm is particularly clean and concrete. Compared to other AI tools—such as large-language models, chatbots, etc.—it is unusually clear what tasks AlphaFold is potentially automating. This makes it easier to interpret subsequent shifts in research pace and direction. Finally, it is an important application. Structural biology and protein folding is a critical area of science. Several Nobel Prizes have been awarded for the discovery of a *single* experimental protein

[3] For example, the first iteration of ChatGPT did not launch until November 2022.

structure, and these structures enable important advances in the understanding of biological processes, disease, and drug design. Moreover, the AlphaFold2 algorithm was awarded the 2024 Nobel Prize, an unusually early indicator of its perceived scientific importance.

The goal of this paper is to understand how this shock impacted the production of science, both in the narrow field of experimental structural biology, and in downstream fields of science and R&D that build on these structures. We first present evidence on how the introduction of AlphaFold affected experimental structural biology. For researchers in this field, the arrival of AlphaFold had the potential to dramatically reshape the production technology for structure determination. Yet we find limited evidence that it has substituted for experimental structure determination thus far. Using the universe of experimental protein structures released in the Protein Data Bank (PDB), we find that since the introduction of AlphaFold and the first wave of predicted structures in 2021, there has been no noticeable decline in the number of experimental structures being solved and uploaded to the PDB. Moreover, the number of papers reporting structural biology experiments has also stayed consistent, including in papers that are published in the top general-interest science journals. Despite the impressive ability of the AI tool to generate accurate structures, scientists are still producing experimental research and publishing it in good venues. Further, we find limited evidence that they are shifting their research towards areas where the AI tool has low accuracy, which might be the case if AlphaFold served as a substitute for some—but not all—structures. We raise two important caveats to these findings. First, we note that this lack of substitution may not be efficient, but rather may reflect researcher incentives—expert scientists, many of whom are tenured, may prefer to continue doing the science they are uniquely trained to perform (Azoulay, Fons-Rosen, and Graff Zivin, 2019). Second, it may also be the case that AI is not *yet* a substitute for this experimental work, but it may one day replace it as the models improve and researchers trust the output more. Still, it is striking that with several years of data, we see no reduction in expensive experimental work.

Despite these apparent non-results, we find strong evidence that structural biologists are using AlphaFold, and that it is *complementing* their experimental work. Experimental structure prediction involves computational steps that can be accomplished faster and more accurately using existing structures as a template. With these methods, researchers can use the template as a starting point in their three-dimensional model, rather than building it entirely from scratch. In the past, scientists would rely on similar experimental structures as templates, limiting their ability to use these more efficient methods in exploratory work about novel proteins. However, AI-predicted structures can also be used as templates. After AlphaFold, we see a sharp uptick in the use of these methods, concentrated among structures that lack a similar experimental structure. This complementarity between experimental and

AI-based research may open opportunities for greater exploration of the protein space.

Second, we turn our attention to the broader impacts of AlphaFold on related disciplines and downstream pharmaceutical R&D. Here we exploit the insight that only a small share of known proteins had an experimental structure model prior to AlphaFold. This creates a natural experiment at the protein level. Proteins with an experimentally-solved structure did not experience much—if any—new structural information from the arrival of predicted structures. This knowledge was largely redundant given the existing high-quality experimental evidence. On the other hand, for proteins *without* an experimentally-solved structure, AlphaFold represented a large shock of new structural information. Thus, we can conceptualize the former group as the "control" group and latter group as the "treated" group. Then we can compare protein-level research activity in broader and downstream fields across these two groups in a difference-in-differences design.

To do this, we use data from the Universal Protein Resource (UniProt), a database of all known proteins. Because of the scale (UniProt contains information on over 200 million proteins) we focus on the more curated 570,000 protein subsample of this data source, known as SwissProt. This includes proteins that are of higher scientific importance (and includes all human proteins). By combining these data with the PDB, we are able to identify which proteins had an experimental structure prior to AlphaFold, and which did not. Even among the selected SwissProt sample, only 8% of proteins had an experimental structure.

The SwissProt subsample also curates a literature review for each protein, linking basic scientific papers that have been written about the protein to their SwissProt entries. This allows us to use scientific research activity as an outcome. We find statistically and economically significant increases in research activity about previously unsolved proteins compared to their peers that already had experimental structures. New papers about protein function, gene expression, protein-relevant disease, and other protein science categories appear after AlphaFold and are disproportionally focused on proteins where AI revealed its structure for the first time. Our estimates suggest that research on unsolved proteins increased by 16 to 25% relative to previously solved proteins, with even larger effects toward the end of our sample period. We also show that new papers about previously unsolved proteins are much more likely to cite core AlphaFold papers, suggesting that the appearance of AI models are a key mechanism driving this shift in research direction. Together, these results suggest that protein structure determination was likely a key bottleneck in complementary research that has been eased by the introduction of AI tools. Moreover, taken with the complementarity results above, these results suggest that AlphaFold is making exploration of the broader protein space easier, having a sort of "floodlight" effect on protein science. This is consistent with the effects observed in other settings after broad increases in data access (Nagaraj,

Shears, and de Vaan, 2020; Nagaraj, 2021; Tranchero, 2024).

We also investigate whether these new structures led to increased R&D activity in the pharmaceutical space, as was initially hoped by observers. While it is too early to expect to see new drugs on the market (or even in clinical trial), we might expect to see increased early-stage drug investigation. Since most drugs work by binding to a protein target, one of the first steps in drug discovery is testing whether small molecules (potential drugs) bind to these targets through bioactivity experiments. We use data on these bioactivity experiments from a source called ChEMBL, which curates them from the scientific literature.[4] We are able to link these experiments back to SwissProt using the protein targets.

In contrast with our basic science results, we see no similar uptick in early stage drug research about previously unsolved proteins. Comparing the number of bioactivities in ChEMBL, we see no change in attention toward AI-illuminated protein structures in the three years since the introduction of AlphaFold, though the results are noisy. This finding potentially speaks to the role of bottlenecks and complementarities in the research process. Our results suggest that more basic scientific research needs to occur before scientists can make use of these new protein structures as potential drug targets.

Related Research

This paper relates to three literatures. First, it connects to the economics of automation and AI, which studies how new technologies reshape labor markets (Autor, Katz, and Krueger, 1998; Autor, 2015; Humlum, 2021; Boustan, Choi, and Clingingsmith, 2024; Feigenbaum and Gross, 2024). The task-based framework developed by Autor, Levy, and Murnane (2003) and Acemoglu and Autor (2011) has been influential in this literature, allowing researchers to study whether technology substitutes for labor in existing tasks, complements workers within those tasks, or creates new tasks altogether. Using the task model, Acemoglu and Restrepo show that automation generates a displacement effect when capital (or AI) takes over tasks previously performed by labor, but technological change can also reinstate labor demand by creating new tasks and reorganizing production (Acemoglu and Restrepo, 2019, 2022; Restrepo, 2024). Recent empirical work on generative AI often finds sizable productivity gains in certain domains,[5] with especially large gains for less experienced workers (Brynjolfsson, Li, and Raymond, 2025; Noy and Zhang, 2023), which could be interpreted as evidence for augmentation rather than one-for-one replacement in the settings studied. Our paper brings this question to a frontier scientific setting in which the potentially automated

[4] Note that private pharmaceutical firms perform bioactivity research that they often don't publish and is therefore not recorded in public datasets such as ChEMBL.

[5] Though not necessarily aggregate, as shown in Humlum and Vestergaard (2026)

task—protein structure determination—is unusually well defined.

Second, this paper relates to a growing literature on AI and science. Agrawal, McHale, and Oettl (2023, 2024) model AI as a tool that improves scientific discovery by prioritizing search over large hypothesis spaces when experimentation is costly. Ludwig and Mullainathan (2024) similarly emphasize that machine learning can contribute to science not only through prediction, but also by generating hypotheses. Jones (2025) places these ideas in a broader R&D framework, arguing that AI's impact depends on the share of research tasks it can perform, the productivity gains on those tasks, and the extent of remaining bottlenecks. Cockburn, Henderson, and Stern (2019) and Mullainathan and Rambachan (2025) go further and argue that algorithms may reorganize scientific production itself, including idea generation and theory formation. Relative to this literature, our contribution is empirical and field-specific. Rather than studying AI as a general-purpose research tool, we examine a sharp technological shock in one scientific domain and trace its effects.

Third, and most narrowly, this paper relates to work on AlphaFold and its effects on science. Within structural biology, Edich et al. (2022) argue that AlphaFold is often used to assist experimental structure solution rather than simply replace it, while Kovalevskiy, Mateos-Garcia, and Tunyasuvunakool (2024) review early evidence of widespread adoption and emphasize that AlphaFold has sped up structure determination and enabled new workflows while leaving important roles for experiment. Varadi and Velankar (2023) provides some case studies of how AlphaFold has impacted downstream science, including drug discovery. In the economics and management literature more specifically, Cavalli (2024) studies how the release of the first AlphaFold algorithm in 2018 impacted the organization of academic labs also working in the protein prediction space, finding that labs run by life science specialists reacted by hiring more computer scientists. Yu (2026) uses the introduction of AlphaFold2 as a shock to study scientists' careers, showing that highly productive structural biologists were more likely to adopt AlphaFold, and that this exacerbated citation polarization between more and less-cited researchers. In work most closely related to this paper, Qian (2026) studies how experimental structural biologists change the proteins they study in response to AlphaFold, finding that researchers pivot away from proteins that are well-predicted, especially if they have limited downstream demand.

The remainder of this paper proceeds as follows. Section 2 describes the institutional background and provides a scientific primer. Section 3 describes the data sources and sample construction. Section 4 describes the empirical design and presents results. Section 5 concludes.

2 Institutional background and scientific primer

2.1 Structural biology and the protein folding problem

Structural biology is the study of the form and function of biological macromolecules, especially proteins.[6] Researchers in this field perform advanced experiments to elucidate the three-dimensional shapes of proteins, which are too small to observe under an optical microscope. Traditional experimental approaches such as x-ray crystallography and cryo–electron microscopy (cryo-EM) are time-consuming and expensive. Solving a single protein structure can take months to years and may cost on the order of $100,000 or more, depending on the method and difficulty of the target (Sullivan, Brennan-Tonetta, and Marxen, 2017).[7] Since the field's development in the 1970s, researchers have solved and publicly deposited around 130,000 unique protein structures.

Protein structures are important because they reveal how proteins function inside the cell. Structural maps allow researchers to see how the protein folds, where it binds to other molecules, and how mutations might alter its behavior. Structural information also enables advances in disease biology and drug design. For example, structural insights into the Cas9 endonuclease were critical to the development and refinement of CRISPR-based gene editing technologies (Jinek et al., 2014). More recently, the rapid determination of the SARS-CoV-2 spike glycoprotein structure enabled the rational design of stabilized spike antigens used in mRNA Covid-19 vaccines (Wrapp et al., 2020).

Proteins are composed of chains of small molecules known as amino acids. In almost all organisms, there are 20 standard amino acids used to build proteins. A protein's three-dimensional shape arises from the physical and chemical interactions among these amino acids, which cause the chain to fold into a specific structure. The amino acid sequence of a protein is coded directly from DNA. As a result, determining a protein's amino acid sequence is comparatively straightforward: it can be inferred directly from genomic sequencing data.[8] By the completion of the Human Genome Project in 2003, essentially the full set of human protein-coding sequences had been identified (International Human Genome Sequencing Consortium, 2004). The advent of next-generation sequencing technologies in the mid-2000s dramatically reduced the cost of DNA sequencing and enabled large-scale se-

[6] Other macromolecules of interest include deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). However, in practice, proteins command the vast majority of scientists' attention and are the focus of this paper.

[7] Although cryo-EM has accelerated structure determination for many large proteins, it requires multi-million dollar instrumentation and historically produced lower-resolution structures than crystallography, particularly prior to the so-called "resolution revolution" in the 2010s (Nogales, 2016).

[8] Each amino acid is coded by three DNA base pairs. For example, the DNA sequence ATG codes for the amino acid methionine.

quencing of other organisms, viruses, and bacteria (Metzker, 2010). Today, public databases contain sequences for hundreds of millions of proteins across all domains of life (The UniProt Consortium, 2023). Despite this explosion in sequence information, only a tiny fraction (less than 0.1%) of known proteins have experimentally determined three-dimensional structures.

Given that amino acid sequences are comparatively cheap and easy to deduce compared to experimental structures, it is not surprising that scientists have long been interested in trying to predict how proteins will fold based on their amino acid sequence alone. Indeed, the so-called *protein folding problem*—determining a protein's three-dimensional structure from its amino acid sequence—has been a central open question in structural biology for more than half a century (Anfinsen, 1973). In practice, however, the mapping from sequence to structure is extraordinarily difficult due to the vast number of possible conformations a protein can take.[9]

2.2 Structure prediction and AlphaFold

Despite the challenges associated with structure prediction, computational biologists launched the Critical Assessment of protein Structure Prediction (CASP) in 1994 as a recurring, blinded evaluation of prediction methods (Moult, Pedersen, et al., 1995; Moult, Fidelis, et al., 2014). Every two years, groups could submit predictions of protein structures whose experimental coordinates had been determined but not yet released publicly. Once the experimental structures were made available, predictions were scored by organizers against the experimentally-determined ground truth. A key accuracy metric is the "Global Distance Test-Total Score" (GDT_TS) which measures how closely a predicted and experimental structure match after optimal superposition. At a high level, it measures the fraction of amino acids that are placed "close enough" to the experimental ground truth (Zemla, 2003). CASP assessors had long treated a GDT_TS score of 90 as equivalent to experimental accuracy (Kryshtafovych, Schwede, et al., 2021).

While the technical details of these prediction models are beyond the scope of this paper, one basic point is important. Modern protein structure prediction models make heavy use of information from evolution. By comparing related protein sequences across many organisms, they can detect statistical patterns that reveal which amino acids are likely to be close together in the folded structure. These patterns help constrain the set of plausible three-dimensional shapes. Importantly, AlphaFold2 does *not* depend primarily on having a closely

[9] In 1969, biophysicist Cyrus Levinthal put this into perspective with "Levinthal's Paradox," which states that a small protein of 100 amino acids has about 10^{47} possible conformations. Even if proteins could sample conformations at the speed of molecular vibration—10^{13} per second—it would take longer than the age of the universe to sample them all (Levinthal, 1969).

related experimental structure already in hand. Instead, it can infer structure directly from the amino acid sequence, together with evolutionary information from related sequences. Consistent with this, Jumper et al. (2021) show that AlphaFold2 performs nearly as well even when experimental structure information is excluded. As a result, AlphaFold can perform well even for proteins without a similar experimental structure in the training data (what economists might call "out of sample").

Progress at CASP was slow, with winning teams' average GDT_TS scores rarely exceeding 40 through 2016. A discrete shift occurred at the 13th edition of CASP in 2018, where Google DeepMind's AlphaFold system appeared for the first time and substantially outperformed competing approaches, achieving an average GDT_TS score of nearly 60 across all targets. When DeepMind returned in 2020 with AlphaFold2 (Jumper et al., 2021), they astonished CASP evaluators, participants, and observers when they achieved an average GDT_TS score of nearly 90 across all competition targets. This led John Moult, the founder of CASP, to say "in some sense, the [protein folding] problem is solved" (Callaway, 2020).

Although DeepMind may have been expected to win, the progress they made—and the declaration that the protein folding problem had been solved—was unexpected by the structural biology community. Coverage of CASP 14 in 2020 describes the breakthrough as "remarkable" and "startling" (Callaway, 2020; Kryshtafovych, Schwede, et al., 2021), with some observers describing an atmosphere of shock and denial (Ball, 2024).

The news of DeepMind's success at CASP was released in November of 2020. By July of 2021, they had released the underlying code for AlphaFold2. Simultaneously, they ran the code on hundreds of thousands of known amino acid sequences, and uploaded the predicted protein structures directly into a freely accessible database.[10] This made access to the structure predictions frictionless for researchers (Varadi and Velankar, 2023).

2.3 Structural biology and drug discovery

Much of the excitement around AlphaFold was focused on the downstream scientific advances it might enable. In Figure 2, we outline a simplified modern drug discovery pipeline to illustrate how AlphaFold predicted structures might play a role in pharmaceutical R&D, and what aspects of the pipeline we can measure with the data sources described in Section 3. While drug discovery is not the only important downstream application of AlphaFold, it is the one that has garnered the most excitement.[11]

[10] In the following months, DeepMind expanded the number of structures covered in this database. We discuss this more in Section 3.

[11] Other important applications include protein engineering, enzyme design, and synthetic protein design, which have uses in chemical manufacturing, waste degradation, and vaccine design. Researchers in these fields modify or generate entirely new amino acid chains. AlphaFold can help by giving a structural starting

Most drugs are small molecules that function by binding to a protein target, inhibiting or modulating its activity (Santos et al., 2017). An example is saquinavir, a drug that targets HIV-1 protease. HIV-1 protease is an enzyme that the virus uses to replicate inside of cells. Thus, blocking this enzyme interrupts the viral life cycle at a critical step (Erickson et al., 1990). With this example in mind, we can walk through the four steps outlined in Figure 2.

The first stage of drug discovery is known as target discovery and validation. This is still basic science. The goal of this step is to determine how a protein (a potential target) is involved in a disease. This typically involves a combination of genetic and cell biology experiments: researchers might use gene knockouts or knockdowns to ask what happens to a cell or organism when a gene is disabled, or they might overexpress a protein to see what goes wrong when there is too much of it. Structure determination is often part of this process. Sometimes, new information about a potential target galvanizes structural biologists to solve a protein's structure. Other times, a protein structure provides additional insight for how a protein might be involved in disease. In the case of HIV-1 protease, researchers established early in the AIDS epidemic that the protease was essential for viral replication—experiments showing that mutations disabling the enzyme produced non-infectious viral particles made it a compelling target. This biological validation motivated structural biologists to solve the enzyme's three-dimensional structure, which was first reported in 1989 (Navia et al., 1989; Wlodawer et al., 1989).

The second stage is known as hit identification and optimization. Some of this research is done in academic labs, though much of it is also done by industry. Once researchers are confident about a protein target's role in disease, they will start testing whether small molecules (candidate drugs) bind to that protein. This is often done through bioassay experiments: at a high level, researchers combine the protein target with different molecules and measure the outcome (does it bind, inhibit, etc.). At this stage, it helps to have a protein structure, as a clear map of the protein can make this step more efficient in at least two ways. First, it enables virtual screening: before any physical experiments, researchers can computationally simulate which small molecules are likely to bind to the protein based on the shape of its binding site, narrowing the field of candidates. Second, once a promising candidate has been identified, structural information guides its optimization—showing researchers exactly how the molecule sits in the binding site and where chemical modifications might improve its fit, potency, or selectivity. For HIV-1 protease inhibitors, the experimental structures revealed a well-defined binding site. Researchers used the shape of this binding site to design

point for where to mutate and what parts of the protein are likely to matter. In the case of entirely new amino acid chains, researchers can run the algorithm directly on the chain to get a prediction of its hypothetical structure (Kovalevskiy, Mateos-Garcia, and Tunyasuvunakool, 2024).

molecules that would fit snugly into it and block the enzyme's activity—an approach known as structure-based drug design.

Promising candidates go to the third step, which involves studies to understand how the drug is metabolized and whether it might be toxic. Much of this research is conducted in animal models. Researchers measure how quickly the drug is absorbed into the bloodstream, how it is distributed across tissues, how the liver metabolizes it, and how it is excreted. They also conduct toxicology studies at a range of doses to identify any harmful effects on organs, as well as tests for genotoxicity (whether the drug damages DNA). If the compound clears these hurdles, the researchers compile their findings into an Investigational New Drug application to the FDA, which, if approved, grants permission to begin human testing.

The final step is FDA clinical trials, which are used to determine whether the drug is safe and effective in humans. For HIV protease inhibitors, Phase III trials demonstrated dramatic reductions in viral load. This process ultimately produced saquinavir, the first HIV protease inhibitor approved by the FDA, in 1995. While the saquinavir example is illustrative, it is worth noting that its timeline was highly unusual. Given the urgency of the AIDS epidemic, regulatory agencies developed accelerated review pathways that compressed the typical timeline from 10-15 years down to only 6. For this reason, we focus on the earlier stages of the process in this paper, as this is where we might realistically expect to see results in the five years since AlphaFold's release.

3 Data sources and construction

In this section, we describe our data and how we construct our main analysis samples. For more detail, please see Appendix A.

