The Influence of AI on Scientific Labor: Evidence from Protein Structure Prediction*

Zhengyi Yu†

December 16, 2024

Abstract

This paper examines the impact of AI on scientific labor through the introduction of AlphaFold2, an AI algorithm capable of accurately predicting protein structures. While AlphaFold2 has minimal impact on the total number of publications by structural biologists, it leads to an 8% improvement in overall research quality, as measured by citation counts. Suggestive evidence indicates that adopters of AlphaFold2 experience an increase in academic performance, while non-adopters face a decline. Furthermore, the technology widens citation polarization: highly cited structural biologists are more likely to adopt AI, achieving 10% citation gains, while their less-cited counterparts see minimal effects. Additionally, I find that less-cited structural biologists reduce their focus on protein structure determination by 12%, and shift their efforts toward other areas of expertise.

*This paper was written using data obtained on July 15, 2024, from Digital Science's Dimensions platform, available at https://app.dimensions.ai.

†Department of Economics, University of Southern California, yuzhengy@usc.edu

1 Introduction

While Artificial Intelligence (AI) is revolutionizing numerous industries, there is a mixture of anticipation and apprehension within society (Acemoglu and Restrepo, 2020b). There are two major concerns related to AI development. The first one is: does AI replace humans? The second one is: does AI increase inequality? While some recent studies, such as Brynjolfsson et al. (2023), Noy and Zhang (2023), Dell'Acqua et al. (2023), and Kanazawa et al. (2022), have demonstrated that AI can augment its users, especially the low-ability workers within specific companies or experimental settings, concerns persist about its net impact on the broader labor market. For example, AI can boost the productivity of its adopters, while simultaneously marginalizing those who fail to integrate this transformative tool. This effect may be particularly pronounced in high-skill labor markets, where swift adaptability to new technologies is crucial to staying competitive.

To study the net benefit of AI on high-skill labor, this paper examines the unexpected development of AlphaFold2, a groundbreaking AI algorithm in structural biology. I find that the emergence of AI leads to an 8% increase in the overall quality of structural biologists' research, as measured by citations. However, suggestive evidence reveals a clear dichotomy beneath the net positive impact on quality: following the debut of AlphaFold2, its adopters achieve gains in both the quantity and quality of their publications, whereas non-adopters experience significant declines in academic output. My findings reveal that high-ability agents capture a larger share of productivity gains. I argue that this occurs because high-ability scientists more quickly integrate this new AI tool into their research.

Proteins are the "workhorses" of biological life, and understanding their structures lays the cornerstone for many advanced studies in life sciences. Experimentally solving a protein's three-dimensional (3D) structure often requires structural biologists to invest significant time—months or even years—and resources. In late 2020, AlphaFold2 revolutionized the field by predicting protein 3D structures directly from amino acid sequences with experimental-level accuracy[1]. The predictions from AlphaFold2 can serve as valuable hypotheses for experimentalists, accelerating structure determination and improving the accuracy of complex and novel protein models (Terwilliger et al., 2023)[2]. In mid-2021, the AlphaFold team published their AI algorithm in *Nature* and released a free database of predicted protein structures (Jumper et al., 2021). To date, its paper has garnered over 28,000 citations, according to *Google Scholar*. In 2024, David Baker and two leaders of the

[1] AlphaFold's initial 2018 release showed potential but lacked transformative accuracy. The 2020 release of AlphaFold2 marked a paradigm shift. This paper uses "AlphaFold" and "AlphaFold2" interchangeably.

[2] I will discuss the role of AI in protein structure determination in greater detail in Section 2.3.

AlphaFold team (Demis Hassabis and John Jumper) shared the Nobel Prize in Chemistry.

Several features of AlphaFold2 and structural biology make it ideal for examining the net impacts of AI on the high-skill labor market. First, AlphaFold2 is a well-established and popular tool in real-world research. Since its release in 2021, structural biologists have widely integrated AlphaFold2 into their protein structure research (Kovalevskiy et al., 2024). Second, structural biology researchers compete vigorously to determine the structure of important proteins, aiming for priority and publication in academic journals (Hill and Stein, 2019, 2021). This setting provides an opportunity to assess the net benefit of AI on the labor productivity in a competitive market. Third, academic research often tackles questions with no prior answers, providing an opportunity to test AI's role in helping knowledge workers solve open-ended problems and shape the direction of innovation.

Three major bibliometric databases enable the delineation of treatment and control groups and facilitate tracking academic outputs. Specifically, I use the PubMed database, which covers nearly all publications in the life sciences; the Dimensions database, which provides author identifiers to track individuals over time; and the Protein Data Bank (PDB), a repository for structural coordinates of biological macromolecules, primarily proteins. The PDB links protein structures to their corresponding PubMed IDs, helping to identify researchers working on protein structure determination. In addition to traditional measures of research productivity, such as publications and citations, I use text analysis of publication abstracts and keywords to assess changes in research direction.

To identify the causal impact of AlphaFold2 on structural biologists, this paper employs a "coarsened exact matching" (CEM) difference-in-differences (DiD) approach. I compare the academic records of structural biologists (treated group) with comparable life scientists in other fields (control group), pre- and post-AlphaFold2's introduction. The sample consists of Principal Investigators (PIs) in life sciences, selected based on their 2013–2016 publication records. Structural biologists are defined as PIs with at least two publications linked to PDB during 2013–2016. Control biologists are those without any PDB publications, matched to treated biologists using CEM based on comparable characteristics to ensure similarity across key attributes. My results are robust to different treatment definitions. The final sample includes $10,865$ control and $1,756$ treated biologists.

I document three major findings in this study. The first set of results shows that the availability of AlphaFold2 leads to a negligible 0.6% decline in PubMed publications but an 8% increase in citation counts for structural biologists. Correlation analysis reveals a clear divergence between structural biologists who adopt AlphaFold2, as indicated by citations of the AlphaFold team's two publications (Jumper et al., 2021; Varadi et al., 2022), and those who do not. Although AlphaFold2's free accessibility makes its adoption a voluntary

choice, event study evidence reveals that adopters and non-adopters exhibited similar trends in publications and citations prior to AlphaFold2's debut. After AlphaFold2's introduction, adopters achieved significant gains in both the quantity and quality of their research, while non-adopters experienced declines in both metrics. These findings indicate that AlphaFold2 serves as a complementary tool that enhances the overall research quality of structural biologists without significantly affecting the quantity of their work. However, the productivity gap between AI adopters and non-adopters began to widen following the release of AlphaFold2.

The second set of results examines the impact of AI on the research direction of structural biologists. Using Term Frequency-Inverse Document Frequency (TF-IDF) analysis on abstracts and keywords from PubMed publications, I find that structural biologists' research becomes 2% more novel following the emergence of AI. To further assess AI's influence on major field choices, I classify structural biologists' publications into two categories: protein structure determination and other fields. The estimates indicate that structural biologists reduce protein structure papers by 7% after the debut of AlphaFold2, without a significant increase in publications in other fields. Suggestive evidence shows that the research directions of AI adopters and non-adopters diverge following AI's debut. AI adopters produce more novel research and increase output in both protein structure determination and other fields. In contrast, non-adopters experience a slight decline in research novelty and publications outside of protein structures, along with a significant drop in protein structure determination.

The third set of results examines the impact of AI on the productivity gap between high-skill and low-skill researchers, as indicated by baseline citation levels. I first document that researchers with citations above the median adopt AI technology faster than their less-cited counterparts. This adoption gap shows no signs of narrowing over time. Overall, the citation gap between highly cited and lowly cited structural biologists widens after the debut of AlphaFold2. Highly cited researchers experience a 10% increase in total citations, while less-cited researchers show no significant effects on total citations. The availability of AI also has heterogeneous impacts on research direction for highly cited and lowly cited structural biologists. Highly cited researchers see a 2.5% improvement in research novelty, with no significant changes in protein structure determination or other publications. In contrast, lowly cited researchers show no effect on research novelty but face a 12% decline in protein structure research, partially offset by a 6% increase in publications in other fields.

