How Artificial Intelligence Shapes Science: Evidence from AlphaFold [*]

Ryan Hill, Northwestern University
Carolyn Stein, UC Berkeley

March 11, 2026

Abstract

We study how a frontier AI model affects scientific discovery by examining the release of the AlphaFold2 algorithm and the impact it had on structural biology and related fields of science. Structural biology is the field of science concerned with understanding the structure and function of proteins, and researchers in this field historically devoted substantial time and resources to experimentally solving three-dimensional protein structures. AlphaFold has the ability to *predict* these structures without running experiments. In July 2021, researchers gained access to hundreds of thousands of AI-predicted protein structures virtually overnight. Yet, we find that the rate of experimental structure determination remains almost unchanged. Instead, researchers appear to use predicted structures to complement experimental structure determination. Looking at downstream science that builds on protein structures, we find evidence of more basic research occurring on proteins that had no structure information prior to AlphaFold. However, we find no evidence so far that more applied, early-stage drug development is targeting these proteins.

Preliminary draft. Comments welcome.

[*] 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. Alexia Witthaus Viñé and Thomas Barden provided excellent research assistance. All errors and omissions are our own. Contact: ryan.hill@kellogg.northwestern.edu, 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. On the other hand, 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 governor 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 structure models relevant for a broad range of basic and applied research fields related to proteins.[1] This was a massive shock to structural biology, the field of research that seeks to determine the three-dimensional structures of proteins. These structural models help us understand how proteins function in cells and how they might be targeted in pharmaceutical therapies. Since the early 1970s, we relied on slow and expensive experimental methods to generate the roughly 150,000 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 running any experiments?

For decades, computational biologists and computer scientists made slow progress on

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

improving the capabilities of computational models. A breakthrough occurred when Google DeepMind's released the AlphaFold2 algorithm in 2021. 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 shock of many in the research community.

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. 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 stage drug design. 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,[2] 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 and what it automates is particularly clean. 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 area. Structural biology and protein folding is an critical area of science. Several Nobel Prizes have been awarded for the discovery of a *single* experimental protein 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

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

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 deposited. 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 no 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. It is important to caveat that this lack of substitution may not be efficient, but rather may reflect researcher incentives. 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. 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, 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 an experimental homolog. This complementarity between experimental and AI-based research seems to be facilitating an increase in the productivity of structural biologists, and may open opportunities for greater exploration of the protein space, unlocking insights into more complex proteins and biological functions.

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 where previously solved proteins have only limited new insights from AI, but previously unsolved proteins have an unexpected and comparatively large shock of new structural information. We can then compare research activity in broader and downstream fields across these two groups of proteins 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 600,000 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, fewer than 7% of proteins had an experimental structure.

The SwissProt subsample of UniProt also curates a literature review for each protein, linking basic scientific papers that have been written about the protein to its SwissProt entry. This allows us to use scientific research activity as an outcome. We find an increase in research activity in related basic research fields 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. This suggests that protein structure was a bottleneck in complementary research that has been eased by the introduction of AI tools.

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 discovery. 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. 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-enabled 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. Although some new opportunities may have been unlocked, it is possible that downstream research has not yet experienced the benefits of AI insights provided further up the research pipeline.

Related Research

This paper relates to three literatures. First, it connects to the economics of AI and automation (Boustan, Choi, and Clingingsmith, 2022; Feigenbaum and Gross, 2024; Humlum and Vestergaard, 2025), which studies whether new technologies substitute for labor in existing tasks, complement workers within those tasks, or create new tasks altogether. In the task-based framework of Acemoglu and Restrepo (Acemoglu and Restrepo, 2019; Acemoglu and Restrepo, 2022; Restrepo, 2024), 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. Recent empirical work on generative AI often finds sizable productivity gains, with especially large gains for less experienced workers (Brynjolfsson, Li, and Raymond, 2025; Noy and Zhang, 2023), which is 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 task—protein structure determination—is unusually well defined.

Second, the paper relates to a growing literature on AI and science. Agrawal, McHale, and Oettl (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. 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 closely, the 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. The closest economics paper is Yu (2026), which uses the introduction of AlphaFold2 as a shock to study scientists' productivity and inequality. Relative to this paper, we shift attention away from scientists as the unit of analysis and toward proteins, research tasks, and downstream knowledge production. In that sense, Yu asks how AlphaFold changed researchers' careers and outputs, whereas we ask how AlphaFold changed the production technology of structural biology itself and how far the resulting knowledge shock propagated into neighboring domains.

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.[3] 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).[4] Since the field's development in the 1970s, researchers have solved and publicly deposited around 150,000 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).

[3] 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.

[4] 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).

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 sequence is comparatively straightforward: it can be inferred directly from genomic sequencing data.[5] 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 sequencing 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. 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.[6]

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

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

[6] 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).

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 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 structural templates are excluded. As a result, AlphaFold can perform well even for proteins without close experimental structural homologs (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 acheived 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). Researchers also spoke about the magnitude of the breakthrough. One of the CASP assessors, evolutionary biologist Andrei Lupas, memorably said of AlphaFold2, "this will change medicine. It will change research. It will change bioengineering. It will change everything."

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. This made access to the structure predictions frictionless for researchers (Varadi and Velankar, 2023).

2.3 Applications of structural biology

As suggested by Lupas, much of the excitement around AlphaFold was focused on the downstream scientific advances it might enable. Knowing a protein's structure opens up new avenues for additional research about the protein that might otherwise not have been possible. Much of this is still very basic science. For example, many proteins' functions inside the cell are unknown. A predicted three-dimensional shape can provide clues (such as a pocket where a small molecule might bind, or a shape that resembles a known protein family) that can be pursued in follow-up lab experiments.

Another important avenue that protein structures can help open is early-stage drug discovery. Most drugs are small molecules that function by binding to a protein target. For example, aspirin, which is often used for pain and fever reduction, binds to COX-1 and COX-2 enzymes, inhibiting their activity. These enzymes help convert arachidonic acid into prostaglandins, which in turn activate pain and fever signaling. Therefore, shutting these enzymes down can be beneficial. A clear map of the binding site—the place where the small molecule drug will attach to the protein—can help researchers better understand what small molecules to try, since the shape of the site and the molecule must match up.

Drug development is a notoriously slow process. Today, taking a drug from early-stage discovery to regulatory approval and market typically takes about 10-15 years. Much of this time is devoted to clinical trials, but even pre-clinical research is slow, with researchers typically spending three to six years on target identification, hit discovery, lead optimization, and animal studies before starting human research in clinical trials. Thus, it is too early to expect to see new approved drugs as a result of AlphaFold. It is also likely too early to expect to see new clinical trials. However, we might expect to see new early-stage, pre-clinical research that has been enabled by predicted structures.