3.1 The Protein Data Bank (PDB)

Our primary data source for experimental structure discoveries is the Protein Data Bank (PDB). Founded in 1972, the PDB is an online repository of biological macromolecular structures that are deposited by structural biologists around the world. Since the 1990s, deposition in the PDB has been mandatory for researchers seeking to publish academic papers (Strasser, 2019). Currently, there are around 250,000 deposits, 95% of which are proteins, and the database is growing by 10% every year. The primary information that the PDB is designed to host is the three-dimensional molecular coordinates that describe the proposed protein structure. They also collect and categorize a rich set of metadata for each project, including protein characteristics, experimental details, and verified quality metrics.

We observe three key dates that help us re-construct the timeline of their research. First is the collection date, which is the date that scientists collect their experimental data. Second, is the deposit date, which is a timestamp for when the project details were first uploaded to the PDB servers. At this point, the data is held confidentially while the authors typically go through the journal submission and revision process. Finally, the data is made public on the release date, usually coordinated with journal publication, or otherwise a maximum of one year after the deposit date. The median time between collection and deposit is 11 months, and the median time between deposit and release is 5 months.

In addition to these dates, we also gather experimental details from the PDB. This includes the broad experimental method that the researchers use to determine the structure (for example, x-ray crystallography or cryo-EM). It also includes the computational techniques they use to turn their data into three-dimensional structures, the software and hardware they used, and any existing protein structures they used as templates.

3.2 AlphaFold Protein Structure Database

When Google DeepMind first used AlphaFold to predict protein structures at scale, they partnered with the European Molecular Biology Laboratory (EMBL) to host the predicted structure data for public use. In July 2021, DeepMind released the first 365,000 predicted protein structures as well as the open-source code for the prediction model. These proteins were selected for their biological relevance, and included almost all human proteins. In December 2021, they released an additional 560,000 structures covering most of SwissProt, and in July 2022, they released a prediction for every known protein (200 million+ structures). We treat July 2021 as the treatment date, because this is when researchers gained access to predicted structures for the first time.

This dataset, shortened to AlphaFold DB, is our primary source of data for AI-predicted structure details. For each structure, AlphaFold DB provides a confidence score known as average predicted Local Distance Difference Test (pLDDT). pLDDT represents how confident the model is in the predicted position of each amino acid in a protein structure, and average pLDDT takes the average across all amino acids in the structure. Multiple studies have shown that pLDDT is strongly correlated with actual prediction accuracy, using proteins that have an experimental structure that was not used in the model training (i.e., in a holdout sample) (Jumper et al., 2021; Tunyasuvunakool et al., 2021; Edmunds, McGuffin, and Genc, 2024).

3.3 UniProt and SwissProt

In order to study the impact of AlphaFold on broader areas of basic protein science, we rely on the Universal Protein Resource (UniProt), a standardized database of protein details, including literature citations. While all of UniProt contains over 200 million protein entries, we focus on a subset of UniProt called SwissProt, a manually curated selection of just over 570,000 proteins that are particularly biologically relevant. There is a variety of criteria used to determine which proteins are included in SwissProt. It includes almost all proteins from humans and other key model organisms or major pathogens.[12] It also includes proteins known to be relevant to medicine and disease, as well as proteins with strong experimental evidence on their function or activity, structure, subcellular location, or other traits.

An advantage of using the literature citations curated by SwissProt (as opposed to trying to match against all publications using protein keywords) is that the citations are extremely high-quality and take advantage of curator expertise. Trained biologists read the literature and screen papers based on several criteria: Is the protein the central focus of the paper?; Is the evidence provided in the paper strong?; Is the paper redundant?; Is the paper a review article or comment? Only a small share of papers that might be flagged by keywords meet these standards (Poux et al., 2017). Thus, these references reflect meaningful new knowledge. In addition, SwissProt pulls some citations directly from other annotated databases.[13] Importantly, SwissProt also provides a paper categorization, that describes the content of the paper in distinct categories, including function, process, disease, sequence, family, etc. One of the categories is structure, which typically denotes a link between the paper and a PDB deposit. Since we are interested in studying the effect of AlphaFold on research beyond experimental structural biology, we drop all structure papers.

An example. For years, biochemist Andrea Pauli had been trying to understand how sperm and egg cells fuse, working with zebrafish.[14] Her lab had found a protein on the surface of zebrafish egg cells known as Bouncer and showed it was essential for fertilization. But they still did not understand how Bouncer recognized sperm cells. AlphaFold revealed the structure of a different protein, known as Tmem81, whose role in reproduction had not been known and whose structure had previously been unsolved. This allowed Pauli and team to hypothesize that 3 sperm proteins—including Tmem81—formed a complex that recognized Bouncer. They conducted additional experiments to validate this hypothesis,

[12] For example, it includes the full proteome of mice, Arabidopsis, yeast, E. coli, HIV, and influenza, among others.

[13] Including the PDB, RefSeqs, Mouse Genome Informatics, and others.

[14] This example is taken from a five-year retrospective of AlphaFold's impact published in *Nature*, (Callaway, 2025).

and published their findings in *Cell* (Deneke et al., 2024).

Deneke et al. (2024) is one of the papers in the SwissProt curated literature. The paper has been manually linked by expert curators to eight different SwissProt proteins, including Tmem81 and Bouncer. Thus in our data, each of these eight proteins gets a new linked publication. Alternatively, if we do not want to count the same paper multiple times, we can assign each linked protein a *fraction* of the paper, splitting the paper evenly among all linked proteins. In this example, since Deneke et al. (2024) is linked to eight proteins, each protein would get 1/8th of the paper.

3.4 OpenAlex

OpenAlex is an open-source database of scholarly works covering hundreds of millions of academic publications (Priem, Piwowar, and Orr, 2022). We use OpenAlex to supplement our SwissProt curated papers with reference lists. This allows us to see which SwissProt papers are citing core AlphaFold references. We focus on the three references that EMBL suggests users cite when using AlphaFold DB: Jumper et al. (2021), Varadi, Anyango, et al. (2022), and Varadi, Bertoni, et al. (2024).

3.5 ChEMBL

ChEMBL is a public database that organizes evidence about how small molecules interact with protein targets. Its core unit is an activity: a single compound's binding affinity against a particular target. The dataset contains information on over 24 million activities. A single experiment (known as an "assay") will include many activities, as scientists will test a single protein target against many different compounds. These activities have been performed on nearly 18,000 different targets, around 10,000 of which are single protein targets. These single-protein targets can be linked back to SwissProt and the PDB via UniProt IDs.

ChEMBL curators extract and standardize these activities primarily from the medicinal chemistry literature (including publications and patents). As part of this standardization, protein targets are mapped back to UniProt IDs. It is important to note that much of this activity occurs inside of pharmaceutical firms and is never published; thus it will be missing from the ChEMBL data which is mostly curated from the scientific literature.

An example. ChEMBL target CHEMBL228 is known as the sodium-dependent serotonin transporter. This is an important human protein that is targeted by many antidepressants. Because of its medical importance, this target has seen a lot of research: a total of 963 different assays have been run on this target. Since each assay tests multiple small molecules,

a total of 16,113 different activities have been run on on this target as part of these 963 assays. ChEMBL also includes its UniProt ID, P31645, and this protein in part of the SwissProt subset of UniProt. Moreover, this structure exists in the PDB: it was first released in 2016 (PDB ID 5I6Z).

3.6 Data construction

Experimental structure data

Our goal is to build a structure-level dataset that will measure the rate of research in experimental structure determination. We start with the universe of PDB structure deposits from 1972 to March 2025—a total of 234,092 structures. We then make a series of restrictions.

First, we drop duplicate structures that are part of group deposits. Occasionally, researchers will deposit hundreds (or even thousands) of near-identical structures by the same team on the same day.[15] Because these submissions do not represent a proportional increase in distinct scientific effort, we retain one deposit in these groups. In some cases, these group deposits are explicitly marked, making this easy. In other cases, we infer group deposits if the protein structures are (1) solved by the identical authors; (2) deposited on the identical day; and (3) have the identical amino acid sequence. Dropping these duplicates leaves us with 171,605 structures.

Second, we drop structures from the SARS-CoV-2 (Covid-19) virus, in an effort to normalize activity around the pandemic. This only impacts 2,713 structures, leaving us with 168,892 structures. Third, we drop non-protein structures (primarily DNA and RNA structures). This leaves us with 164,125 structures.

We focus on structures that were deposited between 2017 and the first quarter of 2024. In this time frame, we have a total of 61,638 structures. We drop the last 12 months of our data, which runs through March 2025. We do this because of the 12-month release lag for experimental structures. Between April 2024 and March 2025, additional structures will be deposited but unreleased (and thus invisible to us), leading us to under-count structures in that time period. Table 1 presents the summary statistics for our analysis sample.

Spillover panel

Our goal is to build a panel dataset that tracks additional research, beyond experimental structure determination, using data from the SwissProt curated literature and ChEMBL.

[15] Research teams may do this for a few reasons. One example is a lab may solve the same protein repeatedly with different bound compounds. Each complex is scientifically useful and gets its own PDB deposits, but much of the experimental setup is shared across deposits.

We start with the list of 570,829 SwissProt-indexed proteins. We drop any proteins that do not have a predicted structure in AlphaFold DB. This includes proteins with fewer than 15 amino acids, proteins with more than 2700 amino acids, and viral proteins, as AlphaFold did not predict these for a variety of reasons.[16] We are left with a sample of 546,646 proteins.

We then attach the SwissProt curated publications to their associated proteins. We drop any publications that are categorized as "structure" publications, as these typically correspond to a PDB deposit, and we are aiming to measure research that occurs beyond structure determination. For every remaining paper, we observe the protein it is linked to and its publication year. Some papers are linked to multiple proteins. We handle this in different ways. One way is to count every paper-protein link. Another is to assign fractional shares of papers to proteins. If a paper is linked to multiple proteins, each protein gets an equal fraction of that paper. We count the number of papers (and fractional papers) linked to each protein from 2017 to 2024, creating a panel. We supplement this with data from OpenAlex, flagging which SwissProt publications cite AlphaFold in their references.

Next, we link ChEMBL activities and assays by their protein target, using the UniProt ID. Of the 10,724 unique single-protein targets, 9,478 are in the SwissProt sample. Activities and assays are dated by the publication date of the paper they are sourced from (or in a smaller share of cases, the date of the patent application they are sourced from). We count the number of activities and assays linked to each SwissProt protein from 2017 to 2024.

Finally, we link data from the full PDB, again using UniProt IDs. This allows us to measure whether a given protein had an experimental structure that was publicly released prior to AlphaFold.

Table 2 presents the summary statistics for this panel. We end our panel in 2024 because for both of these curated datasets, the curation takes time. Thus, despite downloading these datasets in late 2025, the data is sparse (or non-existent) in 2025. See Appendix Figure 1 and Appendix Figure 2. Due to the size of the SwissProt sample, most of our measures are fairly sparse. They are also very skewed. Only 14% of proteins were linked to any non-structure paper from 2017 to 2024. The mean paper count is 1.6 with a standard deviation of 16.3. The mean fractional paper count is 0.6 with a standard deviation of 9.2. If we only count papers that link to a single protein, the mean is 0.4 with a standard deviation of 6.1. This is more extreme for the ChEMBL outcomes: only 1% of SwissProt proteins are linked to any ChEMBL activity. Despite this, the mean number of activities from 2017 to 2024 is 5.3, with a standard deviation of 246. In other words, the vast majority of hit discovery is being done on a very small share of proteins.

[16] Large proteins were omitted for computational reasons, and viral proteins were omitted for safety reasons. Note that this implies that we are dropping any Covid-19 proteins.

4 Empirical strategy and results

4.1 Structural biology

Our first set of results focus on how AlphaFold has impacted the field of structural biology. Has AlphaFold served as a substitute for experimental structure determination, or as a complement to it?

Evidence of substitution

Figure 3 shows simple count statistics over time of experimental work. Panel (a) shows the count of experimentally determined structures in the PDB. We index proteins by their deposit date, which is the date they were uploaded to the PDB (but not publicly released). We drop the last 12 months of data, since authors have up to a year to publicly release their deposits. We see no evidence that the rate of experimental structure solving has slowed post-AlphaFold. Researchers appear to be depositing structures at an indistinguishable (if not slightly *higher*) rate after July 2021. The results are nearly identical if we add back in Covid-19 structures (see Appendix Figure 3). If we plot all protein structures with no sample restrictions, the results are qualitatively similar, though slightly noisier due to the large group deposits (see Appendix Figure 4).

Moreover, it is not the case that scientists are merely finishing up work that they started prior to AlphaFold's release. When we investigate collection dates—the date that scientists collected their experimental data—we find in Appendix Figure 5 that over 60% of deposits near the end of our sample window had collection dates after AlphaFold's release. This implies that researchers are continuing to start new experimental projects.

There also appears to be continued interest in these experimentally-solved structures. Panels (b) and (c) show that these structures continue to publish in journals and in "top journals" (defined as *Cell*, *Nature*, and *Science*) at similar rates before and after AlphaFold. We would not expect to see this if the rest of the scientific community was no longer interested in experimentally-determined protein structures.

Perhaps scientists are still solving a similar number of experimental structures, but are shifting the types of structures that they work on. In particular, we might expect researchers to focus on structures where they have a comparative advantage relative to AlphaFold. The prediction confidence scores that AlphaFold assigns are a convenient measure for testing this theory. We can look at the confidence scores assigned to the predicted analogs of the experimental structures that scientists are solving and depositing. If researchers pivot toward structures that AlphaFold is less confident in, this would show up as lower average

confidence scores among experimental structures. However, Figure 4 suggests that this is not the case: the average predicted confidence of experimentally determined structures remains fairly constant throughout our sample period. Comparing the pre- and post-AlphaFold mean suggests confidence scores dropped by one point, but given that the standard deviation in confidence scores in the PDB sample is over ten, this is a very small effect. Thus, we find limited evidence of even this narrower form of substitution in the aggregate.[17]

Why are scientists continuing to carry out expensive and time-consuming experiments? There are several possibilities. One is simply that structural biologists do not want to give up doing the work that they are highly skilled in performing. This would suggest that the additional value of these new experimental structures is low, above and beyond their predicted counterparts. The fact that these experimental structures continue to publish (and in many cases, publish well), however, provides some suggestive evidence against this explanation. Still, it is important to keep these researchers' incentives in mind—they may continue to produce experimental structures past the point that they are useful.

Another possibility, argued persuasively by many structural biologists, is that while predictions are useful, they are not perfect substitutes for experimental structures. One reason relates to the accuracy of the experimental structures. While highly accurate on average, Terwilliger et al. (2024) found in a careful comparison of around 100 experimental versus predicted structures that some amino acids can be misplaced, even when the confidence scores for that amino acid are very high—they estimate that about 10% of "very confidently" placed amino acids are in fact meaningfully misplaced. Another issue raised by the authors is that AlphaFold typically predicts proteins in their "default" state. Terwilliger et al. (2024) argues that since proteins are flexible and dynamic, this may miss a lot of the interesting information. In some cases, the most important question is not "what shape can this protein take?" but "what shape does it take in this situation?" Experiments can be designed to answer that situation-specific question. Researchers can solve a structure while the protein is bound to a particular drug-like molecule, partner protein, DNA/RNA, metal ion, or membrane-like environment, and under particular chemical conditions (like different salt levels or acidity). Those choices can push the protein into the exact shape that matters for its job in the cell. Ultimately, it may be the case that experimental and predicted structures simply offer slightly *different* information, and both pieces are useful or complementary, reminiscent of the examples provided in Autor (2015), such as the complementarity between bank tellers and ATMs.

[17] This is in contrast to Qian (2026), who finds that *within* narrow protein classes, structural biologists do meaningfully shift away from very well-predicted proteins (> 90 average pLDDT scores).

Evidence of complementarity

If AlphaFold is not serving as a substitute for experimentally-determined structures, is it complementing this work? There have been several accounts of predicted structures enabling researchers to solve experimental structures that had previously eluded them (Kryshtafovych, Moult, et al., 2021). To probe this question, we need to introduce two new scientific concepts: molecular replacement and protein sequence homology.

Molecular replacement. The most common experimental technique in our sample is called x-ray crystallography, with nearly 75% of the structures in our sample using this technique. As outlined in Hill and Stein (2025a), this technique broadly consists of three steps: first researchers crystallize proteins. Second, they take the crystals to specialized synchrotron facilities and beam them with x-rays, generating experimental data known as a "diffraction pattern." Third, they use the experimental data to reverse-engineer the structure that generated it.

This third step can be performed in a variety of ways, but the most common method is known as molecular replacement (MR). Almost 90% of experimental x-ray structures in our sample employ this technique. The core idea behind MR is that it uses a similar known protein structure as the starting point for model building. As discussed by Kim (2025), this makes structure solving by MR easier and faster than by other techniques, especially since the development of specialized software in the early 2000s (McCoy et al., 2007). However, MR can only be employed if a similar structure already exists.

Prior to 2021 and the introduction of AlphaFold, this meant that another similar protein must have been solved experimentally in order for researchers to use MR. However, after the development of AlphaFold and the mass deposition of predicted structures in July 2021, scientists quickly realized that *predicted* structures could also be used as the starting structures in MR (Akdel et al., 2022; Kryshtafovych, Moult, et al., 2021).

Defining proteins with sequence homologs. What does it mean for an unsolved protein to have a "similar enough" experimental structure? The general rule is that if a protein shares at least 30% of its amino acid sequence with another protein, it is a good candidate for MR (Phenix, n.d.; Kim, 2025). Thus, for every protein in our sample, we perform the following computations:

1. We find the pool of all experimentally-solved proteins in the PDB that were released prior to the focal protein's deposit date. These proteins go all the way back to the 1970s and represent all possible (public) structures that a researcher could have used as their starting structure.

2. We compare the focal protein's amino acid sequence to that of every protein in the pool, and calculate the sequence similarity. We are able to do this using specially designed software for this task known as MMseqs2 (Steinegger and Söding, 2017).

3. We find the focal protein's nearest neighbor—the protein in the pool with the highest sequence similarity. This can range from anything between 0 and 100, as shown in Table 1. We call the percent similarity between these two proteins the sequence homology score. Hereafter, we will shorten this to "homology score."

We use this continuous homology score and an indicator for "has homolog" (which equals one if the homology score is greater than or equal to 30) in our subsequent analysis.

Testing for complementarity. We begin by investigating the relationship between homology score and the use of MR, before and after AlphaFold. We restrict to structures solved via x-ray crystallography, since MR is only possible for these structures. Prior to AlphaFold's introduction, we expect to see an increasing probability of using MR as the homology score increases. However, if AlphaFold truly does expand the set of proteins that are amenable to MR, we would expect to see this relationship flatten in the post period.

To test this, we split the PDB sample into five roughly evenly-sized groups based on homology scores (see Appendix Figure 6 for a histogram of the homology scores, noting that there is significant mass at 0 and 100). The first group contains all proteins with a homology score of exactly zero. The second group contains all proteins with a score between 0 and 33, the third all proteins with a score between 33 and 66, and the fourth all proteins with a score between 66 and 100. The final group contains all proteins with a score of exactly 100. We then define a *Post* indicator, which equals one if the protein was deposited after July 22, 2021, the date that the first wave of AlphaFold predicted structures were publicly released. We index structure deposits by i and homology groups by $g \in \mathcal{G} = \{0, 33, 66, 99, 100\}$ and estimate:

$$MR_i = \sum_{g \in \mathcal{G}} \beta_g \cdot D_{ig} + \sum_{g \in \mathcal{G}} \gamma_g \cdot D_{ig} \cdot Post_i + \varepsilon_i \tag{1}$$

where D_{ig} is an indicator equaling one if protein i is in group g.

Figure 5 presents the results. The blue circles show the trend in the pre-period, plotting β_g for each group. For structures with a homology score of zero, we see just under 50% of them using MR. This rises steeply as homology increases: in the $(0, 33]$ group this rises to over 70%, and increases to 90% in subsequent groups.

The red triangles plot the sum of $(\beta_g + \gamma_g)$ coefficients and the 95% confidence interval of the difference between the two series. Across all homology bins, MR becomes more

common in the post-period. The increase is largest at low levels of sequence similarity—precisely where pre-period usage was relatively low. In the zero-homology group, MR usage rises by 21 percentage points. In the $(0, 33]$ group, it increases by 12 percentage points. In contrast, for high-similarity structures—where MR was already used in over 90% of cases pre-AlphaFold—increases are mechanically constrained and range from five to seven percentage points. Overall, the relationship between homology and MR usage is substantially flatter in the post-period compared to the pre-period, suggesting that AlphaFold complements experimental structure determination for low-homology proteins by making them amenable to MR.