This paper contributes to several literature. First, it provides empirical evidence of technology's effects on labor productivity, with a particular emphasis on AI. Economists have extensively examined the impacts of robotics and automation on the employment but the research on AI's impact on labor is still nascent (Restrepo, 2023). Many theoretical and

empirical evidence suggest that the past automation technology reduced employment rates at least in the short run (Autor, 2015; Acemoglu and Restrepo, 2018, 2020a; Feigenbaum and Gross, 2020). Contrary to earlier technology that primarily automated labor, AI demonstrates potential as a augmented tool for workers Agrawal et al. (2019). In fact, by providing AI as a complementary tool to the workers within specific companies or in experimental settings, Brynjolfsson et al. (2023), Noy and Zhang (2023), Caplin et al. (2024), Peng et al. (2023), Dell'Acqua et al. (2023) and Kanazawa et al. (2022) demonstrate that AI can indeed boost the treated labor's productivity. However, evaluating the net impacts of a policy in an experimental setting is challenging unless we randomize the proportion of the treated group across various markets, as shown by Crépon et al. (2013). Leveraging a quasi-experiment setting, this paper indicates that while the benefits to AI users occur concurrently at the expense of non-AI users, the overall quality of research output improves. Furthermore, using a rich panel of bibliographic data on researchers, I analyze outcomes over a longer horizon than previous studies, examining the dynamic impacts of AI.

Second, this paper contributes to the literature on the role of human characteristics in AI adoption and its impact on existing inequalities among workers. Unlike automation, which primarily replaces low-skill workers, AI has the potential to reduce inequality Webb (2019). While most recent empirical studies on AI support the claim that it narrows the productivity gap between high- and low-skill workers, Otis et al. (2023) and Wang et al. (2023) show that the "low-skill bias effect" is context-dependent, with AI potentially widening the gap in non-routine work. However, even if AI benefits low-skill workers more, its impact on the overall productivity gap remains ambiguous if high-ability workers are more likely to adopt AI in voluntary adoption settings, as suggested by recent survey studies (Humlum and Vestergaard, 2024; Bick et al., 2024). My study shows that AI amplifies citation polarization in scientific research. To explain the increased polarization, I find that higher-ability structural biologists are more inclined to adopt AI, as their work on complex proteins aligns more closely with the advantages AI provides.

Third, this paper also links to the growing literature about the role of technology on the productivity and direction of academic research. Furman and Teodoridis (2020) and Kim (2023) shows that the emergence of automation technology can shift the research direction. Cockburn et al. (2018) asserts that AI is transforming the process of innovation by substituting labor-intensive research with automated processes. Ludwig and Mullainathan (2024) and Agrawal et al. (2023) demonstrate that AI's prediction can generate valuable hypotheses for scientific innovation. Charness et al. (2023) and Korinek (2023) discuss the potential impact of AI on economics research and experimental design. My results show that the emergence of AlphaFold2 increases the productivity of high-skill structural biologists, while

causing low-skill researchers to reduce their focus on protein structure determination and shift to other fields of expertise. This paper provides the first causal evidence on how AI impacts scientific productivity and direction in a real world setting.

Admittedly, focusing on specialized AI applications like AlphaFold2 raises concerns about the external validity of the findings, as AI's effects may vary across disciplines. For instance, AI in healthcare often complements doctors (Agrawal et al., 2019), while AI in facial recognition may entirely replace workers who manually identify faces. Nevertheless, AlphaFold2 reflects many augmentative AI applications across professions, embodying two fundamental AI functions: (1) generating valuable hypotheses for testing and (2) simplifying labor-intensive processes. For example, when writing code, one might use ChatGPT to provide a framework (hypothesis), test its validity, and simplify debugging or improve efficiency through iterative interactions with the AI tool. Therefore, while acknowledging the specificity of this context, I argue that the conclusions drawn in this paper may have broader applicability and offer insights relevant to other fields.

2 Background

This section introduces structural biology and the development history of AI in the field.

2.1 Structural Biology

Structural biology is a branch of molecular biology, biochemistry, and biophysics that focuses on the molecular structure and dynamics of biological macromolecules, particularly proteins. Proteins are composed of a linear sequence of amino acids that fold into a 3D structure. Proteins support virtually all biological processes: from transporting oxygen in the blood to sensing the color in the eyes. The 3D structure is key to understanding how proteins interact, function biologically, and cause diseases when they malfunction. Although scientists often understand the amino acid sequence of proteins, they cannot easily deduce the 3D structure from the sequence. Determining the structure typically takes months to years and requires substantial human and financial resources. The journey to solve protein structures started in the 1950s, with the first successful determination of a protein's structure: muscle's myoglobin in 1957. Over the past decades, scientists in structural biology have earned more than a dozen Nobel Prizes, highlighting the field's importance to advanced biological research and medicine design.

2.2 Protein Structure Prediction

Since there is a one-to-one correspondence between a protein's 1D amino acid sequence and its primary 3D structure, the computational prediction of the protein structure has been a pursuit in structural biology for over five decades (Editorial, 2022). Various software and techniques were developed prior to the advent of AlphaFold2 but, limited by computational power and algorithms, they often lacked the accuracy and efficiency needed for practical applications. In 2018, by employing deep learning, AlphaFold1 was able to predict protein structures with enhanced accuracy and won first place in the ranking of the 13th Critical Assessment of Structure Prediction (CASP13)[3]. AlphaFold1's accuracy exceeded 60% in 2018, which was still far from the transformative threshold of 90%.

AlphaFold2 achieved a significant leap in performance in 2020. On November 30, 2020, AlphaFold2 was declared the winner of CASP14, demonstrating an astounding accuracy rate of over 90%. Panel A in Figure 1 shows the comparison between AlphaFold2's in silico predictions and experimentally determined structures in CASP14. AlphaFold2's achievement surpassed most scientists' expectations and was hailed as solving the 50-year-old "protein-folding problem." The AlphaFold team published the algorithm in Nature on July 15, 2021 (Jumper et al., 2021). Additionally, DeepMind open-sourced AlphaFold2's code and offered the scientific community free access to over 350,000 predicted protein structures. Researchers simply need to cite the papers by AlphaFold team if they utilize the data or methodology (Jumper et al., 2021; Varadi et al., 2022). In July 2022, this database underwent a dramatic expansion, growing to include over 200 million protein structures, encompassing nearly all known proteins.

Shortly after the debut of AlphaFold2, RoseTTAFold, a similarly accurate protein prediction software, was introduced by Baker Lab. RoseTTAFold's paper (Baek et al., 2021) was published on the same day as AlphaFold2's in Science. Given the near-simultaneous timelines and comparable accuracy of both AlphaFold2 and RoseTTAFold, the introduction of the latter does not skew our analysis of AlphaFold2's impact. Notably, since the RoseTTAFold team did not release a comprehensive database akin to AlphaFold's, the latter's influence remains significantly larger. This is the primary reason for this paper's focus on AlphaFold2's impact on structural biologists. In fact, in 2024, David Baker, Demis Hassabis, and John Jumper jointly received the Nobel Prize in Chemistry.

Despite the high accuracy of structural predictions, AI cannot fully replace human structural biologists (Terwilliger et al., 2023). Panel B of Figure 1 illustrates that the number of protein structure papers remains stable after the debut of AlphaFold2 in top 3 natu-

[3]CASP is a worldwide competition held every two years since 1994 to foster the study of computational prediction of protein 3D structures.

ral science journals: *Cell*, *Nature*, *Science* (CNS) and journals publishing protein structure research. Two key reasons underscore the continued need for human characterization of protein structures. First, like any predictive AI, AlphaFold2 is not 100% accurate for all protein structures, and experiments conducted by structural biologists remain essential to validate its *in silico* predictions. For example, since AlphaFold2 is trained by the existing protein structure, its accuracy can be diminished for proteins that have little similarity to previously resolved structures, such as certain membrane proteins[4] (Azzaz et al., 2022). Second, AlphaFold2's role is confined to predicting the shape of 3D structures, while structural biologists' work extends beyond this to explore the functions and bio-mechanics explanation underlying protein structures.

2.3 Structural Biology with AI

The debut of AlphaFold2 generated significant excitement among structural biologists by enabling research on complex and novel proteins that were previously difficult to resolve (Subramaniam and Kleywegt, 2022). AlphaFold2 has rapidly become a widely adopted tool in protein studies. Panel C of Figure 1 shows that over 40% of protein structure papers in the top three natural science journals cited AlphaFold2 in 2023.

Before AlphaFold2, structure determination studies involved four major steps (see Panel A in Figure 2). First, structural biologists need to identify functional and stable regions of the protein to produce recombinant protein domains. This step often relies on expert judgment and trial-and-error, which can take months or even years. Second, structural biologists use techniques such as X-ray crystallography or cryo-electron microscopy (Cryo-EM) to gather information about the protein architecture. Third, structural biologists construct a 3D protein model to match the experimental results. This step is time-intensive and requires iterative refinement to ensure accuracy. Finally, structural biologists analyze the structural data to identify active sites, interaction interfaces, or conformational dynamics that explain the protein's biological role.