One important and early pre-clinical step is known as "hit discovery." Once a protein has been identified as a target—in other words, once scientists are confident that it is an important pathway in a disease—scientists begin testing whether small molecules bind to it. At a high level, hit discovery involves combining the target and a candidate drug, and measuring the reaction to see if they bind. A single experiment (known as an "assay") may test hundreds or thousands of candidate drugs, with each reaction known as a single "activity." If researchers have a clear structural map of the target, they can select small molecules in a more rational manner, based on the shape of the binding site. This is known as “structure-based drug discovery.” In the case of HIV-1 protease, insights from the first crystal structure released in 1989 was essential for developing some of the earliest inhibitors (Erickson et al., 1990). These early candidates may become future drugs years down the line.

3 Data sources and construction

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. 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. Importantly, every project is linked to an academic paper that first introduced the experimental structure to the literature, and we can observe the paper and its publication date.

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, we have 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 5.1 months.

The final data that we gather from the PDB are the experimental details. This includes the broad experimental method that the researchers use to determine the structure (for example, x-ray crystallography versus cryo-EM). It also includes the computational techniques they use to turn their data into 3D structures, the software and hardware they used, and any existing protein structures they used as templates.

3.2 AlphaFold DB

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 structure as well as the open-source code for the prediction model. In December 2021, they released an additional 560,000 structures, and in July 2022, they released a prediction for every known protein (200 million+ structures). This dataset, called AlphaFoldDB, is our primary source of data for AI-predicted structure details. For each entry, AlphaFoldDB provides confidence scores for every structure. The predicted Local Distance Difference Test (pLDDT) score represents how confident the model is in the predicted position of each amino acid in a protein structure, and we use the average pLDDT across all positions in the amino acid sequence as the overall measure of confidence. Multiple studies have shown that pLDDT is strongly correlated with actual prediction accuracy (Akdel et al., 2022; Tunyasuvunakool et al., 2021).

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.[7] Papers related to these proteins are manually reviewed for accuracy and relevance, then linked to the entries by curators. We access all of the linked papers to collect information about the authors, journal, and publication dates. 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.

In a five-year retrospective of AlphaFold's impact published in *Nature*, Callaway (2025) highlighted the work of biochemist Andrea Pauli. For years, Pauli had been trying to understand how sperm and egg cells fuse, working with zebrafish. 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 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

[7] This includes virtually all human proteins.

hypothesize that 3 sperm proteins—including Tmem81—formed a complex that recognized Bouncer. They conducted additional experiments to validate this hypothesis, 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. While all of these linked proteins may not be equally important or relevant in the paper, it is difficult for us to distinguish. Thus, we assign each linked protein 1/8 of the paper, which was published in 2024.

3.4 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.

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. This structure in part of the SwissProt subset of UniProt. Moreover, this structure exists in the PDB: it was first released in 2016 (PDB ID 5I6Z).

3.5 Data construction

PDB

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 tens (or even hundreds) of the near-identical structure by the same team on the same day. Because these group deposits don't represent 10x or 100x the scientific effort, we only want to keep 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. We start with the list of 570,829 SwissProt-indexed proteins. We then take the SwissProt curated publications. 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 as its own paper, but drop papers that link to an excessive number of proteins (more than ten). 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.

Next, we link ChEMBL activities and assays by their protein target. Of the 10,724 unique single-protein targets, 9,722 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.

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 is 0.62 with a standard deviation of 9. 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, with a standard deviation of 241. The vast majority of hit discovery is being done on a very small share of 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 1 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.

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 3 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 2 suggests that this is not the case: 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.

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 structure 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 (2025), 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 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 "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 4 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 3 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:

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

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 4, 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 complied. 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 5 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.[8] 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. The scientific literature (e.g., Akdel et al. (2022)) suggests that a similar

[8] 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.

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 empirically.

Why are scientists using predicted structures as an input into continued experimental structure determination, rather than accepting them simply as substitutes? 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 doing. 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 slight *different* information, and both pieces are useful.

In an effort to tease this apart, in the next section we focus on a different group of researchers: the users of these protein structures.

4.2 Spillovers to related fields of basic research

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 nearly 580,000 SwissProt indexed proteins, about 7% had an experimentally-solved structure in the PDB, and 40% had the structure of a close homolog. AlphaFold therefore represented a sudden and unexpected endowment of new structural knowledge for unsolved proteins relative to solved proteins. This 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 characteristics of SwissProt proteins that were solved vs. unsolved and shows clear differences. Previously solved structures have ten times as many papers on average published about them (excluding structure papers) in the period between 2017 and 2020, and are more than twice as likely to have at least one paper. ChEMBL activity is even more skewed, with solved structures having almost 50 times the activity. This imbalance on outcomes is not surprising if research clusters on the most biologically relevant proteins. Our difference-in-difference specification relies on a parallel trends assumption, not random assignment to the solved vs. unsolved groups. On the other hand, AlphaFold prediction confidence is very similar 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 all SwissProt-indexed proteins in years 2017 through 2024. The stark difference in levels in activity in the pre-period motivates us to use a proportional econometric model to compare percent changes in publishing activity. We therfore estimate a Poisson difference-in-differences regression for protein i in year t:

$$\mathbb{E}[Y_{it}|Unsolved_i, t] = \exp\big(\alpha + \lambda Unsolved_i + \sum_{t\neq 2021} \tau_t \cdot D_t + \sum_{t\neq 2021} \kappa_t \cdot D_t \cdot Unsolved_i\big) \quad (3)$$

where Y_{it} in this case is defined as a count of all non-structure papers. 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 avoid double-counting papers, we divide each paper by the number of linked proteins before calculating the total papers for each protein.

Alternatively, we count each protein-paper link but drop papers that link to more than 10 proteins. $Unsolved_i$ is an indicator for whether the protein had an experimentally-solved structure model in the PDB as of 2017. We focus on 2017 as the cut-off for $Unsolved_i$ because we want to observe how the literature develops through the pre-period until 2021. If we instead defined $Unsolved_i$ in 2021, then we might misattribute changes in research attention to AlphaFold that might have been caused by experimental breakthroughs in the treatment year.[9] D_t is an indicator for whether the observation occurred in year t. Standard errors are clustered at the protein level.