To further hone in on this theory, we perform two additional tests that take advantage of the sharp timing of AlphaFold's release. We begin by dividing the sample into two groups: "structures with homolog," defined as experimental structures where the nearest neighbor had at least 30% sequence similarity, and "structures without homolog," defined as the converse. Then, we study how the two groups evolve over time. Letting i again index structure deposits and q index quarter of deposit, we estimate:

$$Y_i = \alpha + \lambda Homolog_i + \sum_{q \neq 2021Q2} \delta_q \cdot D_{iq} + \sum_{q \neq 2021Q2} \theta_q \cdot D_{iq} \cdot Homolog_i + \varepsilon_i \tag{2}$$

where $Homolog_i$ is an indicator for whether structure deposition i has a homolog, and D_{iq} is an indicator for whether i was deposited in quarter q.

In Figure 6, we let the outcome be the use of MR, and restrict to structures that were solved using x-ray crystallography. Panel (a) plots the rates of MR usage by the two groups (equivalent to $\alpha + \delta_q$ for structures without a homolog, and $\alpha + \lambda + \delta_q + \theta_q$ for structures with a homolog). Panel (b) plots the difference $(\lambda + \theta_q)$ and the 95% confidence interval. We can see that the gap in MR usage between structures with and without a homolog only begins to close after the arrival of AlphaFold. At baseline, the difference is about 40 percentage points. By the end of our sample, it has fallen below ten percentage points. The timing suggests that AlphaFold is the causal mechanism behind the closing of the MR gap.

However, our data lets us probe this even more closely. When researchers use MR, they are encouraged (though not required) to report their starting structure and its source in the PDB. About 90% of MR structures in our sample reported a starting model. Prior to AlphaFold, over 98% of reported starting models came from the PDB. However, post-AlphaFold, about 9% of structures reported using an AlphaFold starting structure.

Figure 7 estimates Equation 2 using an indicator for an AlphaFold starting structure as the outcome. We restrict to structures solved via x-ray crystallography, using molecular replacement, and citing any starting structure. Mechanically, in the pre-period, neither

group cites any AlphaFold structures.[18] However, in the post period we rapidly see that researchers are using AlphaFold predicted structures as their starting models. This use is far more concentrated among structures without homologs. By the end of our sample period, structures without experimental homologs are using AlphaFold predicted structures as their starting models over 50% of the time, compared to 10% for structures with experimental homologs. This is consistent with AlphaFold serving as a complement to continued experimental structure determination, especially for structures that are more difficult to solve due to their novelty. [19]

4.2 Spillovers to related fields of basic research

Empirical strategy

We now consider the broader impact of AlphaFold on related areas of science and downstream applied R&D. Here we exploit the fact that AlphaFold predictions represented a shock of structural information about some—but not all—proteins. By mid-2021 when AlphaFold predictions were first posted and the model was released publicly, some proteins had already been solved by experimental methods. Among the full set of roughly 570,000 SwissProt indexed proteins, about 8% had an experimentally-solved structure in the PDB, and 40% had the experimentally-solved structure of a close homolog. AlphaFold therefore represented a sudden and unexpected endowment of new structural knowledge for unsolved proteins relative to solved proteins. Figure 8 illustrates this logic. We show two examples. Prior to 2021, UniProt ID P69905 had an experimental structure already deposited in the PDB. However, UniProt ID Q9NPJ8 did not. After 2021, both of these proteins had a predicted structure. For P69905, the experimental structure was already available, so the appearance of an AlphaFold prediction likely did not add very much additional insight. However, in the case of Q9NPJ8, AlphaFold represents a meaningful increase in information about this protein's structure. Thus, we treat the previously-solved structures as the "control" group, and the previously-unsolved structures as the "treatment" group.[20]

Because some structures in SwissProt are nearly identical, we take this into account

[18] The tiny uptick in using an AlphaFold predicted structure in Q2 of 2021 comes from four experimental structures, and likely represents either a typo or private access to AlphaFold predictions.

[19] The scientific literature (e.g., Akdel et al. (2022)) suggests that a similar phenomenon happens for structures that are determined using cryo-electron microscopy (the other major experimental technique apart from x-ray crystallography, comprising 22% of our sample), through a process called docking. However, this is harder to trace in the data.

[20] This empirical strategy is similar in spirit to Hornbeck (2010), though the two papers study entirely different settings. In both contexts, there is near-universal adoption of a new low-cost technology (protein structure prediction; barbed wire fencing), and treatment intensity is determined by access to a pre-existing substitute (experimental protein structures; local woodland endowment that made wooden fencing viable).

when classifying proteins as either solved or unsolved. Proteins in SwissProt are assigned to mutually-exclusive clusters based on their amino acid similarity, again using the MMseqs2 algorithm. If a structure in SwissProt has not been solved, but a structure that is in the same 100% similarity cluster *has* been solved, we also call the first structure solved. This way, we are not calling structures unsolved if they have a near-identical neighbor with an experimental structure, since researchers could use that related structure.[21]

This information shock provided by predicted structures might be useful for scientists working on questions related to these unsolved proteins. Using the protein-linked literature section of UniProt, we test whether there was a change in research intensity among these previously unsolved proteins that have new AI-predicted structures, relative to those proteins that already had an experimental structure.

We note that these two groups of proteins are observably different. Table 3 compares pre-period characteristics of SwissProt proteins that were solved vs. unsolved and shows clear differences. The first three rows show that previously solved structures have 9-18 times as many papers on average published about them (excluding structure papers) in the pre-period depending on the paper count measure. They are nearly three times as likely to have at least one paper. ChEMBL activity is even more skewed, with solved structures having over 60 times the activity. This imbalance on pre-period outcomes is not surprising if research clusters on the most biologically relevant proteins.

The stark difference in baseline levels also motivates our choice of a proportional (Poisson pseudo-maximum likelihood) specification over a linear model. In a linear difference-in-differences model, parallel trends requires that the two groups would have added similar numbers of papers in the absence of AlphaFold—an implausible assumption when one group begins with over ten times the research activity of the other. Our Poisson specification instead requires that the two groups would have grown at similar rates absent the treatment, which is considerably more credible in our setting. We also note that AlphaFold prediction confidence is nearly identical between both sets of proteins. This aligns with the fact that AlphaFold performs extremely well out of sample. This is important for our research design, because it implies that our results are not being driven by differentially useful predictions.

We compare how publication rates evolve before and after the introduction of AlphaFold differentially for experimentally unsolved and solved proteins. Our regression sample is a panel of 546,646 SwissProt-indexed proteins in years 2017 through 2024. We estimate a Poisson difference-in-differences regression for protein i in year t:

[21] In the appendix, we test robustness to different definitions based on broader categories of similarity at the 30%, 50%, and 90% level.

$$\mathbb{E}[Y_{it}|Unsolved_i, t] = \exp\big(\alpha + \lambda Unsolved_i + \sum_{t\neq 2020} \tau_t \cdot D_t + \sum_{t\neq 2020} \kappa_t \cdot D_t \cdot Unsolved_i\big) \quad (3)$$

where Y_{it} in this case is defined as a count of all non-structure papers. $Unsolved_i$ is an indicator for whether the protein had an experimentally-solved structure model in the PDB prior to AlphaFold's release in July 2021.[22] D_t is an indicator for whether the observation occurred in year t. Standard errors are clustered at the 100% protein cluster level.

Among non-structure papers linked to proteins in UniProt, 43% are linked to more than one protein, and 9% are linked to five or more proteins. To handle this non-unique mapping between papers and proteins, we employ three main publication outcome variables: First, we count all papers linked to a particular protein, allowing the same paper to be counted multiple times. This is an attractive measure if we think that papers linked to many proteins represent large contributions of knowledge in many parts of the protein space. On the other hand, papers focused on fewer proteins might have more insight-per-protein. Therefore, our second measure is fractional papers, where we divide each paper by the number of individual proteins it is linked to.[23] Lastly, in order to focus on cases with a one-to-one mapping, we count papers linked to exactly one protein.

Results

Figure 9 plots the year by treatment interaction coefficients (κ_t's) with standard errors from Equation 3. Each of the three panels tests a different publication outcome as defined above. First, we notice a relatively consistent pre-trend near zero for each outcome, suggesting that publication activity was evolving on a similar trend before treatment. The standard difference-in-differences assumption of parallel trends assumes that this pattern would have continued through the post-period absent the appearance of AlphaFold. Instead, we find a statistically significant increase in publications about unsolved proteins after AlphaFold, suggesting a shift in attention when AI-predicted structures become available. This marked

[22] One subtlety is that we want to assign proteins to *stable* solved vs. unsolved categories, but a protein's solved vs. unsolved status evolves over time. For example, a protein that we code as "solved" as of 2021 may have first been solved in 2019. Thus, for part of a pre-period, this "solved" protein was actually unsolved. In practice, this distinction does not matter much because of the slow pace of experimental research: fewer than 1.5% of proteins in our sample are solved for the first time between 2017 and 2021. Moreover, the results are virtually identical if we define solved vs. unsolved in 2017, at the start of our sample window.

[23] Note that this outcome will take non-integer values. Our Poisson difference-in-differences estimates follow the pseudo-maximum likelihood (PPML) approach of Santos Silva and Tenreyro (2006). Consistency of the Poisson estimator requires only that the conditional mean is correctly specified; it does not require that outcomes be integer-valued.

shift toward research on previously unsolved proteins is apparent across all three publication definitions.

To get a more precise sense of magnitudes, we also report static Poisson difference-in-differences estimates in Table 4. The estimates across Columns 1-3 for each outcome suggest that, in the post-period, research on previously unsolved proteins rose by 16 to 25% relative to previously solved proteins.[24] The results are statistically significant at the 1% level. These estimates are even larger if we focus on the coefficients from the end of our dynamic specification, suggesting 35 to 40% increases by 2024. These results suggest that AlphaFold stimulated significant new basic research about proteins for which we previously lacked an experimental structure model.

We take advantage of the category tags provided for every SwissProt curated paper to analyze how AlphaFold may have impacted a variety of subfields of protein science. We focus on the six largest paper categories: Function (51% of papers), Expression (13%), Phenotypes (8%), Sequences (8%), Disease (8%), and Interaction (4%). Each of these categories represent a major strain of protein research, covering research questions about how proteins function in cells, how they are expressed from DNA, how they affect disease, how they interact with each other, and more. Figure 10 estimates Equation 3, replacing the fractional publication count with specific fractional counts of papers in each category. We find statistically significant shifts in activity toward previously unsolved proteins in five of the six paper categories. Besides phenotype research, all categories have shifts toward unsolved proteins after AlphaFold. The largest category—function papers—reacts most quickly, albeit with a modest magnitude in percentage terms (a 16% increase relative to solved structures). Expression, sequences, disease, and interaction categories increase later starting around 2023, but the percentage shift is large in 2024 (60 to 100%). These results suggest that AlphaFold is having broad impacts on a variety of basic research topics in protein science, with multiple scientific specialties benefiting from the shock of new structural insights.

Analysis of AlphaFold citation rates

One potential threat to our identification strategy would be if an unrelated technology or research demand shock occurred around 2021, leading us to misattribute the treatment effect to AlphaFold instead of the other unobserved but correlated shocks. We can directly test the role of AlphaFold in driving this shift toward unsolved protein by tracking citations to core AlphaFold papers. On their website, EMBL strongly encourages any scientists using the AlphaFold model or referencing structure predictions on the AlphaFold DB to cite three

[24] For example, using the coefficient from Column (2), we compute the percent change by taking $100 \times (e^{0.201} - 1) \approx 22\%$

relevant papers: Jumper et al., 2021; Varadi, Anyango, et al., 2022; and Varadi, Bertoni, et al., 2024.[25] Although these citations are recommended, we recognize that they may be under-provided by authors either for strategic reasons or inattention. Nevertheless, we can compare the citation rates to these three papers after the AlphaFold release separately for papers about solved vs. unsolved proteins.

To implement this, we use the references to SwissProt publications linked from OpenAlex. We then split the papers into a solved or unsolved group based on the proteins linked to the paper. If the paper is linked to a mix of solved and unsolved proteins, we categorize based on the status of the majority of proteins in the paper. Figure 11 shows the share of papers that cite at least one of the three AlphaFold papers separately for both groups. Panel (a) shows the raw counts and panel (b) estimates the difference using a similar specification to Equation 2.

We first notice that by 2024, about 2% of all protein papers cite one of the three AlphaFold papers. On one hand, almost no academic papers reach the level of influence so as to be cited by 2% of papers across a massive swath of science.[26] On the other hand, 2% is a relatively modest share given the massive impact AlphaFold is having in this sphere. This might reflect low adherence to the citation norm encouraged by EMBL. Putting aside the base citation rate, the key fact that we highlight is the stark difference in citation rates between papers about solved and unsolved proteins. By 2024, the unsolved group is citing the core AlphaFold papers at a rate about 50% higher than the solved group. This result underscores that AlphaFold is likely the primary driver in the shift of direction in basic research toward previously unsolved proteins.

4.3 Robustness and alternative specifications

Investigating the Covid-19 pandemic

One potential confounder in this analysis is the intense shift in attention toward the SARS-CoV-2 proteins during the Covid-19 pandemic (Hill, Yin, et al., 2025). SwissProt includes 17 proteins that have SARS-CoV-2 listed as their organism. Although these proteins are a tiny fraction of SwissProt, they are linked to about 1,100 individual papers, about half a percent of all papers published about proteins since 2020. DeepMind and EMBL decided to not post AlphaFold predictions of proteins found in viruses for safety reasons, so we drop these by default from the main analysis. Nonetheless, Appendix Figure 7 shows that including these

[25] See https://www.ebi.ac.uk/training/online/courses/alphafold/accessing-and-predicting-protein-structures-with-alphafold/how-to-cite-alphafold/

[26] Jumper et al., 2021 especially has more than 47,000 citations on Google Scholar, though many of them from computer scientists and computational biologists working on similar AI algorithms.

Covid-relevant papers and proteins does not noticeably impact our results.

Alternative definitions of solved vs. unsolved proteins

We also show robustness to alternative definitions of solved and unsolved in 2021. Our main specification uses a relatively narrow 100% sequence similarity definition to tag proteins that had a structure solution at the time of AlphaFold release. We could instead define a protein as solved as long as a 90% similar protein has been solved, using amino acid sequence similarity. We repeat for different levels of similarity (50% and 30%). Using a broader definition increases the share of proteins that are considered solved. We don't take a strong stand on the correct breadth to define solved, which depends in practice on how exactly scientists are using these structures. In Appendix Table 1, we show that all three publication outcomes are robust to broader definitions of unsolved. The estimated treatment effects are slightly lower for the broader categories (7 to 17% for the 30% similarity definition) but still statistically significant at the 1% level.

Alternative strategy: Proteins that remain unpredicted

As discussed in Section 3, not all proteins in SwissProt have a corresponding predicted structure in AlphaFold. AlphaFold imposed size restrictions, not predicting any protein structures with fewer than 15 amino acids or more than 2700 amino acids. Thus, an alternative empirical strategy is to focus on proteins that were unsolved prior to AlphaFold, and then compare downstream research activity for proteins that do receive a predicted structure versus those that do not. Appendix Figure 8 provides a visual example of this strategy.

We implement this strategy using the 15 amino acid cutoff because there is a dense mass of proteins in SwissProt around this cutoff (see Appendix Figure 9). To make the proteins on either side of the cutoff more comparable, we focus only on proteins with 10 to 50 amino acids. Appendix C provides more details on the implementation of this strategy. Appendix Figure 10 shows the results of Equation 3 where the longer proteins—the proteins that receive a predicted structure—are the treated group. We find large effects with a flat pre-trend, though with very wide standard errors due to the smaller sample. If we average the Poisson coefficients in the post-period, we estimate that research on predicted proteins rose by about 14 times relative to unpredicted proteins.[27] However, the confidence intervals imply an extremely wide range. Focusing only the 2024 coefficient, the confidence interval implies a relative increase of anywhere from 2 times to 200 times. Thus, we interpret the magnitudes extremely cautiously.

[27] Computed by taking the average Poisson coefficient in 2022, 2023, and 2024, which is 2.7. $e^{2.7} - 1 \approx 14$.

Still, this is much larger estimate than we find in our primary empirical strategy, though we hesitate to interpret the difference. For one, it is a fundamentally different comparison. The information shocks in the two strategies are different, and the coefficients have different interpretations. For another, it is a local treatment effect estimated on less than 2% of the data. However, the fact that both strategies estimate positive effects on follow-on research suggests that AlphaFold predictions did causally spur more research on structures where they provided new information.

4.4 Spillovers to downstream pharmaceutical R&D

AlphaFold provides low-cost structure predictions that may also be useful in downstream structure-based drug design. We test this by again comparing research activity about proteins that had previously been solved experimentally to those that had not. Although there are many potential downstream outcomes to focus on in the drug design pipeline, we focus on bioactivity assays, which are an early step in understanding the interaction between target proteins and candidate drugs (see Figure 2). We count "activity" entries in the ChEMBL database and assign them based on their protein target to our SwissProt panel. We then estimate the same Poisson difference-in-differences specification as described in Equation 3, but replace publication counts with ChEMBL activity counts. Figure 12 reports the coefficients for ChEMBL activities and shows no significant increase in attention toward previously unsolved proteins, though the confidence intervals are quite wide. Column 4 of Table 4 reports static Poisson difference-in-differences estimates for ChEMBL activities and we similarly find a statistically insignificant difference in pre-post activity rates for solved and unsolved proteins.

4.5 Discussion

What should we make of the contrasting results between basic science and early-stage R&D? One possibility is that while AlphaFold illuminated a massive number of structures, most of the proteins that represent druggable targets had already been experimentally solved. As a collective community, structural biologists spent five decades slowly tackling the structures of the most disease-relevant proteins. Although they had not fully completed the determination of the entire proteome, they plucked much of the low-hanging fruit that was relevant for many diseases. However, the scientific evidence pushes back against this interpretation. In 2014, the National Institutes of Health launched the Illuminating the Druggable Genome (IDG) Program, explicitly to tackle the fact that pharmaceutical research has clustered on a small set of proteins. The IDG estimates that there are as many as 4,500 proteins susceptible

to targeting by drugs, but only about 15% of these are targeted by FDA-approved drugs (Sharma et al., 2024). Much of the attention remains on proteins that were known prior to the mapping of the human genome in the early 2000s, which likely reflects path-dependence rather than scientific potential (Edwards et al., 2011).

An alternative possibility is that protein structure alone is insufficient to initiate the process of drug discovery. These experimentally unsolved proteins may remain poorly understood, even after the arrival of predicted structures. As Table 3 shows, these papers had far fewer papers written about them prior to AlphaFold. As discussed in Section 2, before researchers will invest in small-molecule binding experiments, a protein's function and disease relevance must be sufficiently established. Establishing this kind of causal disease relevance requires extensive experimental validation in cell and animal models. Historically, structural research and functional or disease research were symbiotic and often proceeded in parallel: evidence implicating a protein in a disease pathway would motivate structural biologists to solve its structure, while a new structure could in turn generate hypotheses about function and mechanism. AlphaFold represented a massive shock of structural information, but without the complementary biological context, new drug targets will not necessarily emerge. Our evidence suggests that this new influx of structural information *is* leading to more basic science on these proteins, and so this biological context may be on its way. For example, Figure 10 shows that basic science about how proteins relate to disease has markedly increased for these newly illuminated proteins. In the short run, however, it may be that this science is playing catch-up. This may imply that the pharmaceutical R&D is coming in the future, but for now is rate-limited by additional human-conducted science that is currently underway. Of course, we caution that all of this remains speculative, and future research will be needed to see how this plays out in subsequent years.

5 Conclusion

AlphaFold offers a clean, early view of how modern AI is already impacting scientific discovery. In the narrow domain where it most plausibly substitutes for humans—experimental structure determination—we find little evidence so far of displacement. Experimental deposits in the PDB and publication outcomes remain stable in the three years after AlphaFold's release, and researchers are continuing to start new experimental work. At the same time, the technology appears to be changing *how* experimentalists work. The rapid expansion of molecular replacement using AlphaFold templates, especially for proteins without close experimental homologs, suggests that humans are using the AI output to make previously challenging experimental work more efficient. Whether or not AI-driven comple-

mentarity will persist as prediction tools improve and expand their capabilities remains to be seen.

Downstream of experimental structure determination, we find a different pattern that might be representative of other scientific settings affected by AI. AlphaFold delivered a broad, low-cost knowledge shock to researchers who consume structural information, and we observe a large relative increase in non-structure publications about previously unsolved proteins in the years after release. AlphaFold seems to have acted as a "floodlight," expanding the set of proteins that scientists are studying.