Panel B of Figure 2 illustrates the typical workflow for protein structure determination with the help of AI. There are two steps in the workflow where AlphaFold2 can involve. First, AlphaFold2's structural predictions play a crucial role as hypotheses, offering researchers a starting point to guide experimental validation and refinement (Goulet and Cambillau, 2022; Kovalevskiy et al., 2024). AlphaFold2's predictions offer structural hypotheses for target proteins, enabling structural biologists to pinpoint potential functional areas and

[4]Membrane proteins are important and comprise approximately $20-30\%$ of all proteins in most organisms, yet they represent only about 2% among all the solved protein structures. This disparity is due to their hydrophobic nature, complex folding patterns, and difficulties in crystallization

avoid unstable domains. By leveraging these hypotheses, structural biologists can design and produce recombinant proteins more efficiently, reducing the time and resources needed.

Second, AlphaFold2 accelerates the process and improves the quality of 3D modeling for both X-ray crystallography and Cryo-EM. X-ray crystallography researchers first crystallize the protein, then subject it to X-ray diffraction, and finally model its three-dimensional structure based on the resulting diffraction patterns. This method, however, often encounters the "phase problem," where information loss during diffraction complicates reconstruction. While molecular replacement (MR) can address this issue, successful MR typically requires a previously solved structure of a similar protein with at least 30% amino acid sequence similarity (Kim, 2023). With AlphaFold2's ability to predict structures for all proteins, structural biologists can now reconstruct phases for novel proteins that lack similar solved structures (McCoy et al., 2022).

Although Cryo-EM does not face the phase problem, AlphaFold2 aids experimentalists in interpreting lower-resolution electron density maps. Cryo-EM researchers observe protein structures by rapidly freezing samples and using electron beams to capture high-resolution images of proteins in their near-native states. Recent advancements in Cryo-EM have significantly improved the efficiency and resolution of structural determination, enabling researchers to resolve complex and novel proteins. However, challenges persist, particularly in achieving high resolution in certain regions due to data collection limitations, which can affect the accuracy of 3D models. By providing high-confidence structural models, AlphaFold2 allows scientists to refine experimental data with greater precision, enhancing the accuracy and completeness of protein structures.

The examples above demonstrate that the use of AlphaFold2 in protein structure research must be integrated with researchers' active efforts. Therefore, users' expertise with proteins is crucial to fully unlocking the tool's potential benefits. In addition to its benefits for structure determination, AlphaFold2 has opened new research avenues for structural biologists, including protein design, protein interactions, and drug discovery. Furthermore, researchers can develop new AI models in various domains based on AlphaFold2's framework. In mid 2024, the AlphaFold team introduced AlphaFold3, a new AI algorithm designed to predict interactions among all biological molecules, with the goal of accelerating drug discovery (Abramson et al., 2024). AlphaFold3 is open-source for non-commercial applications, and its role in the pharmaceutical industry still requires validation.

3 Data, Sample, and Empirical Strategy

3.1 Data Sources

This paper concentrates on the academic performance of scientists in the fields of life science in the period $2017 - 2023$. I mainly use the following databases to collect the data for my study.

3.1.1 PubMed

PubMed is a freely accessible search engine that interfaces with the MEDLINE database (Medical Literature Analysis and Retrieval System Online), encompassing the bibliographic information of a comprehensive collection of life science journals. Within PubMed, each scientific paper is assigned a unique PMID (PubMed ID). Additionally, PubMed categorizes publications through the use of MeSH (Medical Subject Headings) terms, which are standardized keywords representing the scientific fields to which the papers pertain. It is noteworthy that these MeSH terms are assigned by database indexers rather than the authors of the papers, ensuring a consistent and controlled vocabulary across the literature. However, PubMed does not systematically offer an ID to identify authors unambiguously.

3.1.2 Dimensions

In order to trace an author's publication history unambiguously, I leverage the Dimensions database (https://www.dimensions.ai/). This extensive database provides metadata of 150 million publications including author affiliation, publication, citation counts, journal, abstracts, patents and funding. Importantly, Dimensions also offers the author identifier, which enables the unambiguous tracking of an author's publication history, thus facilitating an accurate examination. Furthermore, Dimensions extracts the bibliographic information from various existing databases including PubMed. Therefore, it is convenient to link an author's Dimensions author ID to her publications' PMID. Previous works, such as Jia et al. (2022), have demonstrated the efficacy and reliability of the Dimensions database.

3.1.3 Protein Data Bank

Established in 1971, the Protein Data Bank (PDB) serves as a central repository for protein structures, housing over 200,000 structures as of 2024. It provides detail characteristics and quality of the protein structure, such as length, weight, the amino acid sequence and resolution. The standard protocol for publishing a protein structure paper entails that

authors deposit their characterization of the 3D structure in the PDB shortly before or concurrent with the submission to the academic journals. As a result, the PDB comprises the comprehensive collection of all publicly available protein structures. Once the paper is published in a journal, the PDB links its PMID to the corresponding deposited structure. For the remainder of the paper, I will refer to papers linked to the PDB as "PDB papers," which primarily focus on protein structure research.

3.2 Sample Construction

To evaluate AlphaFold2's influence on the scientific productivity of structural biologists, I construct an individual-level balanced panel comprising structural biologists and comparable control life scientists based on two main criteria. First, control researchers do not engage in protein structure determination but conduct research in other life science fields. Second, control and treated biologists must have comparable key baseline characteristics, including citations, publication age, coronavirus research, and affiliations. The investigation period spans from 2017 to 2023, with 2021 serving as the initial treatment year, following the emergence of AlphaFold2 in late 2020.

To follow the first criterion, I identify journals that published at least two PDB papers during 2013–2016 and query Dimensions to retrieve all PubMed articles published in these journals during the same period. I define structural biologists as last authors[5] of at least two PDB papers from this set, while non-structural biologists are last authors of at least two articles from the set but without any PDB papers. This approach ensures that both groups published in similar academic venues and makes sure no new entrants during the investigation period (2017-2023). I identify $2,441$ structural biologists and $104,129$ non-structural biologists in the author list. Using Dimensions, I collect their complete publication records from 2017 to 2023. I downloaded the data in July 2024.

To meet the second criterion, I implement a "coarsened exact matching" (CEM) procedure (Iacus et al., 2012) to match structural biologists with non-structural biologists based on several key ex ante variables. Although the Difference-in-Differences (DiD) design does not require balanced covariates between treated and control groups, comparable treated and control groups enhance the causal inference of the study. Before the CEM process, I perform several data trimmings. First, I exclude the coauthors of structural biologists to minimize potential spillover effects. Second, I drop control authors who published PDB papers during the pre-treatment period (2017–2020). Third, I remove authors with zero publications during the pre-treatment period to focus on active researchers.

[5]In life sciences, the last author of a paper is conventionally the Principal Investigator (PI).

For the CEM, I begin by performing an exact match on authors' baseline affiliations during 2017–2020, ensuring comparisons between structural and non-structural biologists within the same institution. Additionally, I balance the following three key aspects using coarsened bins: 2016–2020 citations, academic age, and citations of COVID-related papers during 2017–2020. I successfully matched 72% of the structural biologists and details of the matching process are provided in Appendix B. The final panel dataset consists 10,865 control and 1,756 treated biologists.

The overlapping timeline of the COVID-19 pandemic with my investigation period poses a potential threat to the validity of my estimation. However, the CEM procedure described above mitigates some of the potential confounding effects of COVID-19. First, I use exact matching between treated and control biologists based on their affiliations, addressing potential biases stemming from variations in COVID-19 policies across institutions. Second, I perform coarsened matching on baseline coronavirus-related citations to account for potential biases driven by the focus on COVID-19 research.

3.3 Text Similarity

To measure the novelty of the papers, I first query Dimensions for the abstracts and MeSH terms from PubMed papers for the authors in the sample. Then, I process the text by removing common English stop words and punctuation. Finally, I apply a text analysis method to transform the cleaned text into vector representations, as outlined below.

3.3.1 TF-IDF

I use Term Frequency-Inverse Document Frequency (TF-IDF) to represent the text as sparse vectors. Specifically, I apply one-gram TF-IDF, which divides the text into single terms (unigrams) and evaluates their importance within a document relative to a corpus. Term frequency (TF) quantifies how often a term appears in a document, while inverse document frequency (IDF) downweights terms that are frequent across the entire corpus. The product of TF and IDF assigns higher weights to terms that are more informative for distinguishing between documents. The output is a high-dimensional vector for each document, where each dimension corresponds to the weighted importance of a term.