Figure 6 plots the year by treatment interaction coefficients (κ_t's) with standard errors. The left hand panel shows the estimates from equation 3. We find a statistically significant increase in publications about unsolved proteins after AlphaFold, suggesting a shift in attention when AI-predicted structures become available. Importantly, we also notice a slight upward drift in the coefficients in the pre-period. From 2017-2021, scientists were increasingly studying unsolved proteins in a way that is unrelated to AlphaFold, which hadn't been released yet. This upward drift threatens our identification assumption of parallel trends, but we can adjust for this pre-trend in our regression. To do this, we first estimate the Poisson regression in the pre-period only, then calculate the fitted trend component for the whole range of years. We adjust for this fitted trend in the main regression using an offset approach, which essentially tilts the coefficient series to flatten the pre-trend. This adjusted regression is presented in the right panel and shows that AlphaFold led to a significant departure from trend.

To get a more precise sense of magnitudes, we also report static Poisson difference-in-difference estimates in Table 4. Column (2) uses fractional publication counts and implies that, in the post-period, the expected publication measure for previously unsolved proteins rose by about 23% relative to previously solved proteins.[10] These results suggest that AlphaFold may have stimulated new basic research about proteins for which we previously lacked an experimental structure model.

4.3 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

[9] In practice, our results are nearly identical whether we use a 2017 or 2021 cutoff for $Unsolved_i$ because of the slow rate of progress in experimental structure research. Very few structures transition from unsolved to solved in this timeframe.

[10] Computed by taking $100 \times (e^{0.204} - 1) \approx 23\%$

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. 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 7 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 3 of Table 4 reports static Poisson difference-in-difference estimates for ChEMBL activities and we similarly find a statistically insignificant difference in pre-post activity rates for solved and unsolved proteins.

What might drive these low rates of progress in downstream research? One possibility is that these unsolved structures are simply not relevant for disease. 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. Yet, the fact that publications about these proteins increase suggests that they are not wholly uninteresting or irrelevant.

Another possibility is that these experimentally unsolved proteins remain poorly understood, even after the arrival of predicted structures. As Table 3 shows, these structures had far fewer papers written about them prior to AlphaFold. Their role in disease may simply not yet be understood, leaving their therapeutic potential untapped. Perhaps drug targeting will come later, but is currently bottlenecked by the lack of basic scientific research. If this is the case, the results in Figure 6 suggest that these bottlenecks may eventually ease.

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 complementarity will persist as prediction tools improve and expand beyond static monomer structures to complexes and context-dependent states 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 an increase in non-structure publications about previously unsolved proteins in the years after release. 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: the selection of targets, organizational and capital constraints, assay infrastructure, and the long feedback loops inherent to translational pipelines.

In our view, these findings also underscore why the medium-to-long-run consequences of AI for science remain uncertain. One possibility is that the applied impacts are simply lagged: drug discovery timelines are long, and early-stage screening activity is highly skewed toward a small subset of proteins, so it may take time for this deluge of structural information to redirect effort toward "new" targets. Another possibility is that with virtually unlimited structural information, other constraints now bind. Thinking further down the drug development pathway, clinical trials remain a formidable bottleneck. It may be the case that even vast AI-enabled improvements in efficiency along the pathway yield only modest gains drug approvals and human health in the medium-to-long-run.

References

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.* Tech. rep. National Bureau of Economic Research.

Aghion, Philippe, Benjamin F Jones, and Charles I Jones (2017). *Artificial intelligence and economic growth.* Tech. rep. National Bureau of Economic Research.

Agrawal, Ajay, John McHale, and Alexander Oettl (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.

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 (2022). *Automation after the assembly line: computerized machine tools, employment and productivity in the United States*. National Bureau of Economic Research.

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. 203–204.

— (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.

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”. *Available at SSRN 4945566*.

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. 349–365.

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”. *Equilibrium: Evidence from a Creative Goods Marketplace (February 24, 2025)*.

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

Humlum, Anders and Emilie Vestergaard (2025). “The unequal adoption of ChatGPT exacerbates existing inequalities among workers”. *Proceedings of the National Academy of Sciences* 122.1, e2414972121.

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.

Jones, Benjamin (2025). *Artificial intelligence in research and development.* Tech. rep. National Bureau of Economic Research.

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. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26223.

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.Suppl 2, 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.* University of Chicago Press / NBER.

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.

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

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.

Sullivan, Kevin P., Peggy Brennan-Tonetta, and Lori J. Marxen (2017). *Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank.* Tech. rep. Report cited in the structural biology/PDB literature; commonly referenced for replacement-cost and cost-per-structure benchmarks. Rutgers University Office of Research Analytics.

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.

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

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.

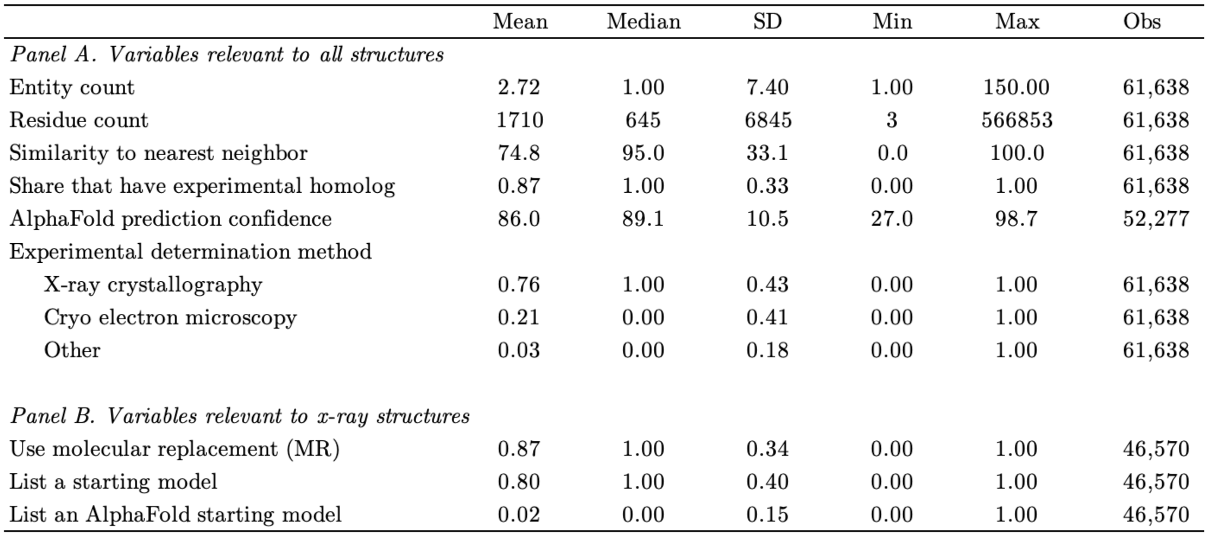
Yu, Zhengyi (2026). "The Impacts of AI at Scale: Evidence from Research Scientists". *Working Paper*.