Yet this apparent broadening of basic research attention does not translate—at least within our current window—into a comparable shift in downstream applied R&D as measured by bioactivity assay activity in ChEMBL. While speculative, this suggests that even when AI dramatically lowers the cost of a key input (here, structural information), downstream progress may remain constrained by slower-moving complements.

In our view, these findings also underscore why the consequences of AI for science remain uncertain. Even as more steps become automated, new constraints will emerge. Thinking further down the drug development pathway, clinical trials remain a formidable bottleneck. While some may point to the possibility of "end-to-end" AI pipelines (Demirer et al., 2026), the timeline and feasibility of these remain uncertain. Thus, in our view, even vast AI-enabled improvements along the basic science to clinical pathway do not guarantee massive gains in drug approval or human health in the short-to-medium run.

References

Acemoglu, Daron and David Autor (2011). "Skills, Tasks and Technologies: Implications for Employment and Earnings". *Handbook of Labor Economics*. Ed. by Orley Ashenfelter and David Card. Vol. 4B. Elsevier. Chap. 12, pp. 1043–1171.

Acemoglu, Daron and Pascual Restrepo (2019). "Automation and new tasks: How technology displaces and reinstates labor". *Journal of Economic Perspectives* 33.2, pp. 3–30.

— (2022). "Tasks, automation, and the rise in US wage inequality". *Econometrica* 90.5, pp. 1973–2016.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). "Combining human expertise with artificial intelligence: Experimental evidence from radiology". *NBER Working Paper*.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones (2017). "Artificial Intelligence and Economic Growth". *NBER Working Paper*.

Agrawal, Ajay, John McHale, and Alexander Oettl (2023). "Superhuman science: How artificial intelligence may impact innovation". *Journal of Evolutionary Economics* 33.5, pp. 1473–1517.

— (2024). "Artificial Intelligence and Scientific Discovery: A Model of Prioritized Search". *Research Policy* 53.5, p. 104989.

Akdel, Mehmet et al. (2022). "A Structural Biology Community Assessment of AlphaFold2 Applications". *Nature Structural & Molecular Biology* 29.11, pp. 1056–1067.

Anfinsen, Christian B. (1973). "Principles that Govern the Folding of Protein Chains". *Science* 181.4096, pp. 223–230.

Autor, David H. (2015). "Why Are There Still So Many Jobs? The History and Future of Workplace Automation". *Journal of Economic Perspectives* 29.3, pp. 3–30.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger (1998). "Computing Inequality: Have Computers Changed the Labor Market?" *The Quarterly Journal of Economics* 113.4, pp. 1169–1213.

Autor, David H., Frank Levy, and Richard J. Murnane (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration". *The Quarterly Journal of Economics* 118.4, pp. 1279–1333.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin (2019). "Does Science Advance One Funeral at a Time?" *American Economic Review* 109.8, pp. 2889–2920.

Ball, Philip (2024). "How AI Revolutionized Protein Science, but Didn't End It". *Quanta Magazine*. Accessed 2026-02-16.

Boustan, Leah Platt, Jiwon Choi, and David Clingingsmith (2024). "Computerized Machine Tools and the Transformation of US Manufacturing". *NBER Working Paper*.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (2025). "Generative AI at work". *The Quarterly Journal of Economics* 140.2, pp. 889–942.

Callaway, Ewen (2020). "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures". *Nature* 588.7837, pp. 203–205.

— (2025). "AlphaFold is five years old — these charts show how it revolutionized science". *Nature* 648.8093. News article; published 26 Nov 2025 (with correction on 27 Nov 2025)., pp. 258–259.

Cavalli, Gabriel (2024). "Responding to Advances in AI: The Impact of AlphaFold on the Organization of Academic Labs". *Working Paper*.

Cockburn, Iain M., Rebecca Henderson, and Scott Stern (2019). "The Impact of Artificial Intelligence on Innovation". *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press.

Cui, Zheyuan Kevin, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz (2025). "The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers". *Working Paper*.

Demirer, Mert, John J Horton, Nicole Immorlica, Brendan Lucier, and Peyman Shahidi (2026). "Chaining Tasks, Redefining Work: A Theory of AI Automation". *NBER Working Paper*.

Deneke, Victoria E. et al. (2024). "A conserved fertilization complex bridges sperm and egg in vertebrates". *Cell* 187.25. Epub 2024-10-17., 7066–7078.e22.

Edich, Maximilian, David C. Briggs, Oliver Kippes, Yunyun Gao, and Andrea Thorn (2022). "The Impact of AlphaFold2 on Experimental Structure Solution". *Faraday Discussions* 240, pp. 184–195.

Edmunds, Nicholas S., Liam J. McGuffin, and Ahmet G. Genc (2024). "Benchmarking of AlphaFold2 accuracy self-estimates as indicators of empirical model quality and ranking: a comparison with independent model quality assessment programmes". *Bioinformatics* 40.8, btae480.

Edwards, Aled M., Ruth Isserlin, Gary D. Bader, Stephen V. Frye, Timothy M. Willson, and Frank H. Yu (2011). "Too many roads not taken". *Nature* 470.7333, pp. 163–165.

Erickson, J., D. J. Neidhart, J. VanDrie, D. J. Kempf, X.-C. Wang, D. W. Norbeck, J. J. Plattner, J. W. Rittenhouse, M. Turon, N. Wideburg, W. Kohlbrenner, R. Simmer, R. Helfrich, D. A. Paul, and M. Knigge (1990). "Design, Activity, and 2.8 Å Crystal Structure of a C2 Symmetric Inhibitor Complexed to HIV-1 Protease". *Science* 249.4968, pp. 527–533.

Feigenbaum, James and Daniel P Gross (2024). "Answering the call of automation: How the labor market adjusted to mechanizing telephone operation". *The Quarterly Journal of Economics* 139.3, pp. 1879–1939.

Goldberg, Samuel and H Tai Lam (2025). "Generative AI in equilibrium: Evidence from a creative goods marketplace". *Working Paper*.

Hill, Ryan and Carolyn Stein (2025a). "Race to the Bottom: Competition and Quality in Science". *Quarterly Journal of Economics* 140.2, pp. 1111–1185.

— (2025b). "Scooped! Estimating Rewards for Priority in Science". *Journal of Political Economy* 133.3, pp. 793–845.

Hill, Ryan, Yian Yin, Carolyn Stein, Xizhao Wang, Dashun Wang, and Benjamin F Jones (2025). "The pivot penalty in research". *Nature* 642.8069, pp. 999–1006.

Hornbeck, Richard (2010). "Barbed Wire: Property Rights and Agricultural Development". *Quarterly Journal of Economics* 125.2, pp. 767–810.

Humlum, Anders (2021). "Robot Adoption and Labor Market Dynamics". *Working Paper*.

Humlum, Anders and Emilie Vestergaard (2026). "Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI". *Working Paper*.

International Human Genome Sequencing Consortium (2004). "Finishing the Euchromatic Sequence of the Human Genome". *Nature* 431.7011, pp. 931–945.

Jinek, Martin, Fuguo Jiang, David W. Taylor, Samuel H. Sternberg, Emine Kaya, Enbo Ma, Carolin Anders, Michael Hauer, Kaihong Zhou, Steven Lin, Matias Kaplan, Anthony T. Iavarone, Emmanuelle Charpentier, Eva Nogales, and Jennifer A. Doudna (2014). "Structures of Cas9 Endonucleases Reveal RNA-mediated Conformational Activation". *Science* 343.6176, p. 1247997.

Jones, Benjamin (2025). "Artificial intelligence in research and development". *NBER Working Paper*.

Jones, Charles I. (1995). "R&D-Based Models of Economic Growth". *Journal of Political Economy* 103.4, pp. 759–784.

— (2005). "Growth and Ideas". *Handbook of Economic Growth*. Ed. by Philippe Aghion and Steven N. Durlauf. Vol. 1. Elsevier. Chap. 16, pp. 1063–1111.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". *Nature* 596.7873, pp. 583–589.

Kim, Soomi (2025). "Navigating the Rugged Data Landscape: The Impact of Data-Extrapolation Technologies on Knowledge Production". *Working Paper*.

Kovalevskiy, Oleg, Juan Mateos-Garcia, and Kathryn Tunyasuvunakool (2024). "AlphaFold two years on: Validation and impact". *Proceedings of the National Academy of Sciences* 121.34, e2315002121.

Kryshtafovych, Andriy, John Moult, Reinhard Albrecht, Geoffrey A. Chang, Kinlin Chao, Alec Fraser, Julia Greenfield, Marcus D. Hartmann, Osnat Herzberg, Inokentijs Josts, Petr G. Leiman, Sara B. Linden, Andrei N. Lupas, Daniel C. Nelson, Steven D. Rees, Xiaoran Shang, Maria L. Sokolova, Henning Tidow, and AlphaFold2 team (2021). "Computational Models in the Service of X-ray and Cryo-electron Microscopy Structure Determination". *Proteins: Structure, Function, and Bioinformatics* 89.12, pp. 1633–1646.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult (2021). "Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round XIV". *Proteins: Structure, Function, and Bioinformatics* 89.12, pp. 1607–1617.

Levinthal, Cyrus (1969). "How to Fold Graciously". *Mössbauer Spectroscopy in Biological Systems*. Ed. by P. Debrunner, J. C. M. Tsibris, and E. Münck. Urbana, IL: University of Illinois Press, pp. 22–24.

Ludwig, Jens and Sendhil Mullainathan (2024). "Machine Learning as a Tool for Hypothesis Generation". *The Quarterly Journal of Economics* 139.2, pp. 751–827.

McCoy, Airlie J., Ralf W. Grosse-Kunstleve, Paul D. Adams, Martyn D. Winn, Laurent C. Storoni, and Randy J. Read (2007). "*Phaser* Crystallographic Software". *Journal of Applied Crystallography* 40.4, pp. 658–674.

Metzker, Michael L. (2010). "Sequencing Technologies — The Next Generation". *Nature Reviews Genetics* 11.1, pp. 31–46.

Moult, John, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano (2014). "Critical assessment of methods of protein structure prediction (CASP)–round x". *Proteins: Structure, Function, and Bioinformatics* 82, pp. 1–6.

Moult, John, Jens T. Pedersen, Robert Judson, and Krzysztof Fidelis (1995). "A large-scale experiment to assess protein structure prediction methods". *Proteins: Structure, Function, and Genetics* 23.3, pp. ii–v.

Mullainathan, Sendhil and Ashesh Rambachan (2025). "Science in the Age of Algorithms". *The Economics of Transformative AI*. NBER.

Nagaraj, Abhishek (2021). "The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry". *Management Science* 68.1, pp. 564–582.

Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan (2020). "Improving Data Access Democratizes and Diversifies Science". *Proceedings of the National Academy of Sciences* 117.38, pp. 23490–23498.

Navia, Manuel A, Paula M D Fitzgerald, Brian M McKeever, Chih-Tai Leu, Jill C Heimbach, Wayne K Herber, Irving S Sigal, Paul L Darke, and James P Springer (1989). "Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1". *Nature* 337.6208, pp. 615–620.

Nogales, Eva (2016). "The Development of Cryo-EM into a Mainstream Structural Biology Technique". *Nature Methods* 13.1, pp. 24–27.

Noy, Shakked and Whitney Zhang (2023). "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence". *Science* 381.6654, pp. 187–192.

Phenix (n.d.). *Molecular Replacement: Overview*. https://phenix-online.org/documentation/reference/mr_overview.html. Phenix documentation page. Accessed 2026-02-21.

Poux, Sylvain, Michele Magrane, Cecilia Arighi, Alan Bridge, Claire O'Donovan, and UniProt Consortium (2017). "On expert curation and scalability: UniProtKB/Swiss-Prot as a case study". *Bioinformatics* 33.21, pp. 3454–3460.

Priem, Jason, Heather Piwowar, and Richard Orr (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". *arXiv*. arXiv: 2205.01833 [cs.DL].

Qian, Jerry (2026). "AI in the Lab: AlphaFold2's Impacts on Human-Produced Knowledge". *Working Paper*.

Restrepo, Pascual (2024). "Automation: Theory, evidence, and outlook". *Annual review of economics* 16.1, pp. 1–25.

Romer, Paul M. (1990). "Endogenous Technological Change". *Journal of Political Economy* 98.5, Part 2, S71–S102.

Santos, Rita, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, and John P Overington (2017). "A comprehensive map of molecular drug targets". *Nature Reviews Drug Discovery* 16.1, pp. 19–34.

Santos Silva, J. M. C. and Silvana Tenreyro (2006). "The Log of Gravity". *The Review of Economics and Statistics* 88.4, pp. 641–658.

Sharma, Karlie R, Christine M Colvis, Griffin P Rodgers, and Douglas M Sheeley (2024). "Illuminating the druggable genome: Pathways to progress". *Drug Discovery Today* 29.3, p. 103805.

Steinegger, Martin and Johannes Söding (2017). "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets". *Nature Biotechnology* 35.11, pp. 1026–1028.

Strasser, Bruno J. (2019). *Collecting Experiments: Making Big Data Biology*. Chicago: University of Chicago Press.

Sullivan, Kevin P., Peggy Brennan-Tonetta, and Lori J. Marxen (2017). "Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank". *RCSB Protein Data Bank* 10.

Terwilliger, Thomas C., Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams (2024). "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination". *Nature Methods* 21.1, pp. 110–116.

The UniProt Consortium (2023). "UniProt: the Universal Protein Knowledgebase in 2023". *Nucleic Acids Research* 51.D1, pp. D523–D531.

Tranchero, Matteo (2024). "Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation". *Working Paper*.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. (2021). "Highly accurate protein structure prediction for the human proteome". *Nature* 596.7873, pp. 590–596.

Varadi, Mihaly, Stephen Anyango, et al. (2022). "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". *Nucleic Acids Research* 50.D1, pp. D439–D444.

Varadi, Mihaly, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. (2024). "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences". *Nucleic Acids Research* 52.D1, pp. D368–D375.

Varadi, Mihaly and Sameer Velankar (2023). "The impact of AlphaFold Protein Structure Database on the fields of life sciences". *Proteomics* 23.17, p. 2200128.

Weitzman, Martin L. (1998). "Recombinant Growth". *The Quarterly Journal of Economics* 113.2, pp. 331–360.

Wlodawer, Alexander, Maria Miller, Mariusz Jaskólski, Bangalore K Sathyanarayana, Eric Baldwin, Irene T Weber, Linda M Selk, Leigh Clawson, Jens Schneider, and Stephen B H Kent (1989). "Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease". *Science* 245.4918, pp. 616–621.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan (2020). "Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation". *Science* 367.6483, pp. 1260–1263.

Yu, Zhengyi (2026). "The Impacts of AI at Scale: Evidence from Research Scientists". *Working Paper*.

Zemla, Adam (2003). "LGA: a method for finding 3D similarities in protein structures". *Nucleic Acids Research* 31.13, pp. 3370–3374.

Zhou, Eric and Dokyun Lee (2024). "Generative artificial intelligence, human creativity, and art". *PNAS Nexus* 3.3, p. 052.

Zhuo, Ran (2025). "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs". *Working Paper*.

Tables and Figures

Table 1: Summary statistics: PDB structures

	Mean	Median	SD	Min	Max	Obs
Panel A. Variables relevant to all structures						
Entity count	2.72	1.00	7.40	1.00	150.00	61,638
Residue count	1710	645	6845	3	566853	61,638
Similarity to nearest neighbor	74.8	95.0	33.1	0.0	100.0	61,638
Share that have experimental homolog	0.87	1.00	0.33	0.00	1.00	61,638
AlphaFold prediction confidence	86.0	89.1	10.5	27.0	98.7	52,277
Experimental determination method						
X-ray crystallography	0.76	1.00	0.43	0.00	1.00	61,638
Cryo electron microscopy	0.21	0.00	0.41	0.00	1.00	61,638
Other	0.03	0.00	0.18	0.00	1.00	61,638
Panel B. Variables relevant to x-ray structures						
Use molecular replacement (MR)	0.87	1.00	0.34	0.00	1.00	46,570
List a starting model	0.80	1.00	0.40	0.00	1.00	46,570
List an AlphaFold starting model	0.02	0.00	0.15	0.00	1.00	46,570

Notes: This table presents summary statistics for the experimental PDB structures in our analysis sample over the 2017 to Q1 2024 time frame. Panel A presents statistics relevant for all 61,638 structures. Panel B presents statistics only relevant for structures solved using x-ray crystallography. The number of observations for AlphaFold prediction confidence is lower because some PDB structures do not have an AlphaFold counterpart.

Table 2: Summary statistics: SwissProt proteins

	Mean	Median	SD	Min	Max	Obs
Unsolved prior to AlphaFold	0.920	1.00	0.27	0.00	1.00	546,646
SwissProt non-structure papers (all paper count)	1.56	0.00	16.28	0.00	2855	546,646
SwissProt non-structure papers (fractions of papers)	0.63	0.00	9.16	0.00	2038.7	546,646
SwissProt non-structure papers (single-protein papers)	0.38	0.00	6.13	0.00	1429	546,646
Any SwissProt non-structure paper	0.14	0.00	0.35	0.00	1.00	546,646
ChEMBL activity count	5.26	0.00	245.76	0.00	56,010	546,646
ChEMBL assay count	0.34	0.00	11.47	0.00	2754	546,646
Any ChEMBL activity/assay	0.01	0.00	0.11	0.00	1.00	546,646
Average pLDDT confidence score	87.51	91.21	10.64	25.97	98.75	546,646

Notes: This table presents summary statistics for the proteins indexed by SwissProt. All SwissProt and ChEMBL variables are counts accrued over the sample period of 2017 to 2024. For fractional papers, if a paper is linked to N proteins, each protein is assigned $1/N$ of the paper. $N = 546{,}646$ SwissProt-indexed proteins. The number of observations for AlphaFold prediction confidence is lower because some SwissProt structures do not have an AlphaFold counterpart.

Table 3: Summary statistics: SwissProt proteins, solved vs. unsolved structures

	Solved structures	Unsolved structures
SwissProt non-structure papers (all paper count)	4.83	0.55
SwissProt non-structure papers (fractions of papers)	2.50	0.17
SwissProt non-structure papers (single-protein papers)	1.59	0.09
Any SwissProt non-structure paper	0.30	0.11
ChEMBL activity count	33.93	0.56
ChEMBL assay count	2.18	0.04
Any ChEMBL activity/assay	0.076	0.004
Average pLDDT confidence score	87.76	87.49
Observations	43,855	502,791

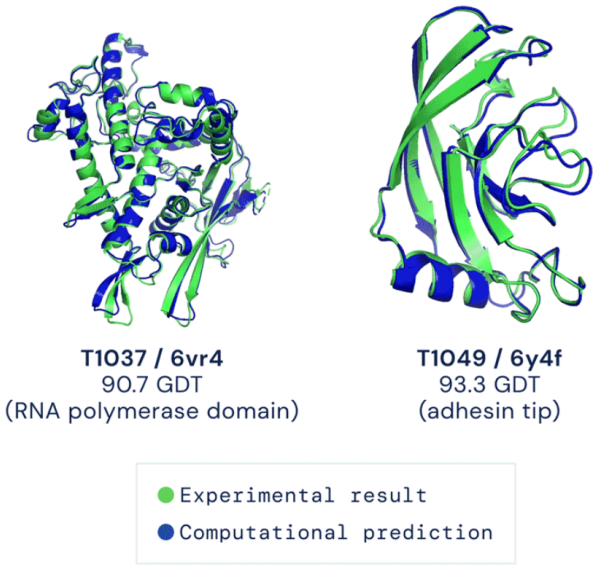
Notes: This table presents means for the proteins indexed by SwissProt, split by whether they have a solved experimental protein structure or not prior to 2021. All variables are counts accrued over the pre-period of 2017 to 2020. For fractional papers, if a paper is linked to N proteins, each protein is assigned $1/N$ of the paper. $N = 546{,}646$ SwissProt-indexed proteins.