3.3.2 Novelty Measure

To measure the similarity between two papers, I calculate cosine similarity using the vector representations generated by TF-IDF. Cosine similarity measures the cosine of the angle between two vectors, offering a robust metric for textual similarity. This metric effectively

captures the degree of similarity between documents while remaining invariant to differences in vector length. The similarity scores range from 0 to 1, where 1 indicates identical vectors and 0 indicates orthogonality. For ease of interpretation, I define the distance as $1 - similarity$ for each pair of papers.

Next, for each author, I compute the average pairwise distance between all papers in each year t after 2016 and the baseline pool of papers from 2013 to 2016. This measure captures how distinct papers in a given year after 2016 are from those in the baseline period. A higher value indicates greater novelty in research during year t relative to the baseline years of 2013-2016. If an author does not publish any papers in a given year, the distance for that year is set to zero, reflecting the absence of (new) research output.

3.4 Summary Statistics

Table 1 reports the annual average characteristics during 2017–2020, categorized by structural biologists and the control group. Scientific productivity is measured using two outcome variables: PubMed publications and citations. PubMed publications represent the total number of researcher i's publications indexed by PubMed in year t, while PubMed citations measure the total citations received by all of researcher i's PubMed publications from year t up to the data collection date. PubMed publications capture the quantity of a researcher's output, while PubMed citations reflect the quality and influence of their work. Structural biologists, on average, produce 0.4 more publications per year than control biologists, but both groups exhibit similar baseline citation levels.

Structural biologists' research is more novel at baseline, as indicated by TF-IDF analysis. To measure the number of researchers involved in a paper, I count its total number of authors. The number of coauthors represents the total count of coauthors across all of researcher i's PubMed publications in year t. Academic age is defined as the difference between 2020 and the year of a researcher's first publication. Both the treated and control biologists exhibit similar academic age and coauthor counts.

The majority of both treated and control researchers are affiliated with universities. To assess the extent of a researcher's work related to COVID, I measure the number of citations an author receives from papers with MeSH terms containing COVID-related keywords such as Coronavirus, SARS-CoV-2, and COVID-19. Since I did not enforce one-to-one matching in the CEM process, differences in group means may arise even for covariates being matched. For example, while I perform exact matching on researchers' affiliations, there are more control biologists from universities because more matched control biologists are assigned per treated biologist in universities than in other organizations.

Since the AlphaFold team requires any research utilizing their database or algorithm to cite their papers (Jumper et al., 2021; Varadi et al., 2022), I use citations of AlphaFold2 papers as a proxy for AlphaFold2 usage. Approximately 40% of structural biologists have cited AlphaFold2 in their research, compared to less than 3% of control biologists. This disparity indicates that AlphaFold2 is significantly more relevant to structural biologists.

3.5 Empirical Strategy

Since the outcome variables used to measure the impact of AlphaFold2 on structural biologists are count data, such as publications and citations, I employ a Poisson DiD regression for the main analysis, following standard practices in the economics of science literature.

$$E[y_{it}|X_{it}] = \exp\left[\sum_{\substack{t=2017 \\ t\neq 2020}}^{2023} \beta_t \cdot \text{Structural_Biologist}_i \times Post_t + \alpha_i + \gamma_t\right] \tag{1}$$

Y_{it} represents the outcomes of interest, such as the number of PubMed publications and citations in year t for author i. $\text{Structural_Biologist}_i = 1$ if author i belongs to the treatment group, and $\text{Structural_Biologist}_i = 0$ otherwise. $Post_t$ is a dummy variable equal to 1 for post-treatment periods ($t > 2020$) and 0 otherwise. α_i denotes the individual fixed effect, capturing time-invariant characteristics of each author, and γ_t represents the year fixed effect.

4 Main Results

4.1 Science Productivity

Figure 3 presents the event-study DiD estimates for two key measures of productivity: the number of PubMed publications and citations. Panel A reports the estimates based on the preferred definition of treatment, which is a dummy variable representing all structural biologists in the sample. The results indicate that the total number of PubMed publications by structural biologists remains unchanged before and after AlphaFold2's introduction. In contrast, the availability of AlphaFold2 leads to a noticeable increase in PubMed citations for structural biologists, with no evidence of pre-trends prior to the emergence of AI. In fact, the positive effect on citations continues to grow, peaking in 2023, suggesting that the observed impact is unlikely to be confounded by the COVID-19 pandemic. This pattern also aligns with the adoption of AlphaFold2 and the roll-out of AlphaFold2 database, as illustrated in

Panel C of Figure 1.

Panel B demonstrates the robustness of these results to alternative treatment definitions. Given that structural biologists often work across various research fields beyond structure determination, I measure treatment intensity using the proportion of PDB papers within a structural biologist i's total publication profile from 2013 to 2016. Table A1 shows the distribution of treatment intensity among structural biologists. A structural biologist is classified as high impact if they have a higher-than-median percentage of PDB papers, and low impact if they have a lower-than-median percentage. Additionally, I use the percentage of PDB papers as a continuous treatment variable.

Across all three treatment definitions, there is no observed effect on publication counts. However, the availability of AlphaFold2 increases citations for structural biologists under both the continuous and high-impact treatments, with no evidence of pre-trends prior to the treatment. In contrast, low-impact structural biologists exhibit minimal effects on citations. Appendix Figure A1 displays a similar pattern for total publications and citations, including both PubMed and non-PubMed research by structural biologists.

I present the pooled DiD estimates using various treatment definitions in Table 2. I exclude observations from 2021 in the pooled DiD estimation for two reasons. First, the AlphaFold team released their paper, algorithm, and database in mid-2021. There is often a lag between the adoption of new technology and its reflection in publications, making it premature to measure the impact within the same year. Second, the COVID-19 pandemic persisted throughout 2021, and excluding this year helps mitigate potential confounding effects of the pandemic on the results.

Columns (1) to (4) indicate no significant impact of AI on publication counts across different treatment definitions. Column (5) indicates that, on average, AlphaFold2's debut increased citations for structural biologists by $e^{0.0802} - 1 = 8.4\%$. The effect is slightly larger for high-impact structural biologists, who experience an increase of approximately 12%. For low-impact structural biologists, the increase is marginally significant at 6%. The continuous treatment reveals the largest effect, with an estimated 20% increase, reflecting the impact of moving from a 0% to 100% share of PDB publications. These results show that AlphaFold2 enhances the overall quality and influence of structural biologists' research but has minimal effect on its overall quantity.

4.2 AI user vs Non-AI user

The difference-in-differences estimate captures the net impact of AI on structural biologists, but not the treatment effect of AI adoption since the adoption of AlphaFold2 in research is

voluntary. However, I present the correlation relationship between the adopter of AlphaFold2 and their academic performance. Citations of AlphaFold2's publications serve as a suggestive indicator of its usage, as the AlphaFold team requires researchers using their results cite their papers (Jumper et al., 2021; Varadi et al., 2022). By cross-referencing the list of PubMed papers citing AlphaFold2 with the publications of structural biologists in my sample, I can identify those who have cited AlphaFold2. A structural biologist is classified as an AI user if they have cited AlphaFold2's publications.

Figure 4 presents the event-study estimates from Equation 1 comparing AI-adopting and non-adopting structural biologists. Panel A shows that both groups exhibit relatively flat patterns in the number of publications prior to AlphaFold2's debut but diverge significantly afterward. Non-adopters experience a decline in publications, while adopters see an increase. I further classify AI users into two groups: early adopters and late adopters. Early adopters are structural biologists who cited AlphaFold2's papers before 2023, while late adopters are those whose first citation of AlphaFold2 appears in 2023. For early adopters, publication increases begin in 2021, while late adopters begin increasing their publications in 2023. Panel B shows a similar pattern in total PubMed citations between AI users and non-AI users.

These estimates need to be interpreted cautiously and should not be construed as the causal effect of AlphaFold2 adoption or non-adoption on structural biologists. However, the results suggest that, beneath the net complementary effect of AI on human researchers lies a pronounced dichotomy between adopters and non-adopters of this new technology.

To further explore how AI supports protein research, Table 3 analyzes the correlation between a PDB paper's protein structure characteristics and its citation of AlphaFold2. All regressions include individual and year fixed effects to account for potential differences between AI adopters and non-adopters, as well as variations across years. The outcome variable in column (1) is a dummy indicating whether the PDB paper contains a novel protein structure[6]. The estimates show that papers citing AlphaFold2 are 22% more likely to solve a novel protein structure. Columns (2) and (3) report results for the natural log of maximum protein length and maximum protein weight in a PDB paper[7], respectively. The findings indicate that protein structures aided by AlphaFold2 are 17% longer and heavier. These results suggest that the use of AlphaFold2 is associated with larger proteins and those that have not been previously solved.