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, pgae052.

Tables and Figures

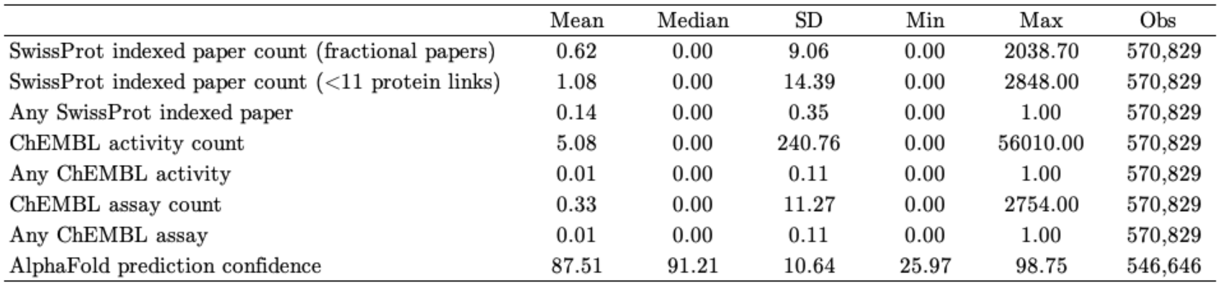
Table 1: Summary statistics: PDB structures



	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.

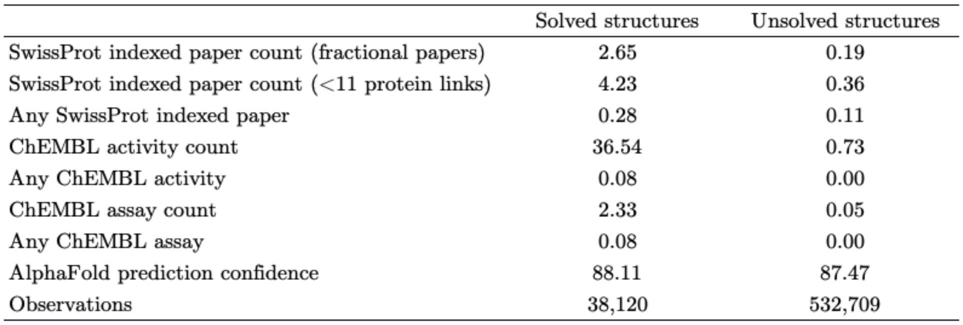
Table 2: Summary statistics: SwissProt proteins



	Mean	Median	SD	Min	Max	Obs
SwissProt indexed paper count (fractional papers)	0.62	0.00	9.06	0.00	2038.70	570,829
SwissProt indexed paper count (<11 protein links)	1.08	0.00	14.39	0.00	2848.00	570,829
Any SwissProt indexed paper	0.14	0.00	0.35	0.00	1.00	570,829
ChEMBL activity count	5.08	0.00	240.76	0.00	56010.00	570,829
Any ChEMBL activity	0.01	0.00	0.11	0.00	1.00	570,829
ChEMBL assay count	0.33	0.00	11.27	0.00	2754.00	570,829
Any ChEMBL assay	0.01	0.00	0.11	0.00	1.00	570,829
AlphaFold prediction confidence	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 variables are counts accrued over the sample period of 2017 to 2024. In the first three rows, structure papers are excluded. For fractional papers, if a paper is linked to N proteins, each protein is assigned $1/N$ of the paper. For proteins with < 11 protein links, each protein-paper is counted as a whole paper. $N = 570,829$ 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 indexed paper count (fractional papers)	2.65	0.19
SwissProt indexed paper count (<11 protein links)	4.23	0.36
Any SwissProt indexed paper	0.28	0.11
ChEMBL activity count	36.54	0.73
Any ChEMBL activity	0.08	0.00
ChEMBL assay count	2.33	0.05
Any ChEMBL assay	0.08	0.00
AlphaFold prediction confidence	88.11	87.47
Observations	38,120	532,709

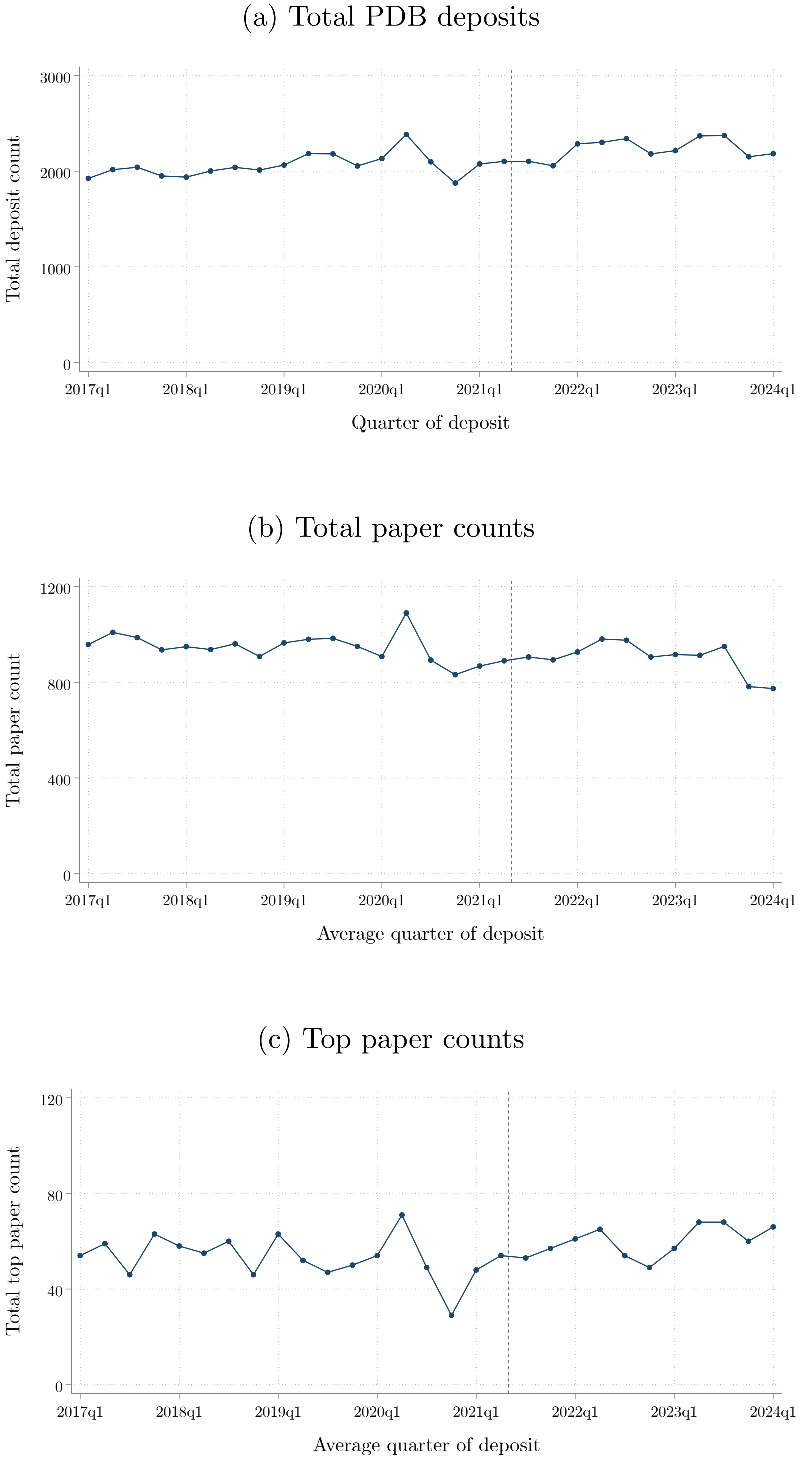
Notes: This table presents means for the proteins indexed by SwissProt, split by whether they have a solved experimental protein structure or not prior to 2017. All variables are counts accrued over the pre-period of 2017 to 2020. In the first three rows, structure papers are excluded. For fractional papers, if a paper is linked to N proteins, each protein is assigned $1/N$ of the paper. For proteins with < 11 protein links, each protein-paper is counted as a whole paper. $N = 570,829$ SwissProt-indexed proteins.