Table 4: Difference-in-differences: Papers in related fields and pharmaceutical R&D

	(1)	(2)	(3)	(4)
Dependent Variable:	Non-structure papers (all paper count)	Non-structure papers (fractions of papers)	Non-structure papers (single protein	ChEMBL Activities
Post	-0.489***	-0.497***	-0.481***	-0.583***
	(0.0102)	(0.0126)	(0.0142)	(0.0653)
Unsolved	-2.173***	-2.656***	-2.822***	-4.135***
	(0.0329)	(0.0398)	(0.0425)	(0.1431)
Post x Unsolved	0.150***	0.201***	0.224***	0.021
	(0.0114)	(0.0149)	(0.0177)	(0.1370)
Observations	4,373,168	4,373,168	4,373,168	4,373,168

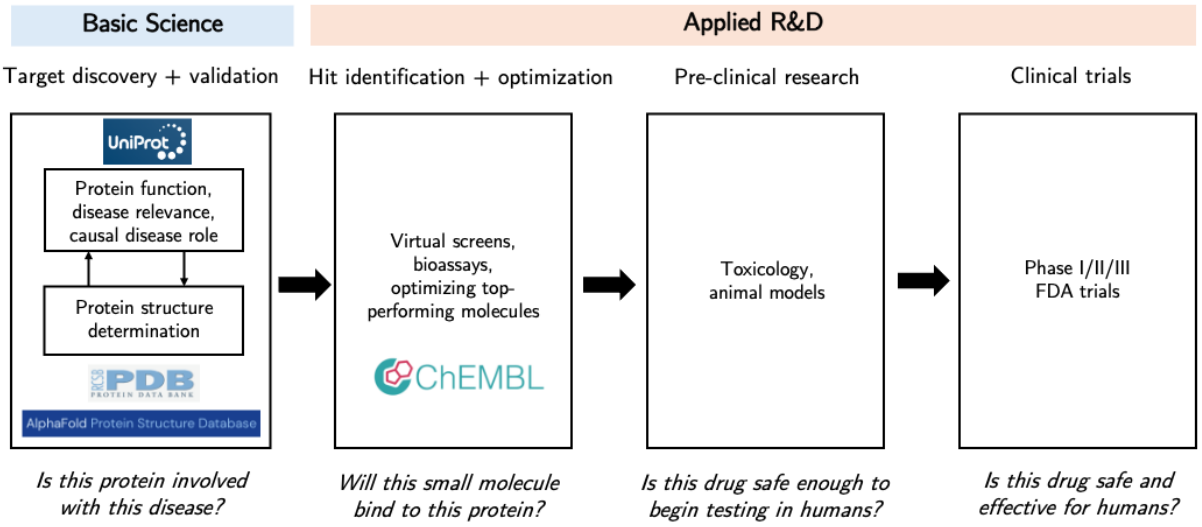
Notes: This table presents Poisson difference-in-differences regression estimates comparing previously solved and unsolved proteins before and after AlphaFold. Post is defined as all papers published in 2022 or later. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to AlphaFold's release on July 22, 2021. Column 1 outcome is counts of all non-structure papers linked to the protein, including those linked to multiple proteins. Column 2 outcome is fractional papers, defined as one divided by the number of distinct proteins linked to the paper. Column 3 outcome is counts of non-structure papers that map uniquely to a single protein. Column 4 outcome is counts of assay activities indexed by ChEMBL. Standard errors are clustered at the protein level. $N = 4{,}373{,}168$ SwissProt protein-years, coming from 546,646 SwissProt proteins. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Figure 1: Example of experimental vs. predicted proteins



Notes: This figure shows two examples of proteins with their experimental structure overlaid on their predicted structure. Source: DeepMind.

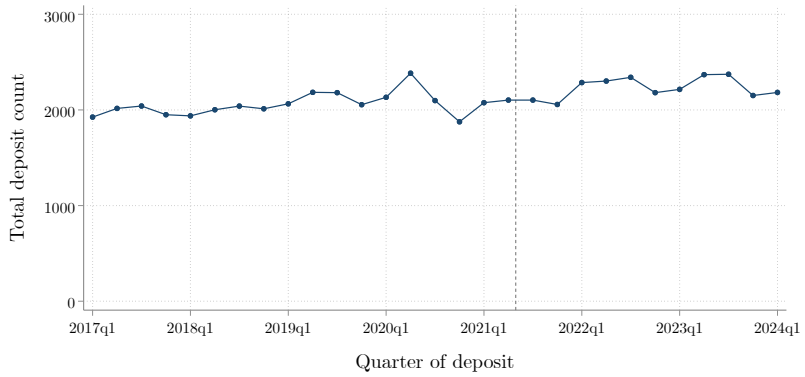
Figure 2: Simplified drug discovery process



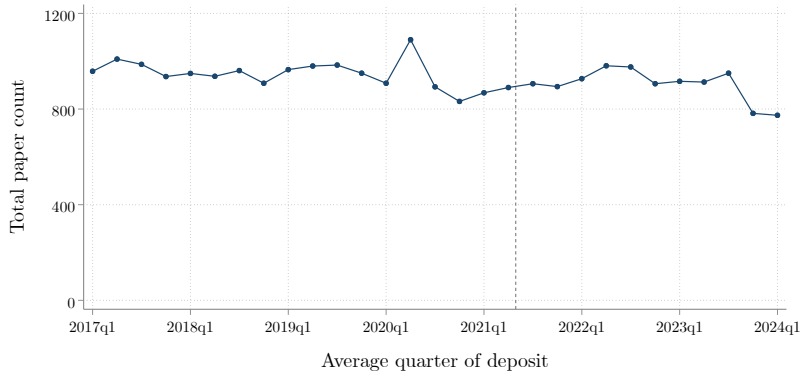
Notes: This figure lays out a simplified four-step schematic of modern drug design and highlights what steps we can observe in our data.

Figure 3: Structural biology experimental output

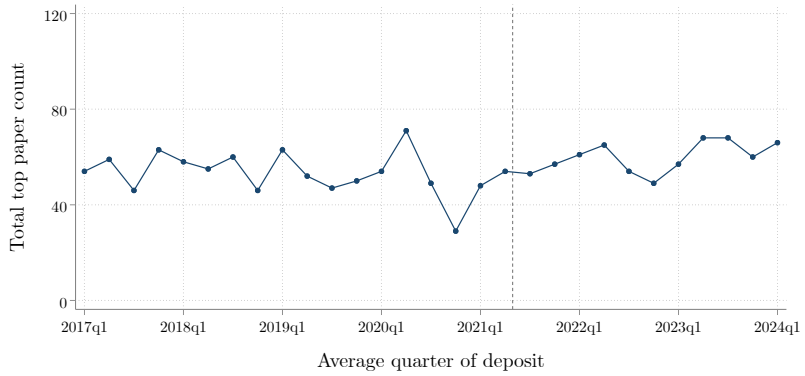
(a) Total PDB deposits



(b) Total paper counts

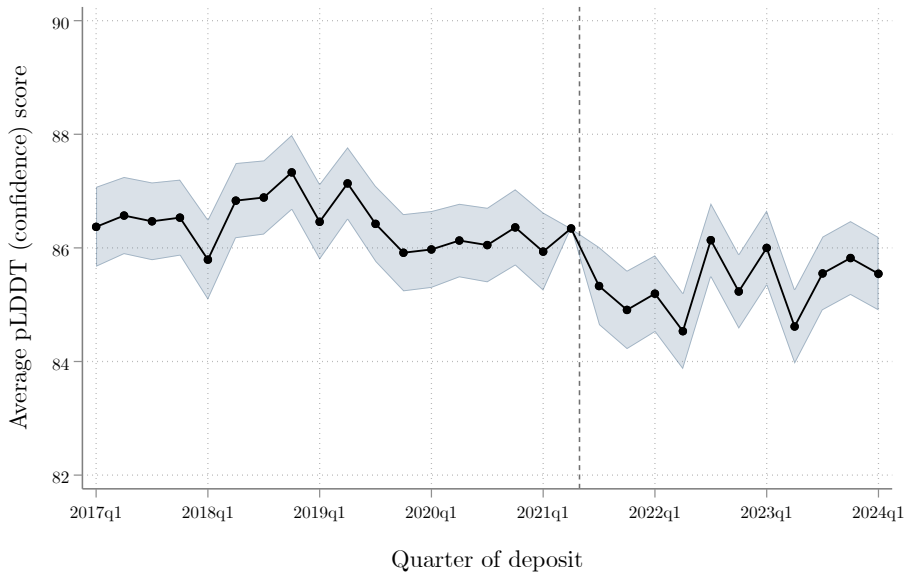


(c) Top paper counts



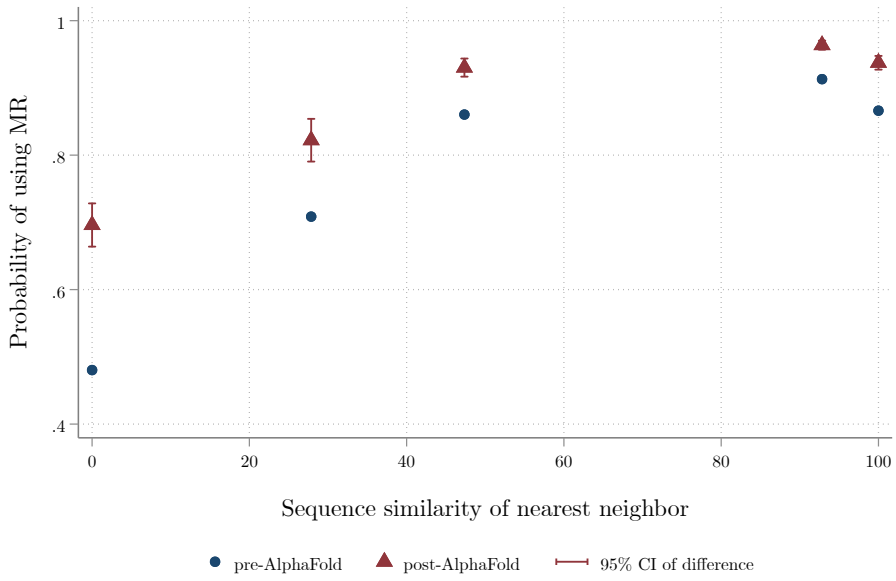
Notes: This figure shows the total counts of PDB structures and papers in our analysis sample over time. Panel (a) plots PDB deposits based on their deposit dates. Panel (b) collapses structure(s) to the paper level. If there are multiple structures per paper, the paper is assigned the average deposit date of the structures. Panel (c) restricts to "top papers," defined as papers published in *Cell*, *Nature*, and *Science*. $N = 61{,}638$ proteins, linked to 26,930 papers.

Figure 4: Average AlphaFold confidence scores of experimentally solved structures



Notes: This figure takes every experimentally solved structure and finds its AlphaFold predicted analog where possible. We then plot the average confidence scores (pLDDT) over time. The shaded area represents the 95% confidence interval of the difference from the omitted quarter. $N = 52{,}277$ proteins that have a corresponding AlphaFold confidence score.

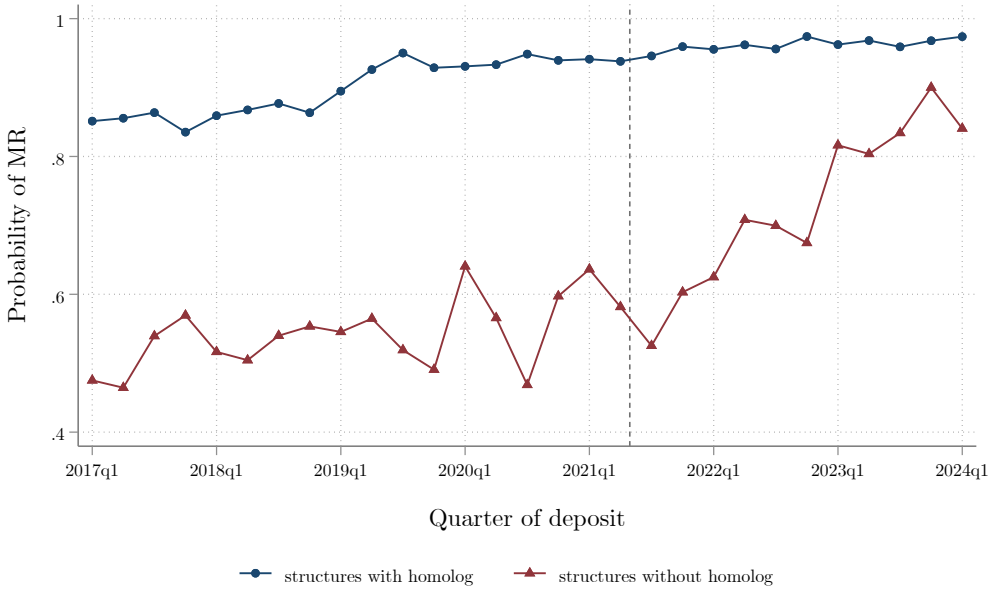
Figure 5: Molecular replacement usage, before and after AlphaFold



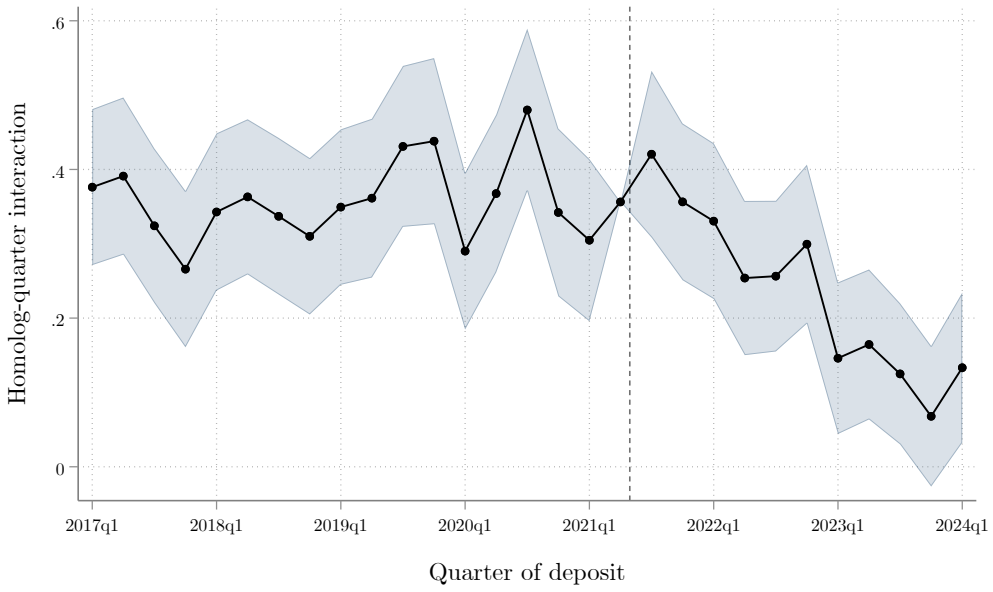
Notes: This figure shows how likely a structure was to be solved by molecular replacement, before and after AlphaFold, following Equation 1 in the text. The x-axis shows the homology score (sequence similarity of a protein's nearest experimentally-solved neighbor). The blue series shows the probability of using molecular replacement before AlphaFold, binned by the midpoint of each similarity group (β_g). The red series shows the probability after AlphaFold ($\beta_g + \gamma_g$). The bars show the 95% confidence interval of the difference (γ_g). $N = 46{,}570$ proteins solved via x-ray crystallography.

Figure 6: Molecular replacement usage over time

(a) Molecular replacement, with versus without homolog



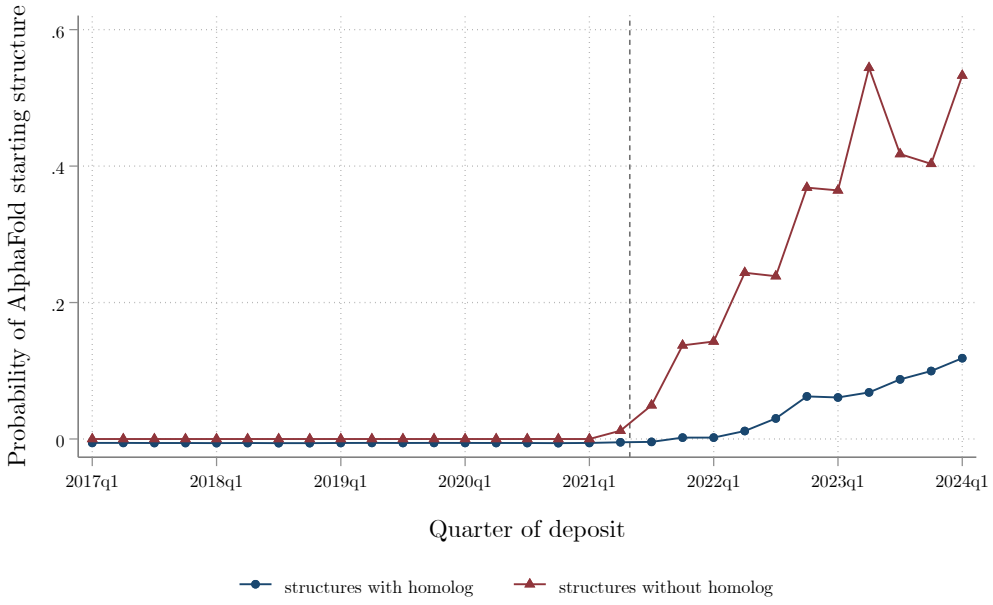
(b) Difference



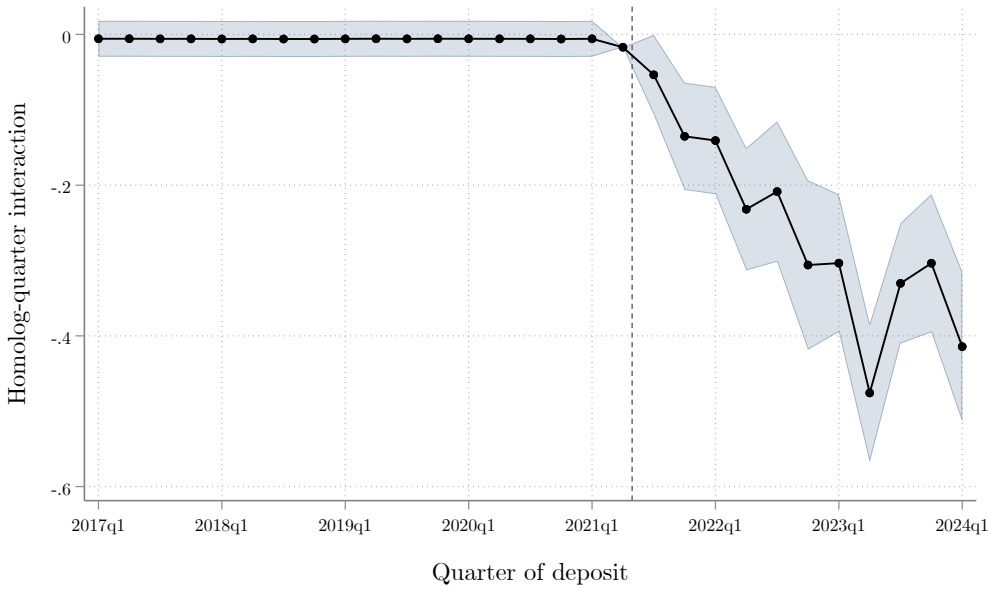
Notes: This figure shows how use of molecular replacement evolved over time, comparing proteins that did and did not have a homolog, following Equation 2 in the text. A protein with a homolog is any protein that had a released protein that was at least 30% similar in terms of amino acid sequence at the time it was deposited. Panel (a) shows the difference in average molecular replacement rates by group, plotting $(\alpha + \delta_q)$ vs $(\alpha + \lambda + \delta_q + \theta_q)$. Panel (b) shows the difference $(\lambda + \theta_q)$, with the shaded area representing the 95% confidence interval. $N = 46{,}570$ proteins solved via x-ray crystallography.