Following Hill and Stein (2021) and Kim (2023), this paper uses R-free and resolution as indicators of the quality of protein structures. R-free, a measure from crystallography,

[6]A protein structure is classified as novel if no previously solved structure exists with at least 30% amino acid sequence similarity.

[7]A PDB paper can contain several protein structures. Therefore, I select the max length and weight of the protein structure in a PDB paper.

evaluates the accuracy of a protein structure model using a subset of reflection data excluded from refinement; lower R-free scores indicate a better model fit to X-ray diffraction data. Resolution, expressed in angstroms (Å), represents the level of detail visible in a molecule's determined structure, with lower values signifying higher resolution and more detailed structural images.

Column (4) in Table 3 reports results for the log-transformed resolution (ResolutionÅ), while column (5) examines the log of the R-free score, limited to structures solved by X-ray crystallography. For PDB papers containing multiple proteins, I select the ResolutionÅ and R-free values for the longest protein. Since more novel, longer, and heavier proteins are more complex to solve, these structures often exhibit lower quality. To account for this, columns (4) and (5) control for protein characteristics from columns (1) to (3). The results show no significant impact of AlphaFold2 on protein structure quality. However, column (6) reveals that PDB publications using AlphaFold2 are correlated with 30% more citations.

These correlations suggest that structural biologists tend to use AlphaFold2 to study more complex proteins, supporting the scientific community's view that AI enables research on proteins that were previously too complex to solve (Subramaniam and Kleywegt, 2022). However, this evidence also implies that encouraging non-adopters to adopt AI is not a straightforward solution, as AI usage is often linked to the study of more complex proteins, which demands expertise in these areas.

4.3 Research Direction

This section examines how AI influences the research direction of structural biologists. Panel A in Figure 5 presents the event-study estimates for the novelty of research publications, measured using TF-IDF on researchers' abstracts and MeSH terms. Since TF-IDF is not count data and includes zeros, I use an OLS regression on the level of TF-IDF. Panel A shows that the availability of AlphaFold2led to a net increase in novelty, as measured by TF-IDF, with no significant pre-trend observed. Column (1) in Table 4 indicates that AlphaFold2's availability resulted in a 2% increase in distinct research by structural biologists. It is important to note that the novelty measurement in this case compares an author's research to their own past work, rather than to the work of others. Therefore, these results suggest that AI induce structural biologists to explore areas they had not previously researched.

Although TF-IDF can detect differences when researchers introduce new terms or words in their abstracts, it cannot clearly distinguish whether structural biologists are shifting from structure determination to other fields or pursuing more innovative research within protein structure. Both scenarios are plausible. The advent of AlphaFold2 might enable researchers

to determine more novel proteins or explore areas beyond proteins, such as drug discovery. Alternatively, it could push biologists to shift to other fields, as AlphaFold2 renders certain aspects of protein structure research less essential.

To assess the impact of AlphaFold2 on structural biologists' decisions to continue protein structure research, I categorize their publications into two groups: PDB publications, which directly solve protein structures, and other PubMed publications. Since control biologists have no PDB publications in the pre-treatment period by design, I use all PubMed publications of control authors to compare with both the PDB and other PubMed publications of structural biologists.

Panel B in Figure 5 shows that, on average, structural biologists slightly reduce the number of PDB publications after the debut of AlphaFold2. Panel C indicates that the number of other PubMed publications by structural biologists remains relatively stable before and after AlphaFold2's introduction. Column (2) in Table 4 reports a 7% decline in PDB publications by structural biologists, while Column (3) shows an insignificant 3% increase in research on other fields. As previously noted, all estimates in Table 4 exclude observations from 2021.

The event studies in Panels A to C of Figure 5 highlight a clear divergence in research direction between AI users and non-AI users. AI users pursue more distinct and innovative research, increasing both PDB publications and publications in other fields after AlphaFold2's debut. In contrast, non-adopters experience declines in research novelty, PDB publications, and publications in other fields. Once again, these results should not be interpreted as the causal impact of using or not using AI on structural biologists. However, these findings align with the increasing popularity of AlphaFold2 usage in PDB papers, as shown in Panel C of Figure 1, and support the widely held view in the academic community that AI augmented research will shape the future of structural biology studies (Subramaniam, 2023).

4.4 AI Adoption and Distributional Effect

As discussed earlier, the skill-biased effects of AI on labor depend on the nature of the tasks AI supports. Previous literature shows that AI can reduce the productivity gap between high- and low-skill workers in tasks like writing, customer support, or taxi driving (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Kanazawa et al., 2022). However, recent studies also suggest that AI can widen the productivity gap in more non-routine jobs, such as running small businesses (Otis et al., 2023). In this context, we examine whether AI amplifies the performance differences between high- and low-skill agents in scientific research.

4.4.1 Who Are More Likely Using AI?

As shown above, the correlation results suggest that AI adopters and non-adopters exhibit significantly different performance patterns after the debut of AI. A natural question arises: is there a difference in AI adoption between high-skill and low-skill researchers? To measure a researcher's skill level, I use their baseline citation count from 2017 to 2020 while controlling for age[8]. Specifically, I group structural biologists into mutually exclusive five-year age bins. Researchers with baseline citations above the median within their age group are classified as high-skill, while those below the median are classified as low-skill.

Figure 6 shows that by 2022, 30% of highly cited researchers had cited AlphaFold2, compared to only 10% of lowly cited researchers. By 2023, half of the highly cited researchers had adopted AlphaFold2, while the adoption rate among lowly cited researchers remained below 30%. These patterns indicate that highly cited researchers adopt the new AI technology at a faster rate, with a persistent and significant gap in adoption rates between highly cited and lowly cited researchers over time.

To further investigate why highly cited researchers adopt new technology more quickly in a voluntary setting, Table 5 analyzes the *ex ante* characteristics of structural biologists that may influence their decision to cite AlphaFold2 in their research. All columns present the results of OLS regressions, where the dependent variable is a binary indicator equal to 1 if an author cites AlphaFold2. Independent variables include $HighCitation$, a dummy variable equal to 1 if an author is highly cited as defined earlier; $HighNovelty$, which equals 1 if a researcher's percentage of novel proteins is above the median; $HighLength$, which equals 1 if the average length of proteins solved by a structural biologist is longer than the median; and $HighWeight$, defined similarly to $HighLength$, but for the average weight of proteins. Additional variables include $CryoEM$, which equals 1 if the structural biologist used cryo-EM in their protein structure research, and $HighAge$, indicating authors whose academic age is above the median. The analysis also considers whether an author is affiliated with a university. For university-affiliated researchers, I use the *QS World University Rankings*[9] from 2020 to distinguish those affiliated with universities ranked above the median.

Column (1) shows that highly cited structural biologists are 26% more likely to cite AlphaFold2 in their research compared to their less-cited counterparts. Consistent with the correlation between AI usage and complex proteins, columns (2) to (4) indicate that structural biologists solving more complex proteins at baseline are more likely to cite Al-

[8]The results are similar without controlling for age, as tenure in academia does not significantly correlate with AI adoption, as shown later.

[9]The rankings include the top 1200 universities globally. I code authors affiliated with universities outside the top 1200 as having a rank lower than 1200.

phaFold2. Column (5) reveals that structural biologists who use cryo-EM are also more likely to cite AlphaFold2, consistent with cryo-EM's frequent application in analyzing large and complex proteins. Column (6) shows that more senior researchers are 2% less likely to cite AlphaFold2, though this difference is not statistically significant. Column (7) indicates that researchers from higher-ranked institutions are 7% more likely to adopt AI.

Column (8), which regresses on all characteristics, identifies highly cited researchers as the strongest predictor of AI usage. Expertise in novel proteins and the use of advanced equipment, such as cryo-EM, are the second and third strongest predictors of AI adoption. Appendix Table A2 further supports the finding that highly cited researchers are more likely to solve complex proteins and employ advanced methods such as cryo-EM. Taken together, these results suggest that highly cited researchers are more likely to adopt AI, potentially due to their expertise in solving complex proteins, where AlphaFold2 provides significant assistance.

4.4.2 Impacts by Structural Biologist Skill

Figure 7 presents the event study estimates for highly cited structural biologists and their lowly cited counterparts. Panel A shows that both groups maintain a flat pattern in the number of PubMed publications before and after the debut of AlphaFold2. Panel B reveals that the emergence of AlphaFold2 increases PubMed citations for highly cited structural biologists, with a stable pattern prior to the treatment. In contrast, lowly cited researchers show a relatively flat citation pattern in 2022 and 2023, accompanied by a dip in 2021.