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

	(1)	(2)	(3)
Dependent Variable:	Non-structure Papers (<11 Protein Links)	Non-structure Papers (Fractions of Papers)	ChEMBL Activities
Post	-0.517***	-0.509***	-0.618***
	(0.0126)	(0.0135)	(0.0691)
Unsolved	-2.459***	-2.623***	-3.901***
	(0.0400)	(0.0419)	(0.1293)
Post x Unsolved	0.191***	0.204***	0.189
	(0.0143)	(0.0154)	(0.1222)
Observations	4,566,632	4,566,632	4,566,632

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 non-structure papers that are linked to fewer than 11 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 assay activities indexed by ChEMBL. Standard errors are clustered at the protein level. $N = 4,566,632$ SwissProt protein-years. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Figure 1: Structural biology experimental output

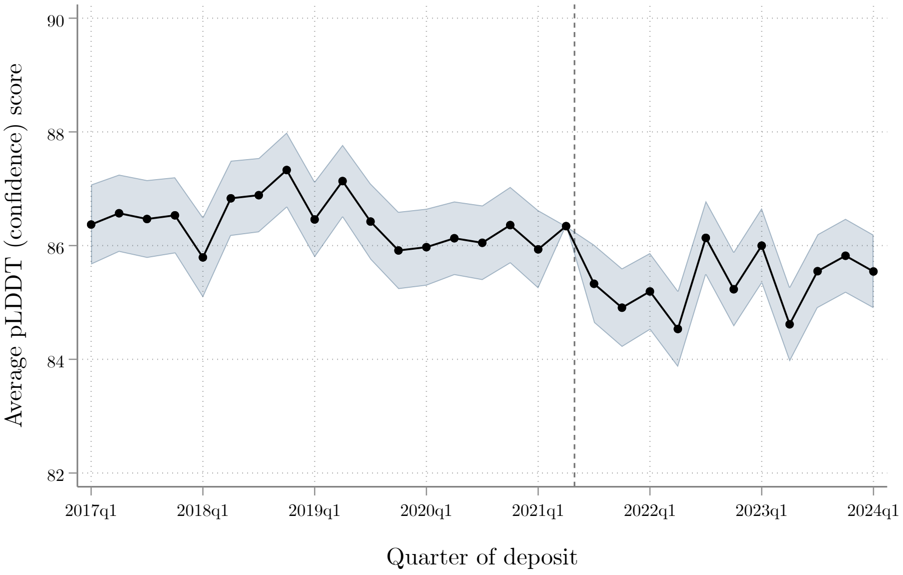

(a) Total PDB deposits

(b) Total paper counts

(c) Top paper counts

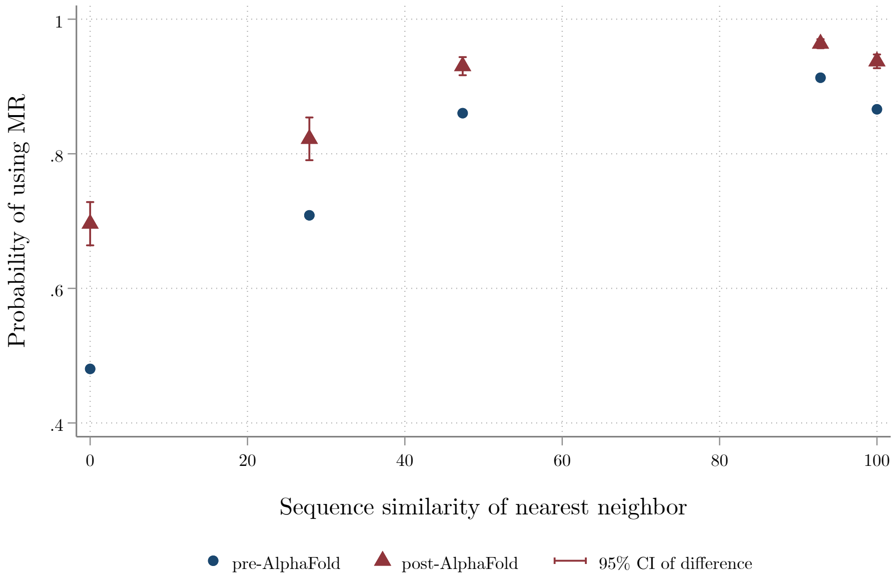
Notes: This figure shows the total counts of PDB structures and papers in our analysis sample over time. Panel (a) plots PDB deposits based on their deposit dates. Panel (b) collapses structure(s) to the paper level. If there are multiple structures per paper, the paper is assigned the average deposit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

Figure 2: Average AlphaFold confidence scores of experimentally solved structures