Figure 7: AlphaFold usage over time

(a) AlphaFold usage, with versus without homolog

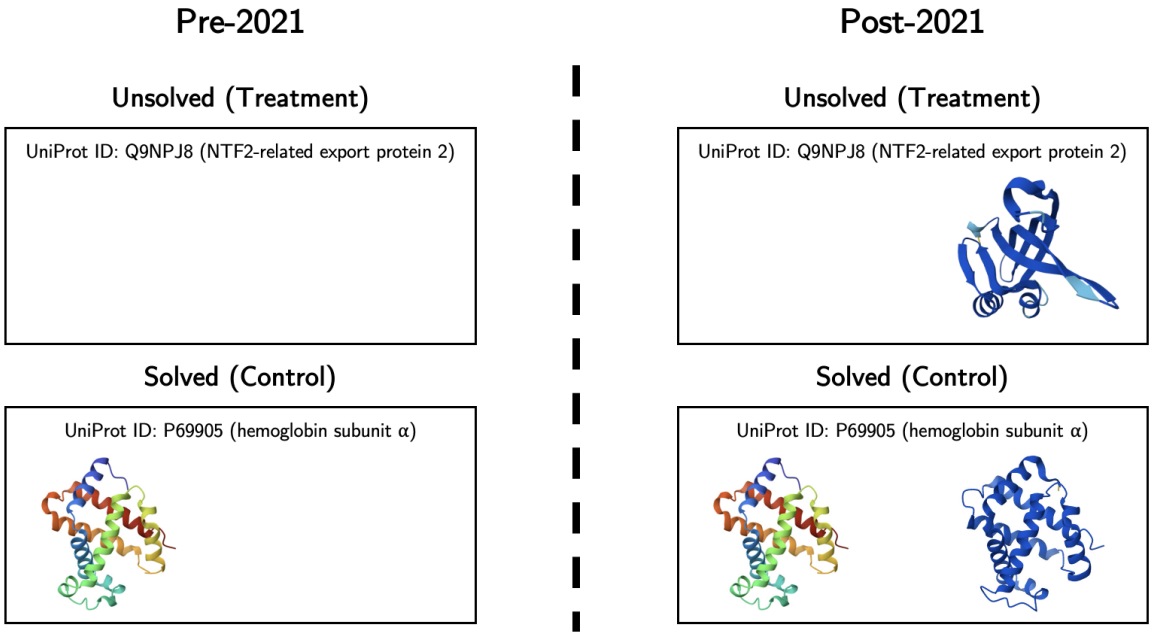


(b) Difference



Notes: This figure shows how use of AlphaFold predicted structures as starting structures for molecular replacement evolved over time, comparing proteins that did and did not have a homolog, following Equation 2 in the text. A protein with a homolog is any protein that had a released protein that was at least 30% similar in terms of amino acid sequence at the time it was deposited. Panel (a) shows the difference in average AlphaFold usage by group, plotting $(\alpha + \delta_q)$ vs $(\alpha + \lambda + \delta_q + \theta_q)$. Panel (b) shows the difference $(\lambda + \theta_q)$, with the shaded area representing the 95% confidence interval. N = 36,242 proteins solved via x-ray crystallography, using molecular replacement, and listing a starting model.

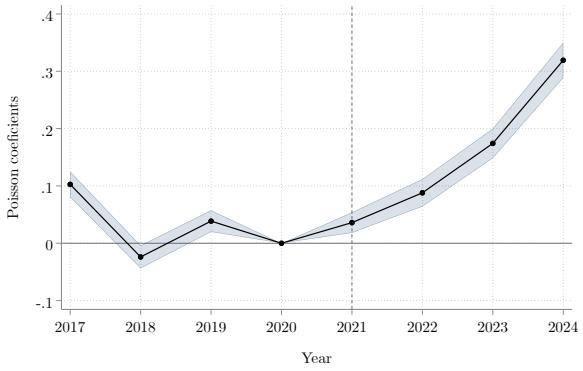
Figure 8: Difference-in-differences strategy



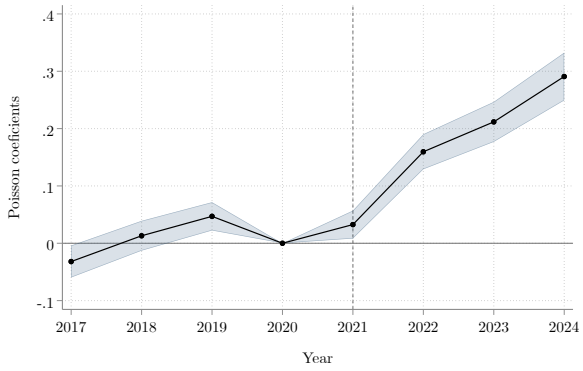
Notes: This figure explains our difference-in-differences strategy. Prior to 2021, UniProt ID Q9NPJ8 had not been solved, so researchers had no structural information about it. Meanwhile, UniProt ID P69905 had been solved and researchers had access to an experimental structure. Post 2021, both structures had a predicted structure. For UniProt ID Q9NPJ8, this represented a large shock to our knowledge of its structure. For UniProt ID P69905, this represented a much smaller knowledge shock, given that we already had the experimental structure. Experimental structure taken from PDB ID 1A01. Predicted structures taken from AlphaFold IDs AF-Q9NPJ8-4-F1 and AF-P69905-F1.

Figure 9: Non-structure publications: solved vs. unsolved proteins

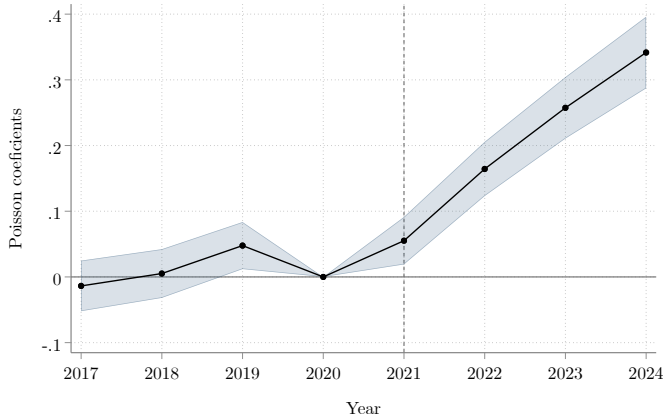
(a) All paper count



(b) Fractions of papers

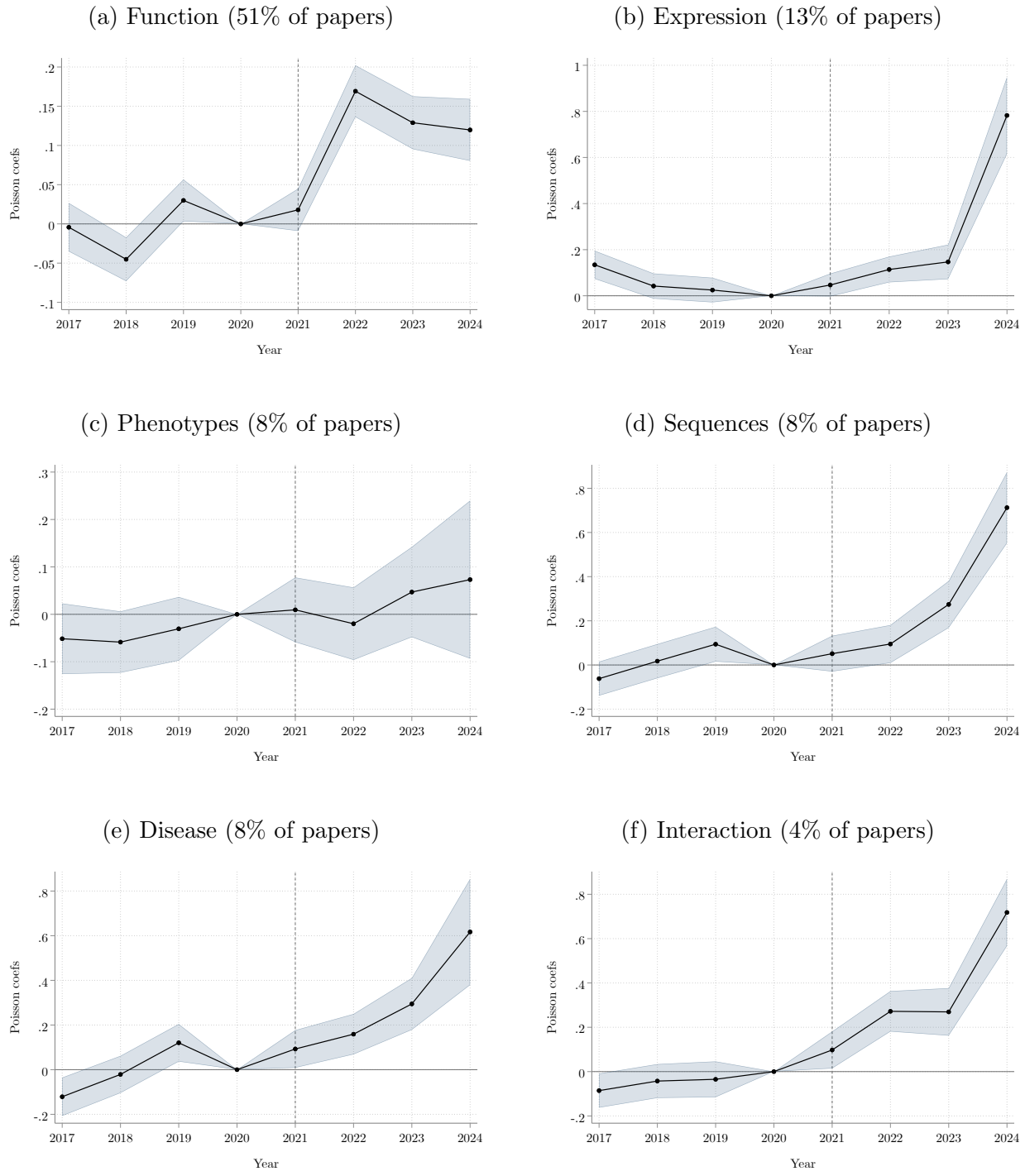


(c) Single protein paper count



Notes: This figure reports the Poisson difference-in-differences coefficients comparing non-structure paper counts for solved vs. unsolved proteins before and after AlphaFold, following Equation 3 in the text. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to AlphaFold's release on July 22, 2021. Panel (a) outcome is counts of all non-structure papers linked to the protein, including those linked to multiple proteins. Panel (b) outcome is fractional papers, defined as one divided by the number of distinct proteins linked to the non-structure paper. Panel (c) outcome is counts of non-structure papers that map uniquely to a single protein. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the 100% cluster level. N = 4,373,168 SwissProt protein-years, coming from 546,646 SwissProt proteins.

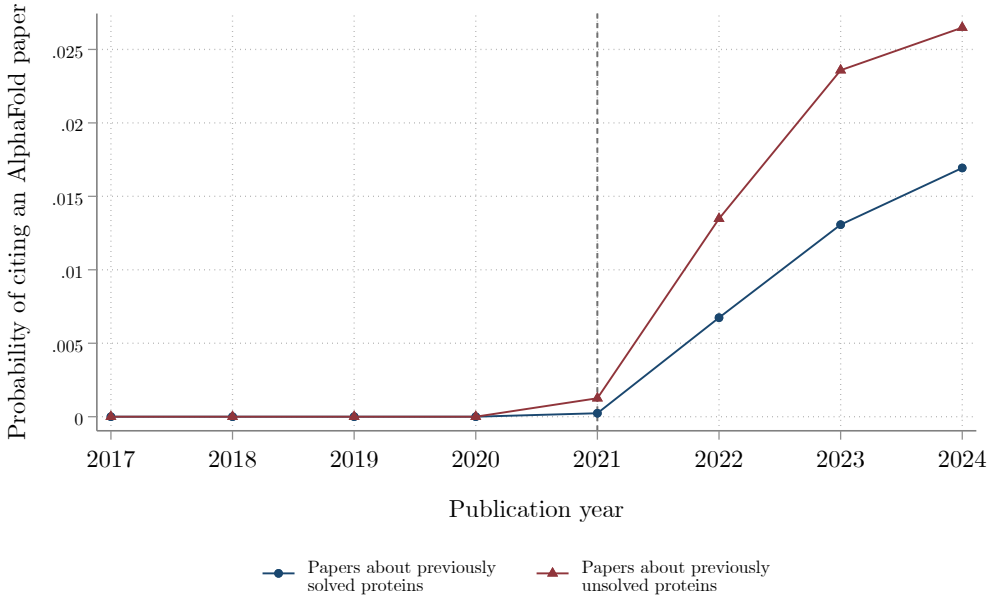
Figure 10: Categories of publications: solved vs. unsolved proteins



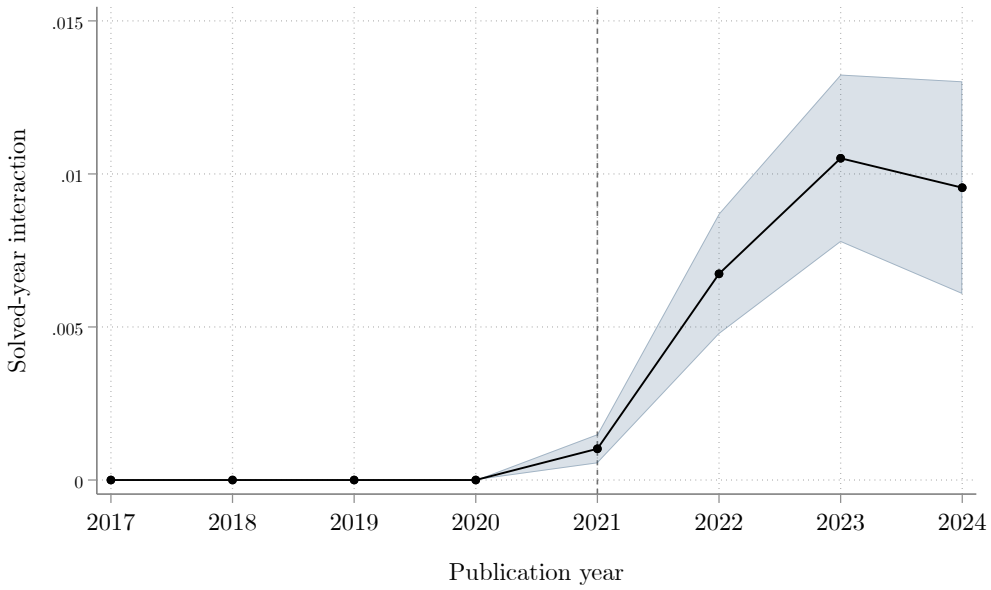
Notes: This figure reports the Poisson difference-in-differences coefficients comparing fractional paper counts for solved vs. unsolved proteins before and after AlphaFold, following Equation 3 in the text. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to AlphaFold's release on July 22, 2021. Each panel shows a different paper category as assigned in the UniProt literature curation process based on the main basic research topics covered in the paper. In parentheses, we report the share of all papers in the time period that are tagged with that category (smaller categories are excluded from this plot). Confidence intervals are shaded in blue and are calculated with standard errors clustered at the 100% cluster level. N = 4,373,168 SwissProt protein-years, coming from 546,646 SwissProt proteins.

Figure 11: Citations to AlphaFold papers

(a) Citation share by year, solved vs. unsolved proteins

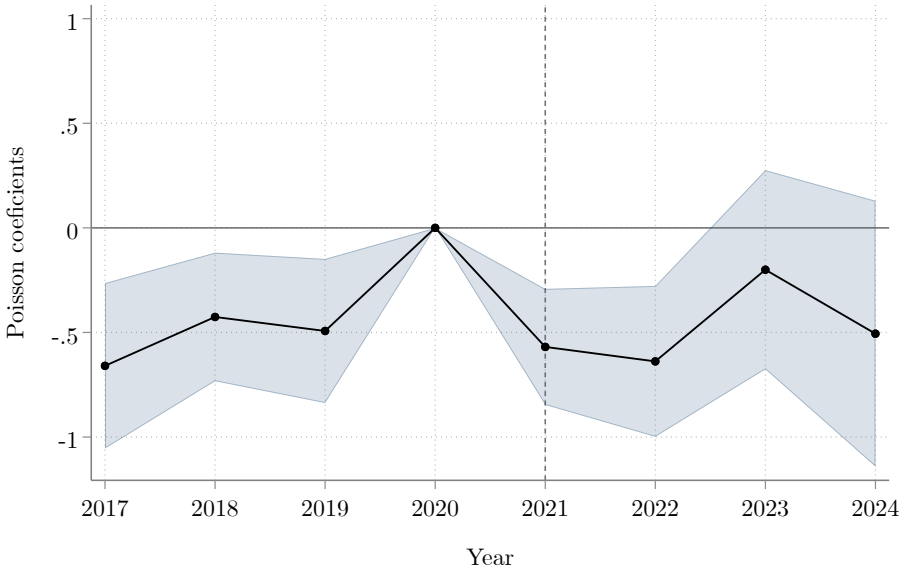


(b) Difference



Notes: This figure shows how citations to AlphaFold papers evolved over time, comparing papers about proteins that were previously solved vs. unsolved. If the paper is linked to a mix of solved and unsolved proteins, we categorize based on the status of the majority of proteins in the paper. We define AlphaFold citations as those to any of the three papers EMBL requests authors to cite when using the AlphaFold models or database: Jumper et al., 2021; Varadi, Anyango, et al., 2022; and Varadi, Bertoni, et al., 2024. The regression specification is a version of Equation 2 in the text, adapted to the paper sample. Panel (a) shows the difference in AlphaFold citation rate by group, plotting $(\alpha + \delta_q)$ vs $(\alpha + \lambda + \delta_q + \theta_q)$. Panel (b) shows the difference $(\lambda + \theta_q)$, with the shaded area representing the 95% confidence interval. N = 355,556 papers indexed by UniProt from 2017-2024.

Figure 12: Early-stage drug discovery: solved vs. unsolved proteins



Notes: This figure reports the Poisson difference-in-differences coefficients comparing ChEMBL bioactivity activity counts for solved vs. unsolved proteins before and after AlphaFold, following Equation 3 in the text. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to AlphaFold's release on July 22, 2021. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the 100% cluster level. N = 4,373,168 SwissProt protein-years, coming from 546,646 SwissProt proteins.

Appendix

A Data appendix

In this section we describe where we sourced, how we cleaned, and how we constructed the data for this paper. All of the data used in this project is publicly available.

A.1 Protein Data Bank (PDB)

The Protein Data Bank (PDB) was established in 1971. Today, the archive is managed by the Worldwide Protein Data Bank (wwPDB), an international consortium that maintains a single, globally standardized archive of experimentally determined macromolecular structures (Berman et al., 2007). The wwPDB includes regional PDB partners such as the Research Collaboratory of Structural Bioinformatics Protein Data Bank (RCSB PDB), the Protein Data Bank in Europe (PDBe), and the Protein Data Bank in Japan (PDBj), and collaborates closely with related structural biology archives including the Biological Magnetic Resonance Data Bank (BMRB) and the Electron Microscopy Data Bank (EMDB). New structure depositions are submitted through the wwPDB OneDep system and are processed by wwPDB deposition centers according to shared standards, with responsibilities divided geographically (wwPDB consortium, 2019). Additional details about the PDB data can be found on RCSB PDB website.[1] All structures deposited in the PDB are publicly released after a holding period, at which point full metadata and coordinates become accessible. We access the data via the RCSB PDB RESTful API. Documentation can be found at `https://data.rcsb.org/#data-api`.

A.1.1 Data structure

An important point that we abstract away from in the paper is that some protein *structures* are made up of multiple *entities*. A single protein chain folds up into an entity, but in some cases multiple entities bind together to form a multi-entity structure. This complicates some of the data construction in ways we will highlight when relevant.

A single protein structure represents a fairly consistent unit of scientific effort, and so we want to perform our PDB analyses at the structure level. At several points, we need to aggregate some entity-level measures up to the structure level. However, AlphaFold and

[1] `https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction`

UniProt/SwissProt are indexed at the entity level. Thus, when building our spillover panel, we structure out analysis around UniProt IDs, which are at the entity level.

A.1.2 Data access and download

We collected PDB data via two separate API endpoints covering all indexed records. Structure-level metadata (also called "entry-level" metadata) including experimental method, deposition and release dates, and linked publication identifiers were downloaded in May 2025 from `https://data.rcsb.org/rest/v1/core/entry/{pdb_id}`.

Entity-level metadata, including the amino acid sequences and UniProt identifiers were downloaded in May 2025 from `https://data.rcsb.org/rest/v1/core/polymer_entity/{pdb_id}/{entity_id}`.

A.1.3 Key variables extracted

At the structure level, we extract: structure title, deposit date, release date, experimental data collection date, experimental method, structure solution method for x-ray structures, the source of the starting model used to initialize refinement (template structure), PubMed identifier for linked paper, journal name, and group deposit identifiers.

At the entity level, we extract: UniProt identifier, RCSB cluster similarity measures, standardized amino acid sequence, and sequence length.

A.1.4 Sample construction

The analysis sample is constructed from the merged structure and entity datasets, applying four sequential restrictions:

First, we drop redundant structures. Structures explicitly flagged by depositors as group deposits are thinned to one per group (retaining the first by PDB identifier). We then identify implicit group deposits: any structure whose entities share the same deposit date, same complete author list, and same RCSB 100%-identity cluster assignment (this is a measure of the similarity of the amino acid sequence, described in more detail below). Since the similarity clusters are at the entity level, we compute a structure-level redundancy share and drop structures where more than 50% of entities are flagged as redundant as long as the date and author conditions are met. The motivation for dropping all but one of these group deposits is that we are trying measure *effort* expended on structure determination. These group deposits lead to hundreds or thousands of near-identical structures, and the effort required is closer to that of a single structure than 100x or 1000x the effort of a single structure.