Panel C shows that highly cited researchers produce more distinct research, as measured by TF-IDF, following the availability of AlphaFold2. In contrast, lowly cited researchers exhibit no significant change in research novelty. Panel D indicates that highly cited structural biologists maintain a stable number of studies on protein structure, while lowly cited researchers reduce their focus on protein structure research. Finally, Panel E reveals that lowly cited researchers are more likely to increase research in areas outside protein structure determination, whereas highly cited researchers show no significant shift toward research beyond protein structure determination.

Table 6 compares the pooled DiD estimates between highly cited and lowly cited structural biologists. All estimates exclude observations from 2021, as previously discussed. Columns (1) and (2) indicate no significant change in the number of publications for either group following the emergence of AlphaFold2. However, Columns (3) and (4) reveal increasing citation polarization among structural biologists: highly cited researchers experience a 10% increase in citations driven by the availability of the AI tool, while lowly cited researchers see minimal impact on their citation counts. Columns (5) and (6) show that

highly cited researchers produce more novel research with the availability of AI, whereas lowly cited researchers exhibit no change in research novelty. Columns (7) and (9) indicate that highly cited researchers maintain a consistent composition of research between protein structure studies and other fields after the introduction of AI. In contrast, Columns (8) and (10) reveal that lowly cited researchers reduce 13% of their protein structure research but increase 6% in other areas of expertise.

Overall, the divergence in academic performance between highly cited and lowly cited researchers following the availability of AI aligns with earlier findings that highly cited researchers adopt AI technology more quickly. These results suggest that highly cited researchers are more likely to leverage AI tools like AlphaFold2 to address complex proteins, thereby increasing their research novelty and impact. In contrast, lowly cited researchers tend to shift their focus away from protein structure research, reallocating their efforts to other areas. These findings underscore the importance of prior research abilities in driving the heterogeneous impacts of AI.

4.5 Research Team Size

My main analyses thus far focus on principal investigators (i.e., the last authors) in structural biology research, as they can represent the overall performance of the lab. Although we do not observe any reduction in publications by principal investigators, AlphaFold2's ability to accelerate and simplify certain processes in protein structure determination may replace specific roles within research teams, such as research staff, PhD students, and postdocs. To examine the impact of AlphaFold2 on the team size of structural biologists, I analyze the number of authors listed on their publications.

Figure 8 illustrates the impact of AlphaFold2 on the total number of coauthors across all PubMed papers. To account for the influence of publication volume, I include the number of PubMed papers as a control variable. Panel A shows no significant change in the total number of authors after the emergence of AlphaFold2. Panel B presents the event study for AI users versus non-AI users. Among AI users, there is an increase in the number of coauthors, suggesting that AI adoption does not correlate with a reduction in team size. In contrast, non-AI users experience a decline in the total number of authors. Finally, Panel C reveals that highly cited structural biologists maintain a stable pattern in the number of coauthors. While lowly cited researchers experience a dip in 2021, their coauthor numbers recover swiftly thereafter.

Table 7 presents the pooled estimates for the number of coauthors, excluding observations from 2021. Overall, there is no significant reduction in coauthors for the average structural

biologist, nor is there notable heterogeneity between highly cited and lowly cited structural biologists. Taken together, while AI accelerates the workflow of structure determination, we do not yet observe a significant change in team size within structural biology research.

5 Conclusion

AI technology is rapidly becoming an integral part of society, transforming the way we work and live. Among its pioneering applications was AlphaFold2, a breakthrough in protein structure research. This paper provides an initial exploration of AlphaFold2's impact on the scientific productivity of structural biologists. By analyzing publication records of individual structural biologists and their control counterparts, I find that structural biologists' academic influence, as measured by citations, has increased. However, the evidence suggests that these positive effects are primarily confined to structural biologists who incorporate AI into their research. Conversely, those who have not yet adopted AI experience a decline in academic performance.

The emergence of AI appears to widen the citation gap between highly cited and less-cited structural biologists. I argue that highly cited researchers adopt this new AI tool more quickly because their expertise aligns with research areas that benefit more from AI. The emergence of AI also reveals heterogeneous impacts on research direction choices between highly cited and less-cited researchers. While lowly cited researchers are reducing their focus on protein structure research, highly cited researchers maintain a similar level of engagement in this area. Nonetheless, the broader implications of AI for the research workforce remain unsettled, particularly as the technology continues to evolve.

My paper focuses exclusively on the impacts of AlphaFold2 on structural biologists, as they are the most directly affected by this technology. However, AlphaFold2's influence extends far beyond protein structure determination. Future research could examine its impact on other areas of biology that have historically relied on human protein characterization but now benefit from computational predictions.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3.

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American economic review*, 108(6):1488–1542.

Acemoglu, D. and Restrepo, P. (2020a). Robots and jobs: Evidence from us labor markets. *Journal of political economy*, 128(6):2188–2244.

Acemoglu, D. and Restrepo, P. (2020b). The wrong kind of ai? artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society*, 13(1):25–35.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Agrawal, A. K., McHale, J., and Oettl, A. (2023). Artificial intelligence and scientific discovery: A model of prioritized search. Technical report, National Bureau of Economic Research.

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30.

Azzaz, F., Yahi, N., Chahinian, H., and Fantini, J. (2022). The epigenetic dimension of protein structure is an intrinsic weakness of the alphafold program. *Biomolecules*, 12(10):1527.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.

Bick, A., Blandin, A., and Deming, D. J. (2024). The rapid adoption of generative ai. Technical report, National Bureau of Economic Research.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative ai at work. Technical report, National Bureau of Economic Research.

Caplin, A., Deming, D. J., Li, S., Martin, D. J., Marx, P., Weidmann, B., and Ye, K. J. (2024). The abc's of who benefits from working with ai: Ability, beliefs, and calibration. Technical report, National Bureau of Economic Research.

Charness, G., Jabarian, B., and List, J. A. (2023). Generation next: Experimentation with ai. Technical report, National Bureau of Economic Research.

Cockburn, I. M., Henderson, R., and Stern, S. (2018). The impact of artificial intelligence on innovation: An exploratory analysis. In *The economics of artificial intelligence: An agenda*, pages 115–146. University of Chicago Press.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2013). Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics*, 128(2):531–580.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Editorial (2022). Method of the year 2021: protein structure prediction. *Nat. Methods*, 19:1.

Feigenbaum, J. and Gross, D. P. (2020). Automation and the fate of young workers: Evidence from telephone operation in the early 20th century.

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

Goulet, A. and Cambillau, C. (2022). Present impact of alphafold2 revolution on structural biology, and an illustration with the structure prediction of the bacteriophage j-1 host adhesion device. *Frontiers in Molecular Biosciences*, 9:907452.

Hill, R. and Stein, C. (2019). Scooped! estimating rewards for priority in science.

Hill, R. and Stein, C. (2021). Race to the bottom: Competition and quality in science.

Humlum, A. and Vestergaard, E. (2024). The adoption of chatgpt. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2024-50).

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Jia, R., Roberts, M. E., Wang, Y., and Yang, E. (2022). The impact of us-china tensions on us science. Technical report, National Bureau of Economic Research.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kanazawa, K., Kawaguchi, D., Shigeoka, H., and Watanabe, Y. (2022). Ai, skill, and productivity: The case of taxi drivers. Technical report, National Bureau of Economic Research.

Kim, S. (2023). Shortcuts to innovation: The use of analogies in knowledge production.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Kovalevskiy, O., Mateos-Garcia, J., and Tunyasuvunakool, K. (2024). Alphafold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121(34):e2315002121.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

McCoy, A. J., Sammito, M. D., and Read, R. J. (2022). Implications of alphafold2 for crystallographic phasing by molecular replacement. *Acta Crystallographica Section D: Structural Biology*, 78(1):1–13.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*.

Otis, N., Clarke, R. P., Delecourt, S., Holtz, D., and Koning, R. (2023). The uneven impact of generative ai on entrepreneurial performance. *Available at SSRN 4671369*.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.

Restrepo, P. (2023). Automation: Theory, evidence, and outlook.

Subramaniam, S. (2023). Structural biology in the age of ai. *Nature Methods*, pages 1–2.

Subramaniam, S. and Kleywegt, G. J. (2022). A paradigm shift in structural biology. *Nature Methods*, 19(1):20–23.