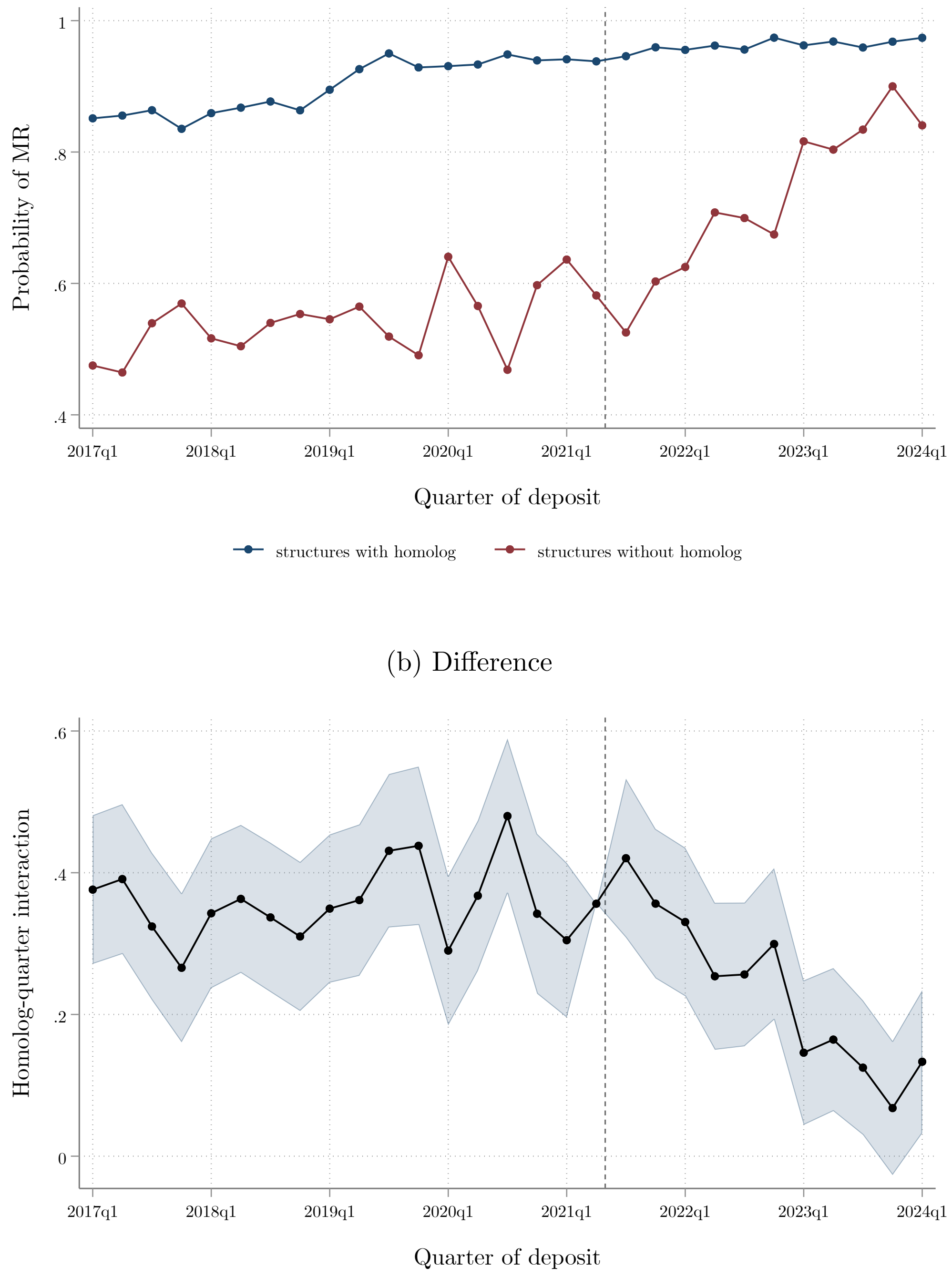
Notes: This figure takes every experimentally solved structure and finds its AlphaFold predicted analog where possible. We then plot the average confidence scores (pLDDT) over time. The shaded area represents the 95% confidence interval of the difference from the omitted quarter. $N = 52,277$ proteins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

Figure 3: 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 4: Molecular replacement usage over time



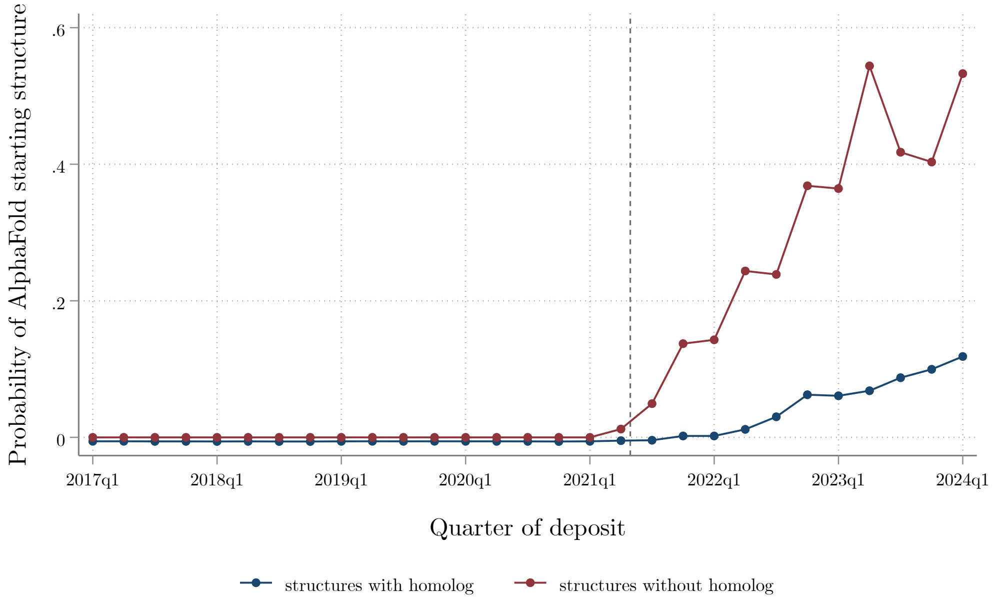
(a) Molecular replacement, with versus without homolog