Second, we drop COVID-19-related structures, identified by keyword matching on the structure title ("covid", "cov-2", "sars" combined with "cov", "coronavirus").

Third, we restrict to protein structures—keeping only entries classified as "Protein (only)", "Protein/NA", or "Protein/Oligosaccharide."

Fourth, we restrict to structures deposited between January 1, 2017 and March 31, 2024, dropping the most recent twelve months of data to account for the release lag between deposition and public availability. After these four restrictions, the analysis sample contains 61,638 structures.

A.2 Sequence homology and MMseqs2

To classify PDB structures by whether their protein sequences had a solved experimental homolog at the time of deposit, we compute pairwise sequence similarity scores using MMseqs2, a widely used tool for fast protein sequence search and clustering (Steinegger and Söding, 2017). More information, as well as download links for this software, can be found at https://github.com/soedinglab/MMseqs2.

A.2.1 Input sequences

Sequences are drawn from the PDB entity-level data (entitySequenceStd). Before running MMseqs2, we apply three restrictions: we retain only entities classified as proteins, drop entities shorter than 10 residues, and drop entities whose sequences consist entirely of the ambiguous residue code "X."

A.2.2 Computing sequence homology on a rolling basis

For each entity in the PDB, we take its amino acid sequence. Each sequence is compared against a dynamic reference pool consisting of the sequences from all entities from structures released prior to that entity's own deposit month. This rolling definition ensures the reference pool is contemporaneous with the time of deposit. We use MMseqs2 to perform these comparisons, and it returns the single best match for each query, along with percent of the sequence that is identical.

A.2.3 Aggregation to the structure level

Entity-level similarity scores are aggregated to the structure level by averaging over all entities within a structure, weighting by the number of amino acids in the entity. A structure is classified as having a homolog if the weighted mean identity meets or exceeds 30%, the

conventional threshold in structural biology for being able to use a structure as a template in molecular replacement.

A.3 AlphaFold Protein Structure Database

The AlphaFold Protein Structure Database (AlphaFold DB), maintained by DeepMind and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), provides computational three-dimensional structure predictions for hundreds of millions of proteins (Varadi, Anyango, et al., 2022). Each prediction is accompanied by per-residue confidence scores (pLDDT) ranging from 0 to 100. Data can be linked back to the PDB using UniProt identifiers. Note that unlike the PDB, these data are at the entity level.

A.3.1 Data access and download

Metadata on all 200 million plus predicted structures are hosted on Google BigQuery. Because of the size of the dataset, we only download the observations we need. To do this, we query AlphaFold DB with the list of all UniProt IDs linked to our PDB entities to create a dataset that links to our PDB data. More information is available at `https://github.com/google-deepmind/alphafold/blob/main/afdb/README.md`. The BigQuery data can be accessed directly at `gs://public-datasets-deepmind-alphafold-v4`. We queried the data in April 2025.

In addition, there is an option to direct download the subset of AlphaFold DB that links to SwissProt (discussed below). This download is available at `https://alphafold.ebi.ac.uk/download`. We downloaded it in March 2025.

A.3.2 Key variables extracted

The primary variable we use from AlphaFold DB is average predicted local distance difference test (pLDDT). pLDDT is a per amino acid measure of local confidence. It is scaled from 0 to 100, with higher scores indicating higher confidence and usually a more accurate prediction. Average pLDDT averages over all amino acids in the structure.

When merging to the PDB at the structure level, in cases of multi-entity structures, we have a many-to-one merge. When this happens, we average the confidence score, weighting by the number of amino acids in each entity. This preserves pLDDT as the average confidence across each amino acid.

A.4 UniProt and SwissProt

UniProt is the primary repository for protein sequence and functional annotation. It contains information on over 200 million known proteins (The UniProt Consortium, 2025). It consists of SwissProt, which contains manually curated entries reviewed by expert biocurators, and TrEMBL, which contains computationally annotated entries (Bairoch and Apweiler, 2000). Our analysis uses exclusively the SwissProt database.

A.4.1 Data access and download

We downloaded the full SwissProt protein list in February 2024 from the UniProtKB search interface, filtering to reviewed entries: https://www.uniprot.org/uniprotkb?query=*&facets=reviewed%3Atrue. This yields 570,829 UniProt accession numbers, which serves as the universe of proteins for the spillover panel.

UniProt curators link each protein entry to papers providing evidence for its annotations. We downloaded all publication records associated with SwissProt entries in November 2025 via the UniProt REST API: https://rest.uniprot.org/uniprotkb/{uniprot_id}/publications. This produces a protein-paper linked dataset.

A.4.2 Citation types and category filtering

The API returns citations with a citation type field. We retain only citations of type "UniProt indexed literatures"—papers directly linked to the protein entry by curators or through computationally assisted curation. We exclude large-scale background references that are associated with the database header rather than specific proteins.

Each citation carries one or more category labels indicating the aspect of protein biology it informs: Structure, Function, Sequences, Expression, Subcellular Location, Interaction, Phenotypes & Variants, PTM/Processing, Disease & Variants, Names, and Family & Domains. We parse these and construct binary indicators. Our primary spillover outcome excludes papers tagged with the Structure category, restricting to non-structure papers to ensure the outcome captures downstream scientific engagement rather than structure-determination activity itself.

A.4.3 Papers linking to multiple proteins

A single paper can be linked to multiple SwissProt proteins. To avoid double-counting when aggregating to the protein-year level in our panel, we assign each paper a fractional weight equal to one over the number of SwissProt proteins to which it is linked. For robustness and for count-based specifications, we also construct raw counts for papers linked to at most 1,

5, 10, or 15 proteins. In these cases, papers will be counted more than once if they link to multiple proteins.

A.5 OpenAlex

OpenAlex is a free, open-source index of scientific publications, maintained by the nonprofit OurResearch. It catalogues papers, authors, institutions, and citation relationships, drawing primarily on data from Crossref, PubMed, and the now-defunct Microsoft Academic Graph (Priem, Piwowar, and Orr, 2022). We use an OpenAlex snapshot downloaded by us in April 2025 and hosted on Google BigQuery.

A.5.1 Data access and download

We begin by taking the full list of SwissProt-indexed papers, and query this list against OpenAlex, using the PubMed IDs provided by SwissProt. We use this to merge on the OpenAlex ID. Call this "crosswalk 1." We repeat this process with the three key AlphaFold papers that users are encouraged to cite: Jumper et al. (2021), Varadi, Anyango, et al. (2022), and Varadi, Bertoni, et al. (2024). Call this "crosswalk 2."

Then, we query OpenAlex using OpenAlex IDs, asking to find cases where papers on crosswalk 1 cite papers on crosswalk 2. We can merge this resulting list back to our SwissProt-indexed papers to understand which of the SwissProt papers are citing AlphaFold.

A.6 UniRef and UniClust sequence clusters

To cluster SwissProt proteins based on amino acid sequence, we use sequence-based clustering the UniRef and UniClust databases. UniRef is part of UniProt, and it provides clustered sets of sequences based on the overlap of the protein's amino acid sequence using MMSeqs2. For example, the UniRef90 cluster data is built by clustering amino acid sequences with at least 11 amino acids that have at least 90% sequence identity and 80% overlap with the longest amino acid sequence in the cluster. We could perform this clustering ourselves using MMSeqs2 as described above. However, given the very large size of this database, we prefer to use these pre-constructed cluster measures.

A.6.1 Data access and download

We download the 100%, 90%, and 50% clusters for all structures in SwissProt. These files are sourced from the Uniprot ID Mapping tool, the documentation of which can be found

here: https://www.uniprot.org/help/id_mapping. We downloaded these data in April 2024.

UniRef does not host data on 30% clusters, so we sourced this from a UniClust snapshot. Our snapshot of clusters of all SwissProt structures was downloaded in March of 2024. However, the snapshot itself was from 2018. However, because the selection of SwissProt proteins has been very stable over time, we are missing 30% clusters for less than 1% of SwissProt proteins. The most recently-available snapshots are available here: https://uniclust.mmseqs.com/.

A.7 ChEMBL

ChEMBL is a large public database of drug-discovery information maintained by EMBL-EBI. It combines data on small molecules, biological targets (often proteins), assays (experiments), and measured bioactivities (outcomes), with much of the content manually curated and standardized from the medicinal chemistry and pharmacology literature (Zdrazil et al., 2024). In practice, ChEMBL records which compounds were tested in which experiments, against which targets, and with what measured activity, while also linking compounds to standardized chemical structures and many protein targets to external resources such as UniProt. These features make it useful for studying the experimental and pharmacological activity associated with a given protein target across a wide range of compounds and assay types.

A.7.1 Data access and download

The entire ChEMBL database can be downloaded from https://chembl.gitbook.io/chembl-interface-documentation/downloads. We downloaded the entire 36th release in December 2025. Given the size of the data (around 35 gigabytes uncompressed), we hosted the SQL files on the Haas School of Business high-performance computing cluster.

The data itself is comprised of many different SQL tables, which are extremely well-documented at https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_36_schema.png and https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/schema_documentation.html.

A.7.2 Key variables extracted

Given the size of the ChEMBL data, we built an SQL query that extracted only observations and variables that were relevant for our analysis. They query joins data from several tables in the ChEMBL relational schema.

Target dictionary. This table is that the target level. The backbone of the query is the biological targets. For each target, we select the target ID, name, and type. We restrict to single-protein targets in ChEMBL so that we can link back to our SwissProt sample.

Target components. This table is at the component level (analogous to entities in the PDB). Each target ID links to one or more components. We keep the component IDs.

Component sequences. This table is also at the component level. For each component ID, we merge on the UniProt IDs. We filter to components where the UniProt ID is non-missing.

Assays. This table is at the experimental assay level. Each assay has a target. Thus, we merge all assays to their target using the target ID. We keep the assay ID, assay type, and the assay source.

Activities. This table is at the activity level. Each entry records a single bioactivity measure from a specific assay. We merge activities to their respective assays using the assay ID. We keep the activity ID, the document ID, the activity's outcome and unit of measurement, and notes about data validity and potential duplicates.

Documents. This table is at the document level. Documents are the source from which the data is extracted (typically papers, patents, or datasets). We merge these to their respective activities using the document ID. One document links to many activities. We keep the source ID, journal, the publication year, the document type, the PubMed ID, the DOI, the patent ID, and the ChEMBL release ID.

Source. This table is at the source level. We merge this based on the source ID. For each source, we keep the source description which provides additional information on where a document was ingested from (e.g., ChEMBL literature curation or other data sources such as BindingDB or PubChem).

ChEMBL Release. This table the level of the ChEMBL release. We merge this to documents using the ChEMBL release ID. We keep a variable which provides the date of the ChEMBL release that first incorporated the document.

A.7.3 Data cleaning

At a high level, the resulting data build provides information on bioactivities, when these activities were published, and the proteins that these bioactivities target. The final dataset links over 10 million bioactivities to 10,731 unique single-protein targets.

We exclude records flagged by ChEMBL as potential duplicates. We assign activities to the date that their respective document was published. About 40% of activities do not have dates; this usually happens when the activities are sourced from a third-party database (such as BindingDB or PubChem). Occasionally these database activities do have a publication date, but this is often inaccurate as it represents the single date the dataset was published or incorporated, rather than the date that the underlying science was done. Thus, we re-code these dates as also missing.

A.7.4 Integration into the spillover panel

With activities and assays now assigned dates, we can collapse the counts of activities and assays at the protein-year level. We can then merge these protein-year counts into our spillover panel using UniProt IDs. Just over 90% of the unique UniProt IDs in ChEMBL merge to SwissProt (the remaining are in the uncurated TrEMBL portion of the data). For any UniProt ID-years in our spillover panel that do not merge, we impute zeroes.

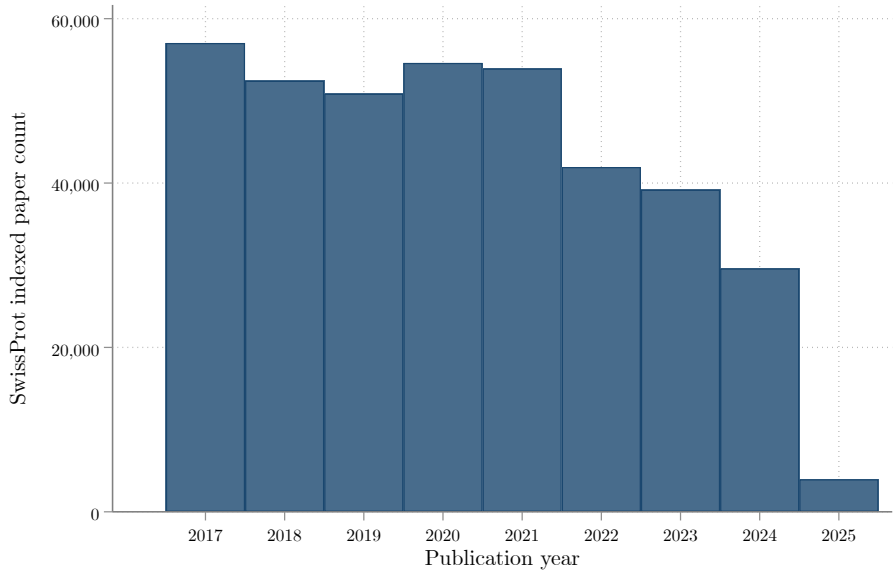
B Supplemental tables and figures

Appendix Table 1: Difference-in-differences for papers in related fields: Robustness by solved definition

Dependent Variable:	(1) Non-structure papers (all paper count)	(2) Non-structure papers (fractions of papers)	(3) Non-structure papers (single protein papers only)
Panel A. "Unsolved" defined with 30% sequence similarity			
Post	-0.423***	-0.438***	-0.421***
	(0.0092)	(0.0123)	(0.0140)
Unsolved	-0.929***	-1.472***	-1.616***
	(0.0368)	(0.0435)	(0.0468)
Post x Unsolved	0.064***	0.139***	0.154***
	(0.0113)	(0.0151)	(0.0186)
Observations	4,356,960	4,356,960	4,356,960
Panel B. "Unsolved" defined with 50% sequence similarity			
Post	-0.449***	-0.463***	-0.445***
	(0.0093)	(0.0120)	(0.0136)
Unsolved	-1.227***	-1.738***	-1.890***
	(0.0368)	(0.0433)	(0.0460)
Post x Unsolved	0.101***	0.168***	0.186***
	(0.0110)	(0.0150)	(0.0182)
Observations	4,373,168	4,373,168	4,373,168
Panel C. "Unsolved" defined with 90% sequence similarity			
Post	-0.470***	-0.480***	-0.464***
	(0.0103)	(0.0128)	(0.0143)
Unsolved	-1.825***	-2.313***	-2.474***
	(0.0335)	(0.0403)	(0.0431)
Post x Unsolved	0.126***	0.181***	0.202***
	(0.0115)	(0.0150)	(0.0176)
Observations	4,373,168	4,373,168	4,373,168
Panel D. "Unsolved" defined with 100% sequence similarity			
Post	-0.489***	-0.497***	-0.481***
	(0.0102)	(0.0126)	(0.0142)
Unsolved	-2.173***	-2.656***	-2.822***
	(0.0329)	(0.0398)	(0.0425)
Post x Unsolved	0.150***	0.201***	0.224***
	(0.0114)	(0.0149)	(0.0177)
Observations	4,373,168	4,373,168	4,373,168

Notes: This table presents Poisson difference-in-differences regression estimates comparing previously solved and unsolved proteins before and after AlphaFold. Column 1 outcome is counts of all non-structure papers linked to the protein, including those linked to multiple proteins. Column 2 outcome is fractional papers, defined as one divided by the number of distinct proteins linked to the paper. Column 3 outcome is counts of non-structure papers that map uniquely to a single protein. Panels A through D use different definitions for "unsolved" based on varying degrees of sequence similarity. The broadest definition is 30% similarity in Panel A. Note the smaller sample size is because 30% clusters were sourced from an older dataset with slightly fewer clusters available. The narrowest definition is 100% similarity in Panel D, which is the preferred definition in the main text. Standard errors are clustered at the protein level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Appendix Figure 1: SwissProt-indexed publications by publication year

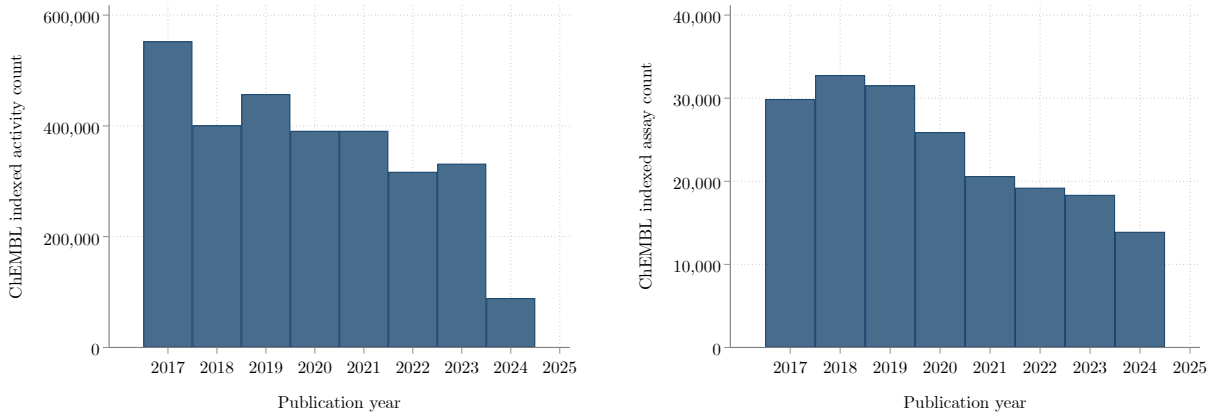


Notes: This figure shows the histogram of publication years for SwissProt-linked papers. $N = 383,625$ papers published between 2017 and 2025.

Appendix Figure 2: ChEMBL-indexed activities and assays by publication year

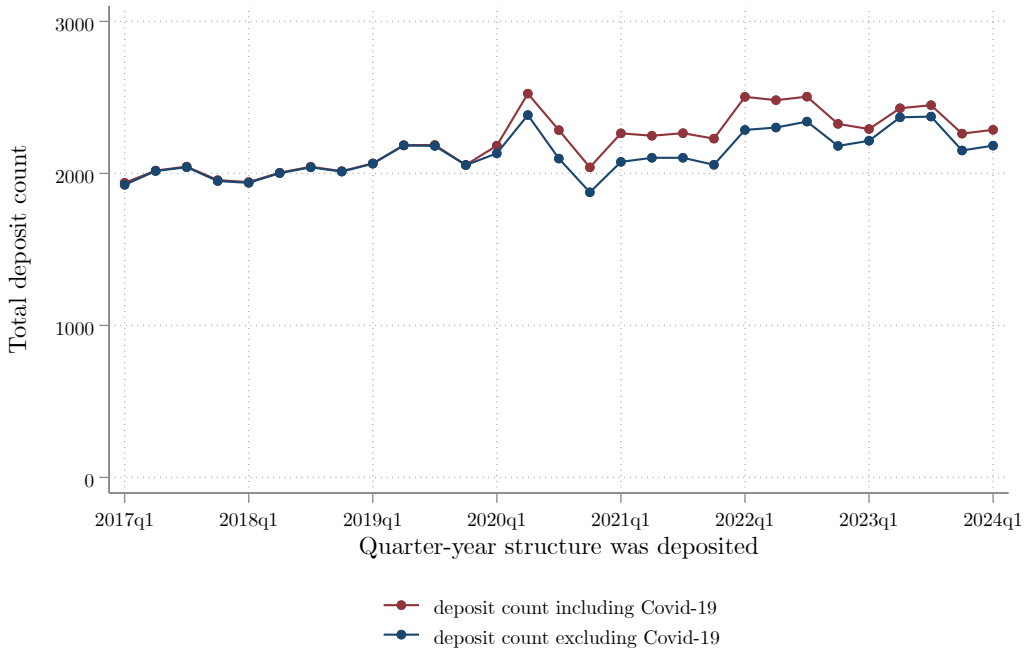
(a) Activities

(b) Assays



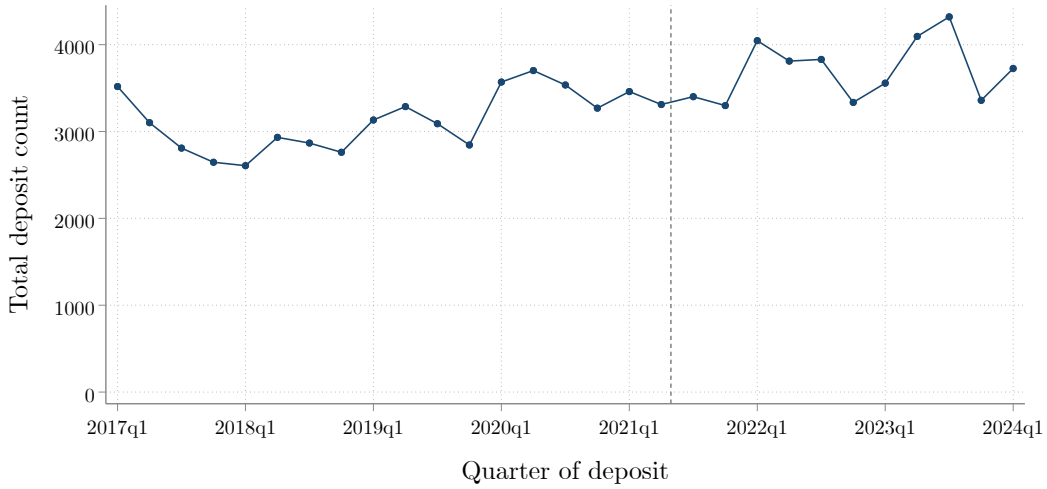
Notes: This figure shows the histogram of publication years for ChEMBL-linked activities and assays. $N = 2,930,804$ activities and 192,230 assays published between 2017 and 2025.