Terwilliger, T. C., Liebschner, D., Croll, T. I., Williams, C. J., McCoy, A. J., Poon, B. K., Afonine, P. V., Oeffner, R. D., Richardson, J. S., Read, R. J., et al. (2023). Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, pages 1–7.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., 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):D439–D444.

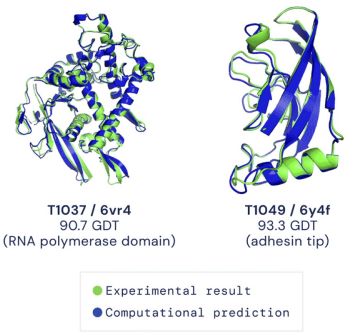
Wang, W., Gao, G., and Agarwal, R. (2023). Friend or foe? teaming between artificial intelligence and workers with variation in experience. *Management Science*.

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.

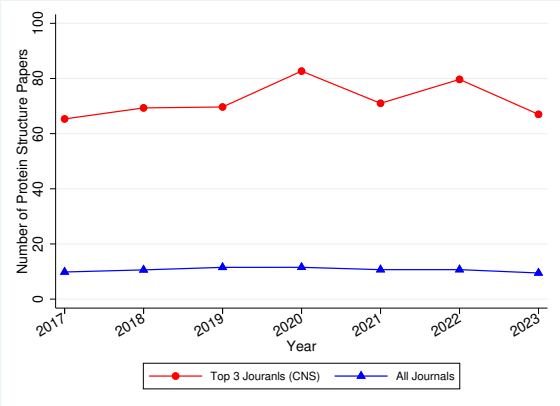
Figures & Tables

Figure 1: AlphaFold2 and Protein Structure Papers

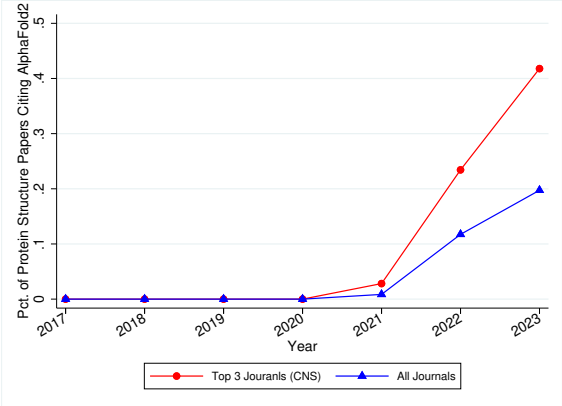
A: AlphaFold2's Prediction of CASP14 Target Protein



B: Protein Structure Papers

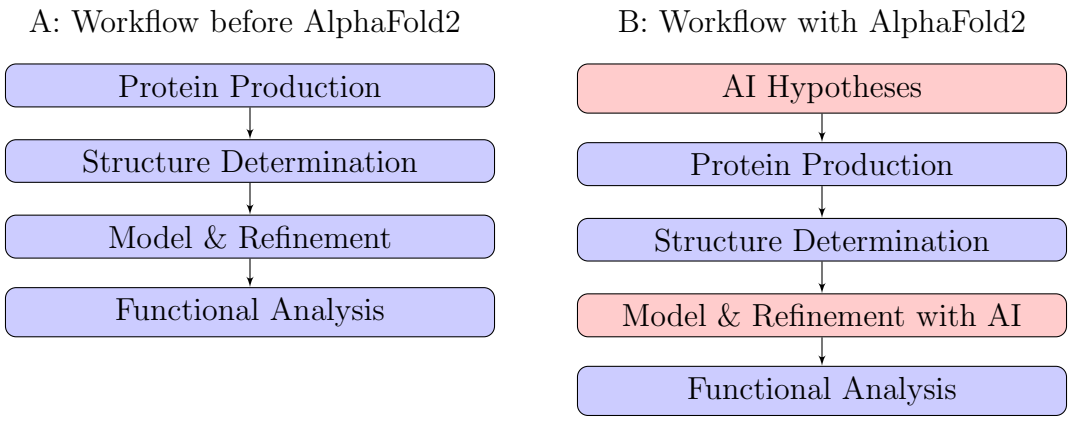


C: Proteins Structure Papers Using AI



Note: Panel A shows the comparison between the AlphaFold2's prediction and the structure characterized by experiment in CASP14. The left one is CASP14 target T1037 (PDBID: 6VR4). The right one is CASP14 target T1049 (PDBID: 6Y4F). (source: https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/). Panel B presents the annual average number of protein structure papers from 2017 to 2023 in *Cell*, *Nature*, *Science*, and all journals that published at least two protein structure papers between 2013 and 2016. Panel C shows the percentage of protein structure papers citing AlphaFold2 from 2017 to 2023 in *Cell*, *Nature*, *Science*, and the same set of journals. The list of protein structure papers are downloaded from Protein Data Bank.

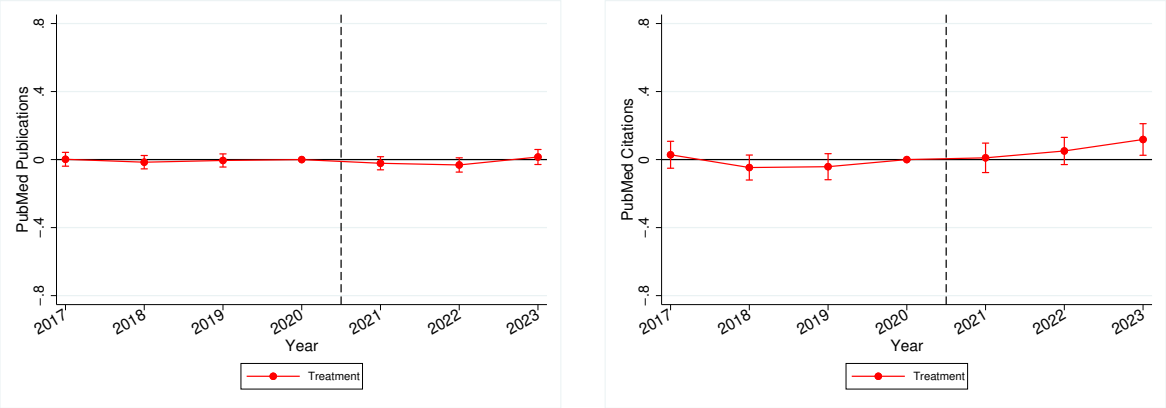
Figure 2: Protein Structure Determination Workflow



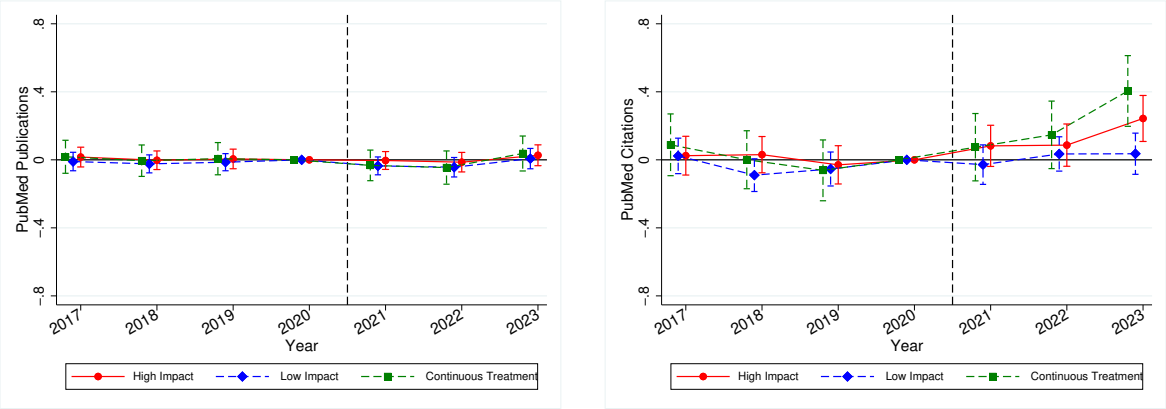
Note: Panel A illustrates the workflow of protein structure determination prior to the advent of AI. Panel B depicts the workflow of protein structure determination with AI assistance.