(b) Difference

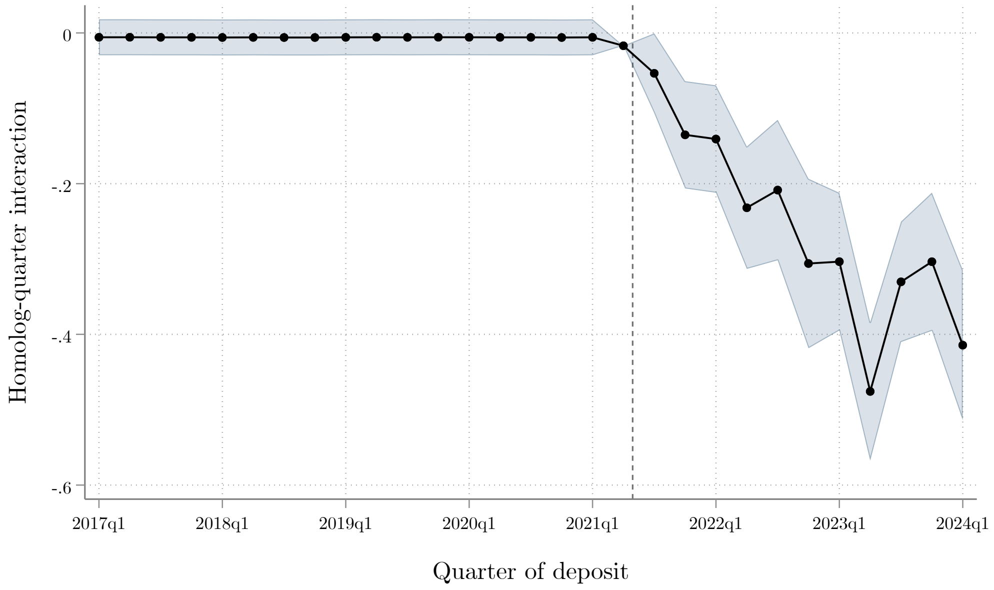
Notes: This figure shows how use of molecular replacement evolved over time, comparing proteins that did and did not have a homolog, following Equation 2 in the text. A protein with a homolog is any protein that had a released protein that was at least 30% similar in terms of amino acid sequence at the 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 5: AlphaFold usage over time

(a) AlphaFold usage, with versus without homolog

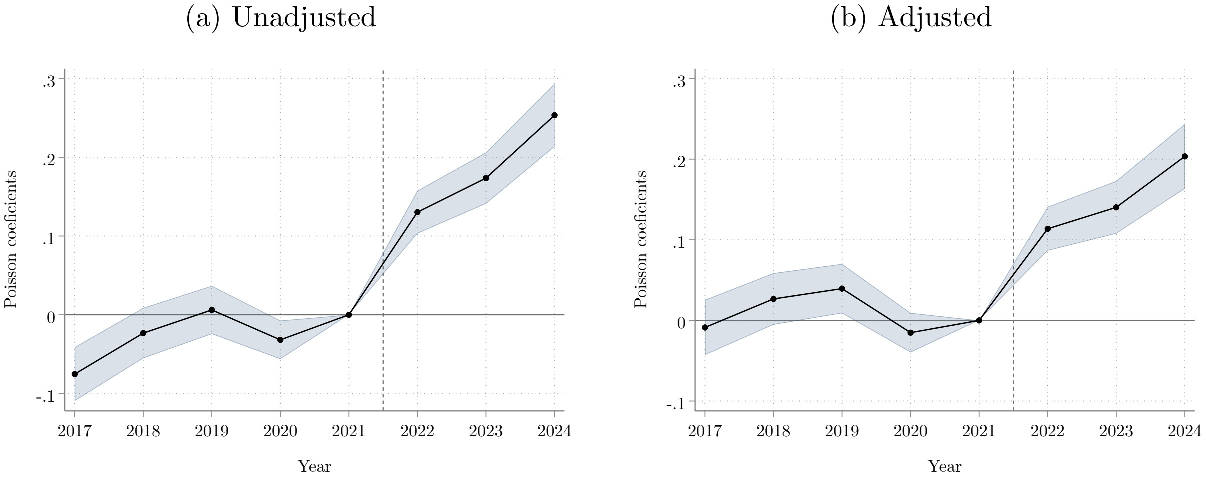


(b) Difference



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.

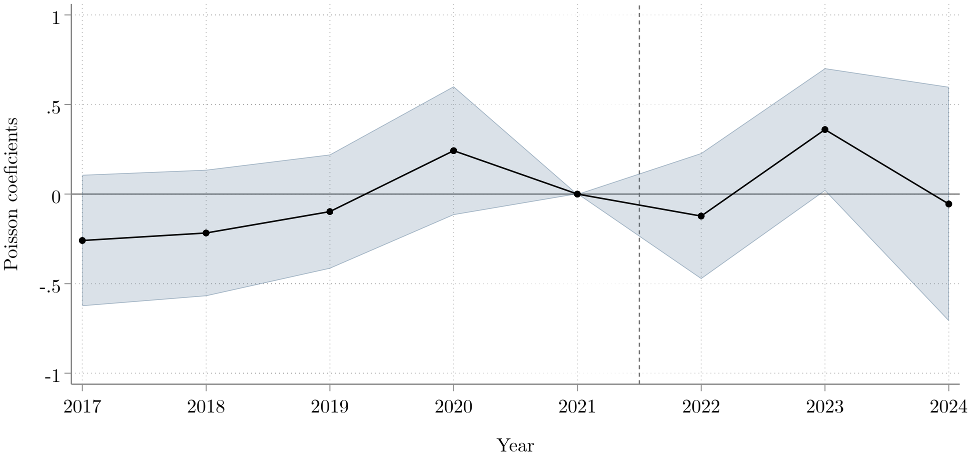
Figure 6: Non-structure publications: solved vs. unsolved proteins


(a) Unadjusted

(b) Adjusted

Notes: This figure reports the Poisson difference-in-differences coefficients comparing non-structure paper counts for solved vs. unsolved proteins before and after AlphaFold. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to 2017. The right panel reports adjusted regression coefficients to control for the pre-period trend as described in the text. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the protein level. $N = 4,566,632$ SwissProt protein-years.

Figure 7: Early-stage drug design: ChEMBL activities

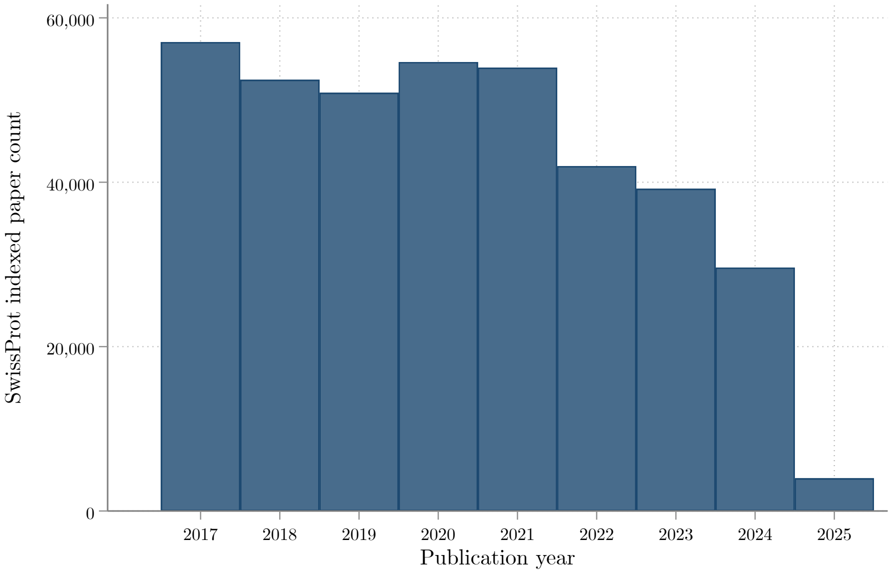


Notes: This figure reports the Poisson difference-in-differences coefficients comparing ChEMBL bioactivity activity counts for solved vs. unsolved proteins before and after AlphaFold. Solved structures are defined based on whether the protein had an experimental structure deposited in the PDB prior to 2017. Confidence intervals are shaded in blue and are calculated with standard errors clustered at the protein level. $N = 4,566,632$ SwissProt protein-years.