Appendix Figure 3: Total PDB deposits, including Covid-19 structures



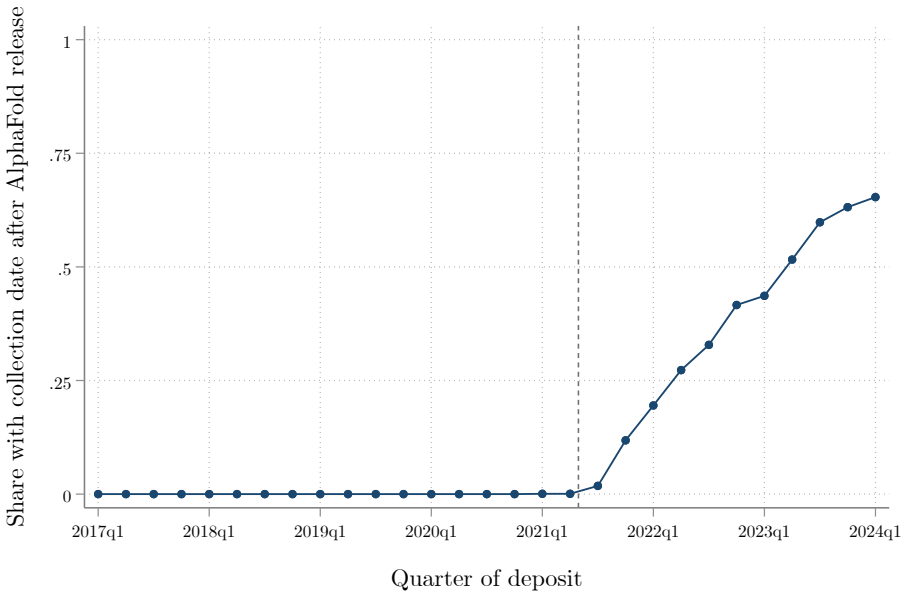
Notes: This figure shows the total counts of PDB structures in our analysis sample over time, adding back in Covid-19 structures. $N = 64,035$ structures.

Appendix Figure 4: Total PDB deposits, no restrictions



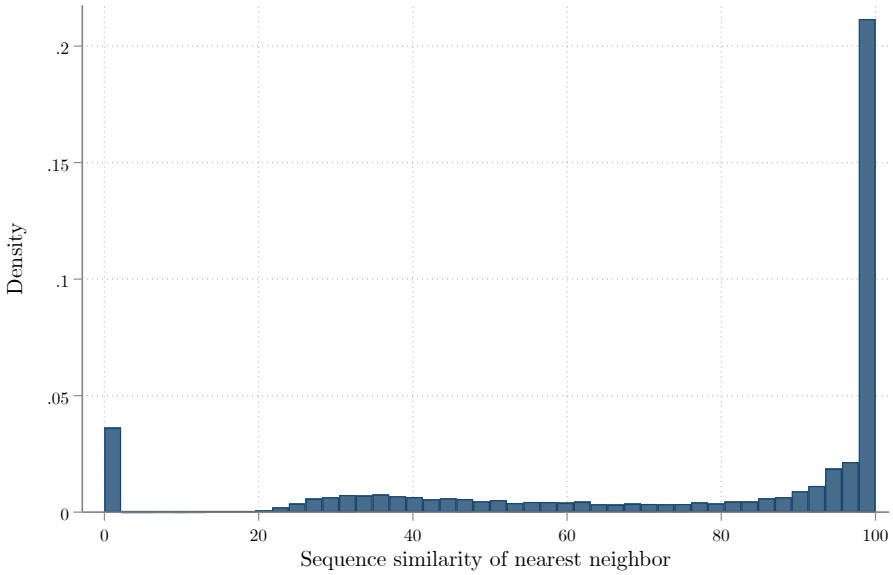
Notes: This figure shows the total counts of PDB structures in our analysis sample over time, making zero sample restrictions. $N = 97,225$ structures.

Appendix Figure 5: Share of structures collecting data after AlphaFold release



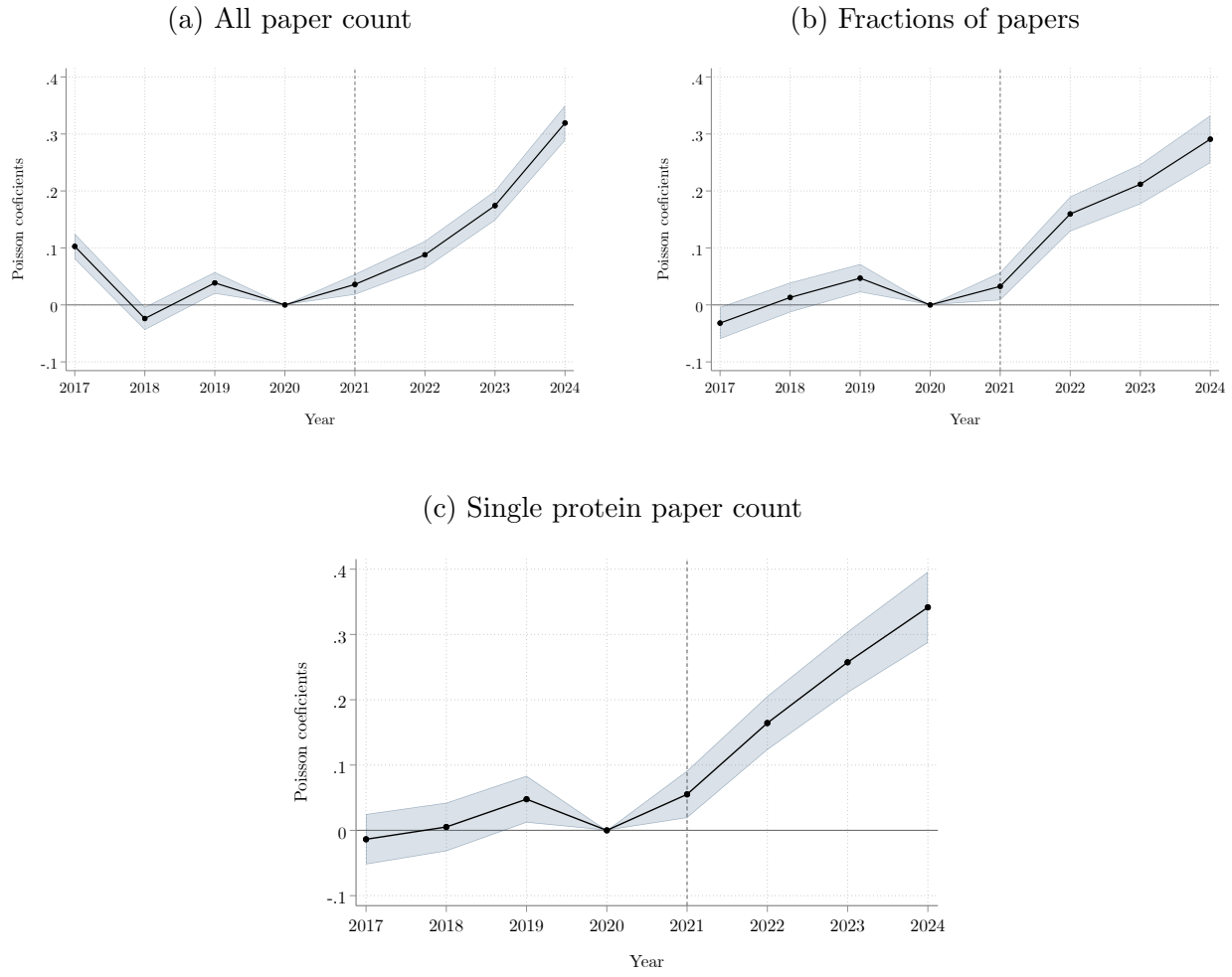
Notes: This figure shows the share of PDB structures that collected their experimental data after AlphaFold's release, by quarter of deposit. $N = 46,711$ structures that report a collection date.

Appendix Figure 6: Homology scores of PDB-deposited structures



Notes: This figure shows the distribution of homology scores (sequence similarity of a protein's nearest experimentally-solved neighbor). $N = 46{,}570$ proteins in the analysis sample solved via x-ray crystallography.

Appendix Figure 7: Non-structure publications: solved vs. unsolved proteins including SARS-CoV-2

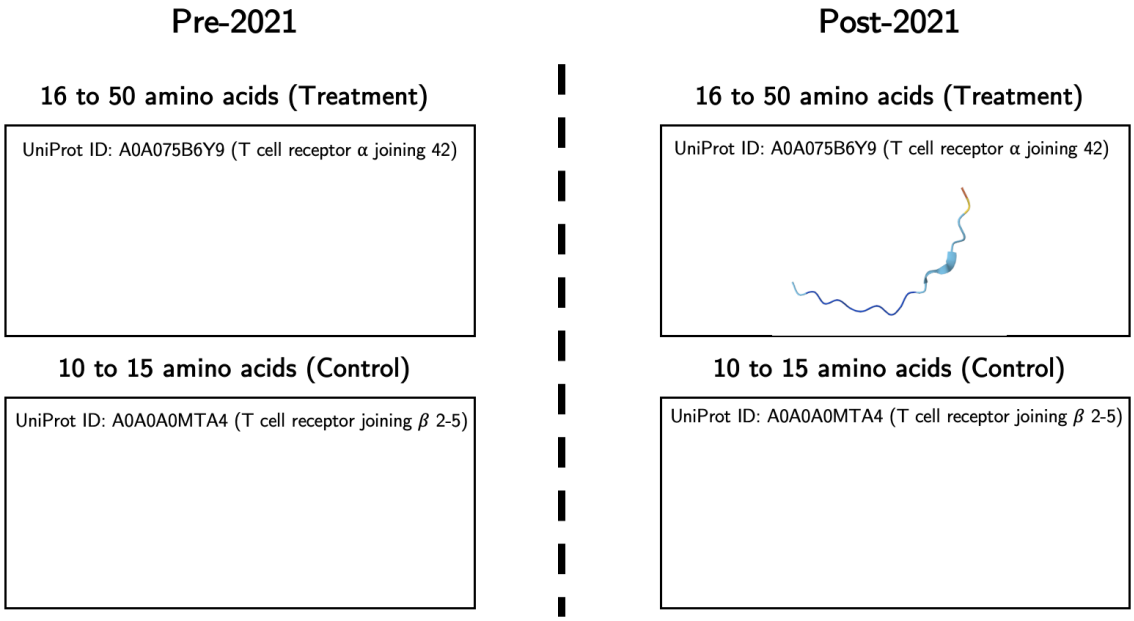


Notes: This figure reports the Poisson difference-in-differences coefficients comparing non-structure paper counts for solved vs. unsolved proteins before and after AlphaFold, following Equation 3 in the text. Here we include the 17 proteins in SwissProt that belong to the SARS-CoV-2 organism. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to AlphaFold's release in 2021. Panel (a) outcome is counts of all non-structure papers linked to the protein, including those linked to multiple proteins. Panel (b) outcome is fractional papers, defined as one divided by the number of distinct proteins linked to the non-structure paper. Panel (c) outcome is counts of non-structure papers that map uniquely to a single protein. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the 100% cluster level. N = 4,373,304 SwissProt protein-years, coming from 546,663 SwissProt proteins.

C Alternative empirical strategy: Proteins that remain unpredicted

In this section, we discuss our implementation of an alternative empirical strategy that takes advantage of the fact that AlphaFold does not provide predicted structures for all SwissProt proteins. One restriction AlphaFold implemented is that they do not predict any protein structures with fewer than 15 amino acids or more than 2700 amino acids. This motivates an empirical strategy illustrated in Appendix Figure 8.

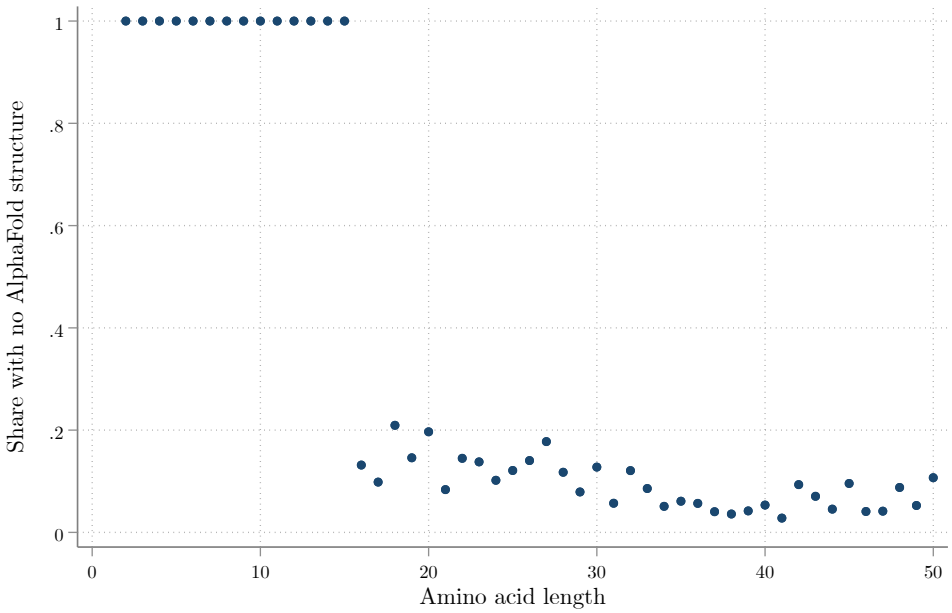
Appendix Figure 8: Alternative difference-in-differences strategy



Notes: This figure explains our alternative difference-in-differences strategy. Prior to 2021, neither structure had been solved. Post 2021, UniProt ID A0A075B6Y9 does get a predicted structure, because it has more than 15 amino acids. UniProt ID A0A0A0MTA4 does not get a predicted structure, because it has 15 or fewer amino acids. For UniProt ID A0A075B6Y9, this represented a large shock to our knowledge of its structure. Thus, the former gets an information shock post-AlphaFold, and latter does not. Predicted structure taken from AlphaFold ID AF-A0A075B6Y9-F1.

We choose the 15 amino acid cutoff because there is a dense mass of proteins around this cutoff. Appendix Figure 9 shows that cutoff is indeed binding.

Appendix Figure 9: AlphaFold coverage by amino acid count



Notes: This figure shows the share of proteins in SwissProt with 50 or fewer amino acids that do not have an AlphaFold predicted structure. $N = 13,106$ proteins.

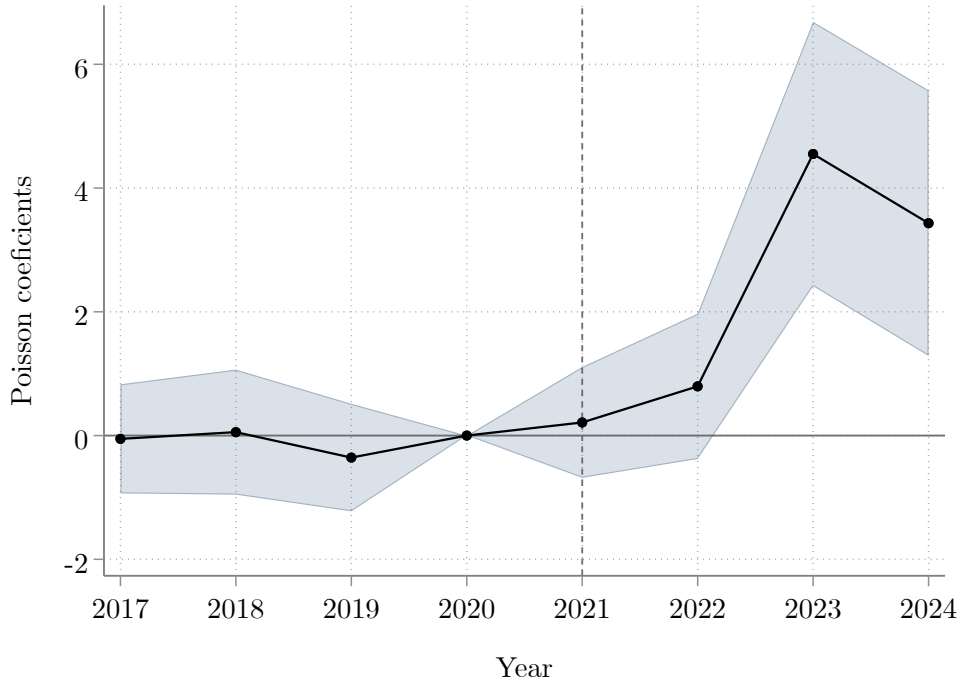
In an effort to focus on comparable proteins on either side of the cutoff, we restrict to proteins that are between 10 and 50 amino acids long, which is about 2% of the full SwissProt sample.[2] We then further restrict to proteins that were unsolved prior to AlphaFold.[3] This leaves us with a sample of 10,280 proteins.

Appendix Figure 10 shows the results of Equation 3 where the longer proteins—the proteins that receive a predicted structure—are the treated group.

[2] The mean length is 362 amino acids, the median is 295, and the standard deviation is 342. We avoid using proteins with fewer than 10 amino acids because we worry these are so small they are not comparable.

[3] We also drop the small number of proteins that have more 15 amino acids but no predicted structure, since we don't know why AlphaFold chose not to predict these.

Appendix Figure 10: Non-structure publications: predicted vs. unpredicted proteins



Notes: This figure reports the Poisson difference-in-differences coefficients comparing non-structure paper counts for predicted vs. unpredicted proteins before and after AlphaFold, following Equation 3 in the text. We restrict to proteins that had no experimental structure prior to 2021, and to structures with 10-50 amino acids. Unpredicted structures are structures with 10-15 amino acids. Predicted structures have 16 or more. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the protein level. $N = 82,240$ SwissProt protein-years, coming from 10,280 SwissProt proteins.

The effect sizes we estimate are extremely large, with the average Poisson point estimates in post-period being 2.7. This implies a relative increase in unsolved protein research of around 14x relative to solved protein research. However, the confidence intervals do not allow us to rule out much smaller (or larger effects).

References

Bairoch, Amos and Rolf Apweiler (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000". *Nucleic Acids Research* 28.1, pp. 45–48.

Berman, Helen, Kim Henrick, Haruki Nakamura, and John L. Markley (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data". *Nucleic Acids Research* 35.Database issue, pp. D301–D303.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". *Nature* 596.7873, pp. 583–589.

Priem, Jason, Heather Piwowar, and Richard Orr (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". *arXiv*. arXiv: 2205.01833 [cs.DL].

Steinegger, Martin and Johannes Söding (2017). "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets". *Nature Biotechnology* 35.11, pp. 1026–1028.

The UniProt Consortium (2025). "UniProt: the Universal Protein Knowledgebase in 2025". *Nucleic Acids Research* 53, pp. D609–D617.

Varadi, Mihaly, Stephen Anyango, et al. (2022). "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". *Nucleic Acids Research* 50.D1, pp. D439–D444.

Varadi, Mihaly, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. (2024). "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences". *Nucleic Acids Research* 52.D1, pp. D368–D375.

wwPDB consortium (2019). "Protein Data Bank: the single global archive for 3D macromolecular structure data". *Nucleic Acids Research* 47.D1, pp. D520–D528.

Zdrazil, Barbara et al. (2024). "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods". *Nucleic Acids Research* 52.D1, pp. D1180–D1192.