Appendix

A Supplemental tables and figures

Appendix Figure 1: SwissProt-indexed publications by publication year

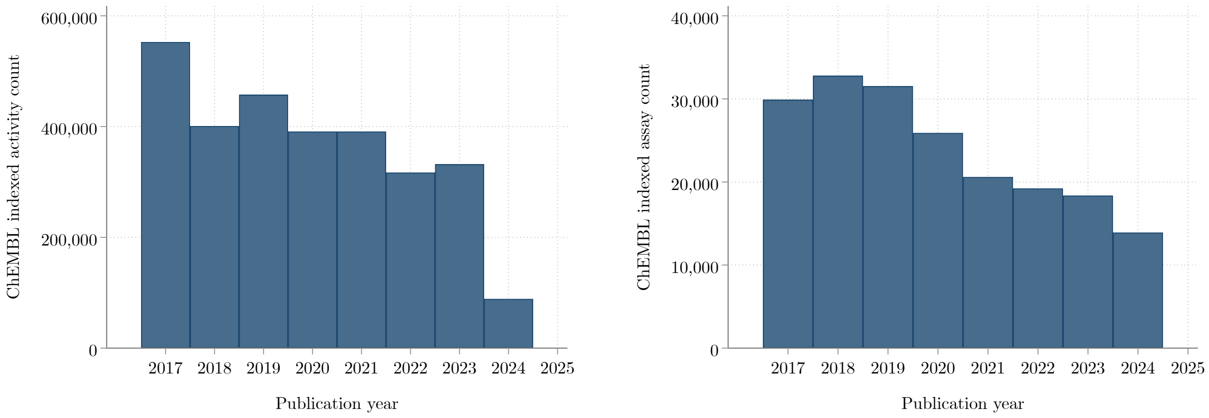


Notes: This figure shows the histogram of publication years for SwissProt-linked papers. $N = 383,625$ papers published between 2017 and 2025.

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

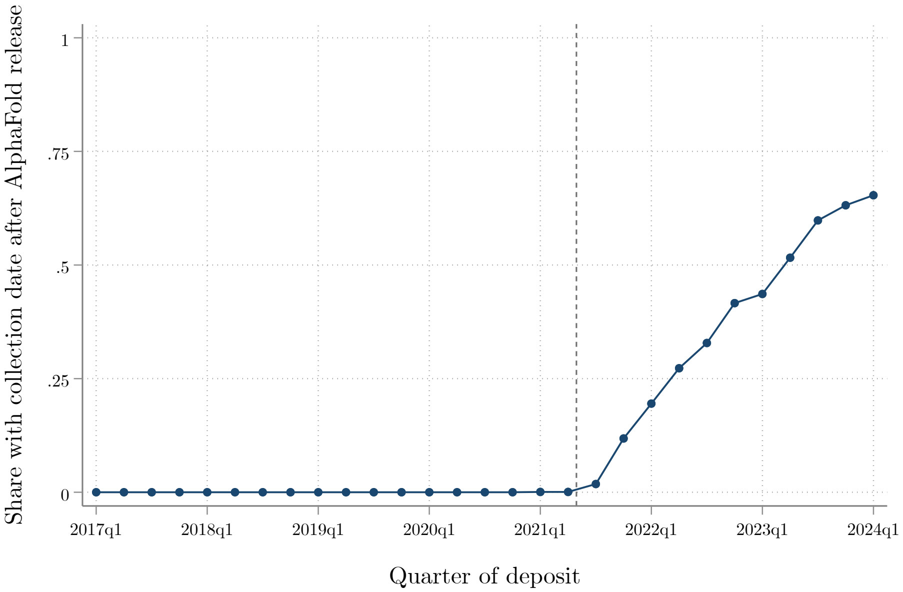
(a) Activities

(b) Assays



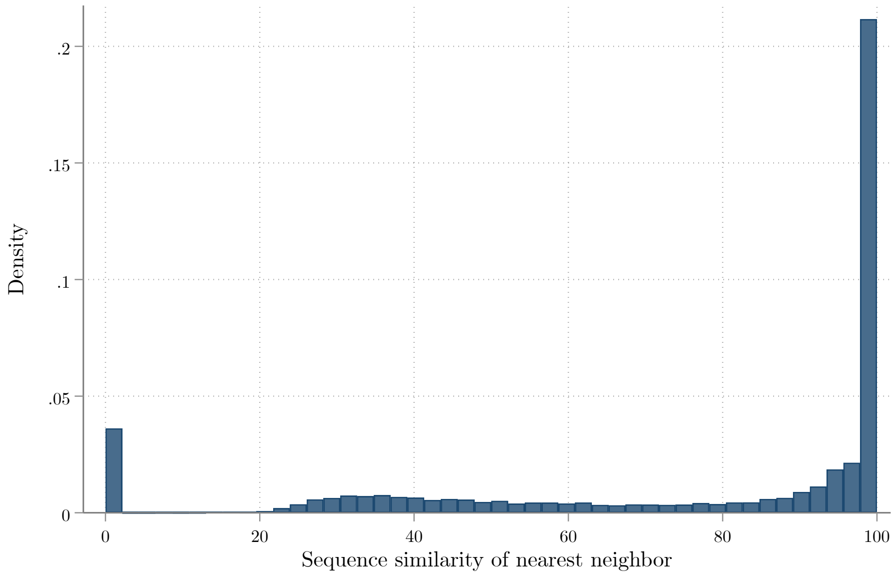
Notes: This figure shows the histogram of publication years for ChEMBL-linked activities and assays. $N = 2,930,804$ activities and $192,230$ assays published between 2017 and 2025.

Appendix Figure 3: 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 4: 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.