Figure 3: Impacts on PubMed Publications and Citations

A: Default Treatment

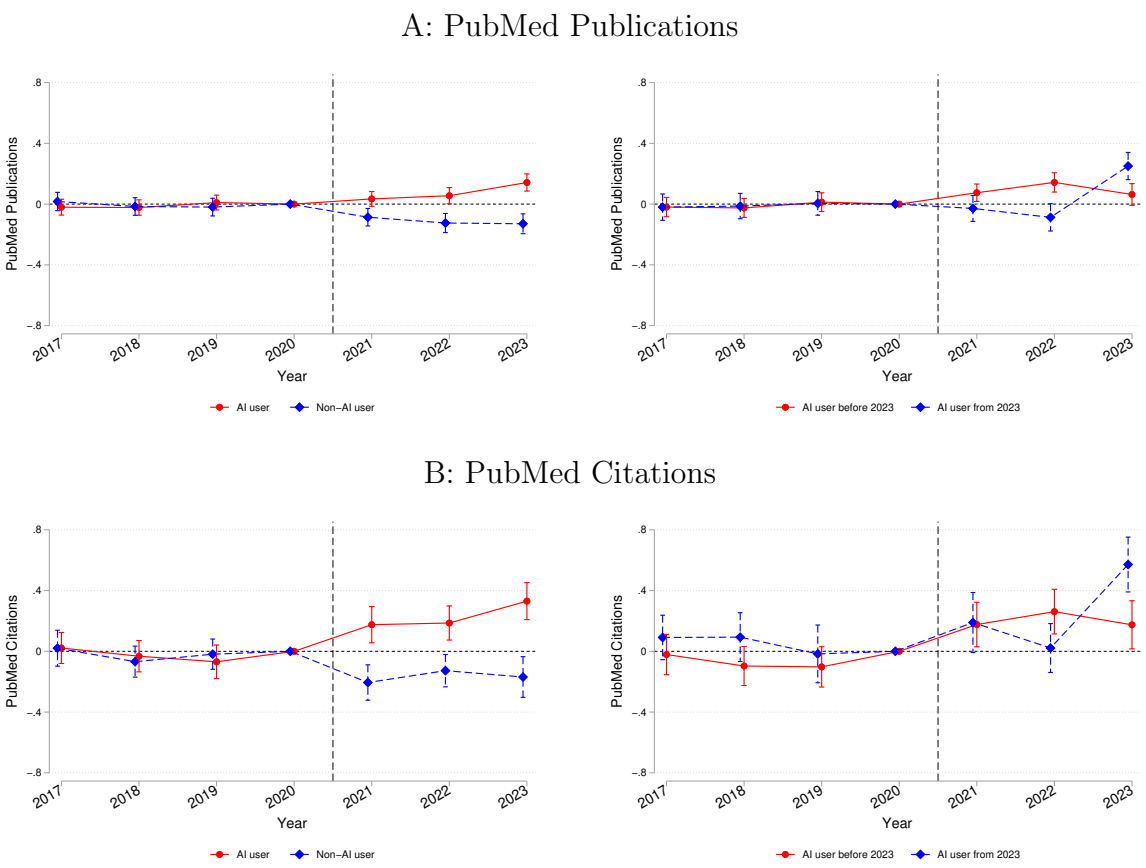


B: Alternative Definitions of Treatment



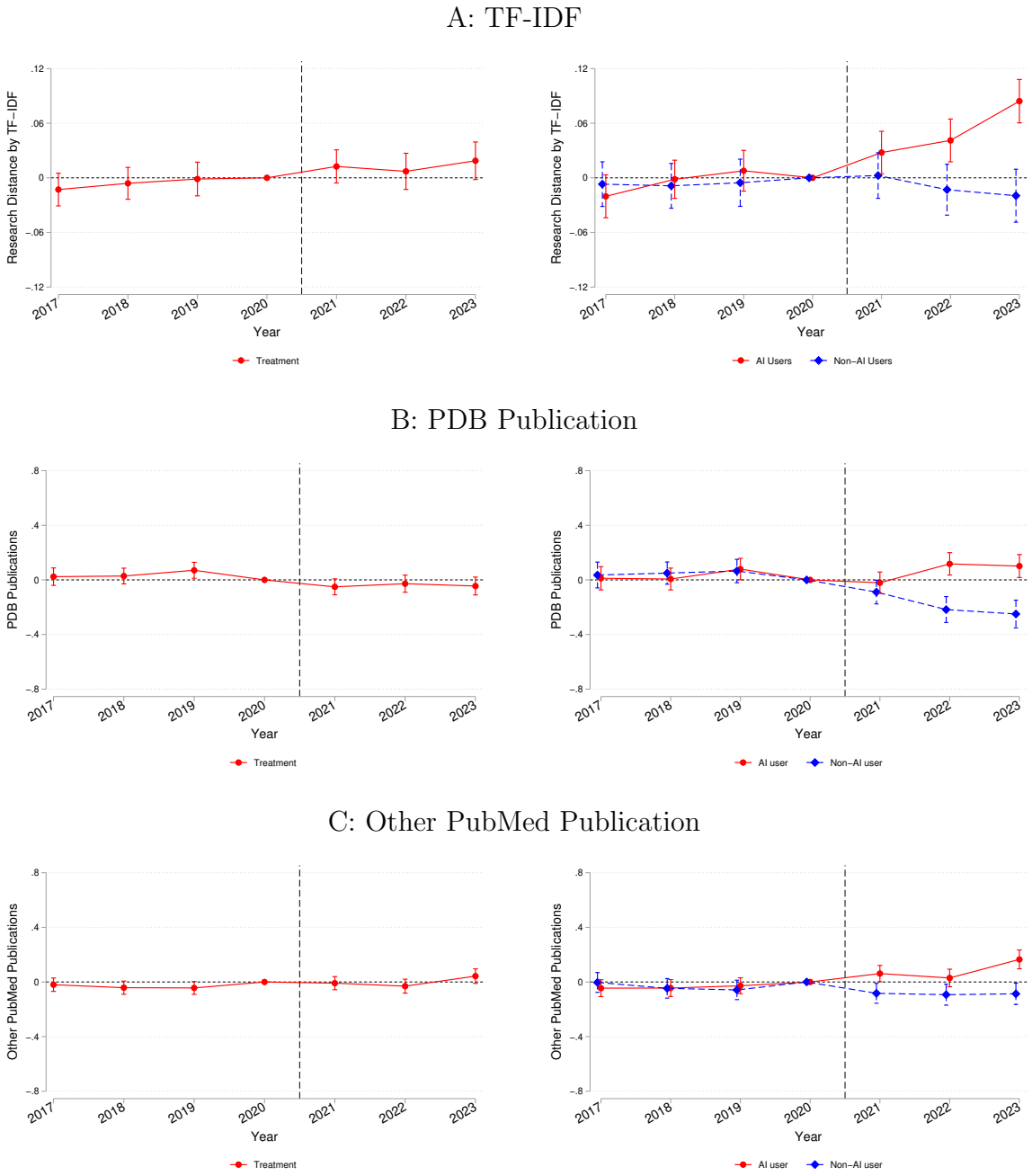
Note: These figures plot the event-study estimates of AI's impact on structural biologists' academic performance. The left side of the figure presents results for PubMed publications, while the right side shows results for PubMed citations. Panel A reports estimates using a dummy treatment variable representing all structural biologists. Panel B provides results based on various treatment definitions: high impact, low impact, and continuous treatment. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Figure 4: Impacts on PubMed Publications and Citations: AI vs. Non-AI Users



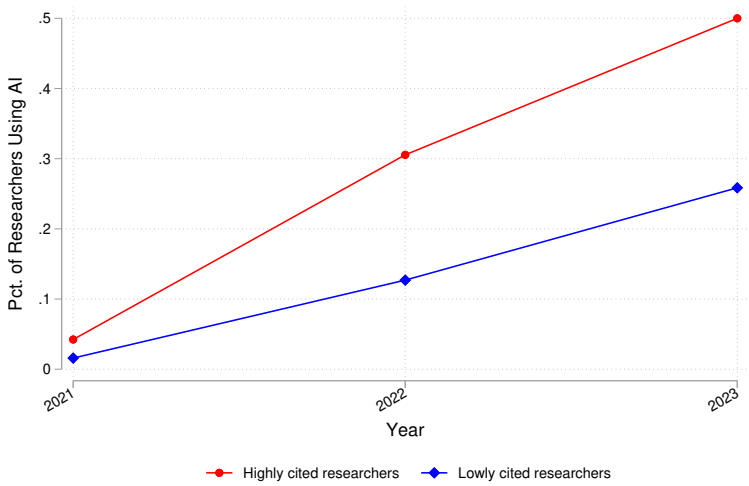
Note: These figures plot the event-study estimates of academic performance for AI users and non-AI users. The left side compares AI users with never-takers, while the right side compares early adopters, who began using AI before 2023, with late adopters, who started using AI in 2023. Panel A presents the results for PubMed publications, and Panel B shows the results for PubMed citations. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Figure 5: Impacts on Publication Direction by TF-IDF and SciBERT



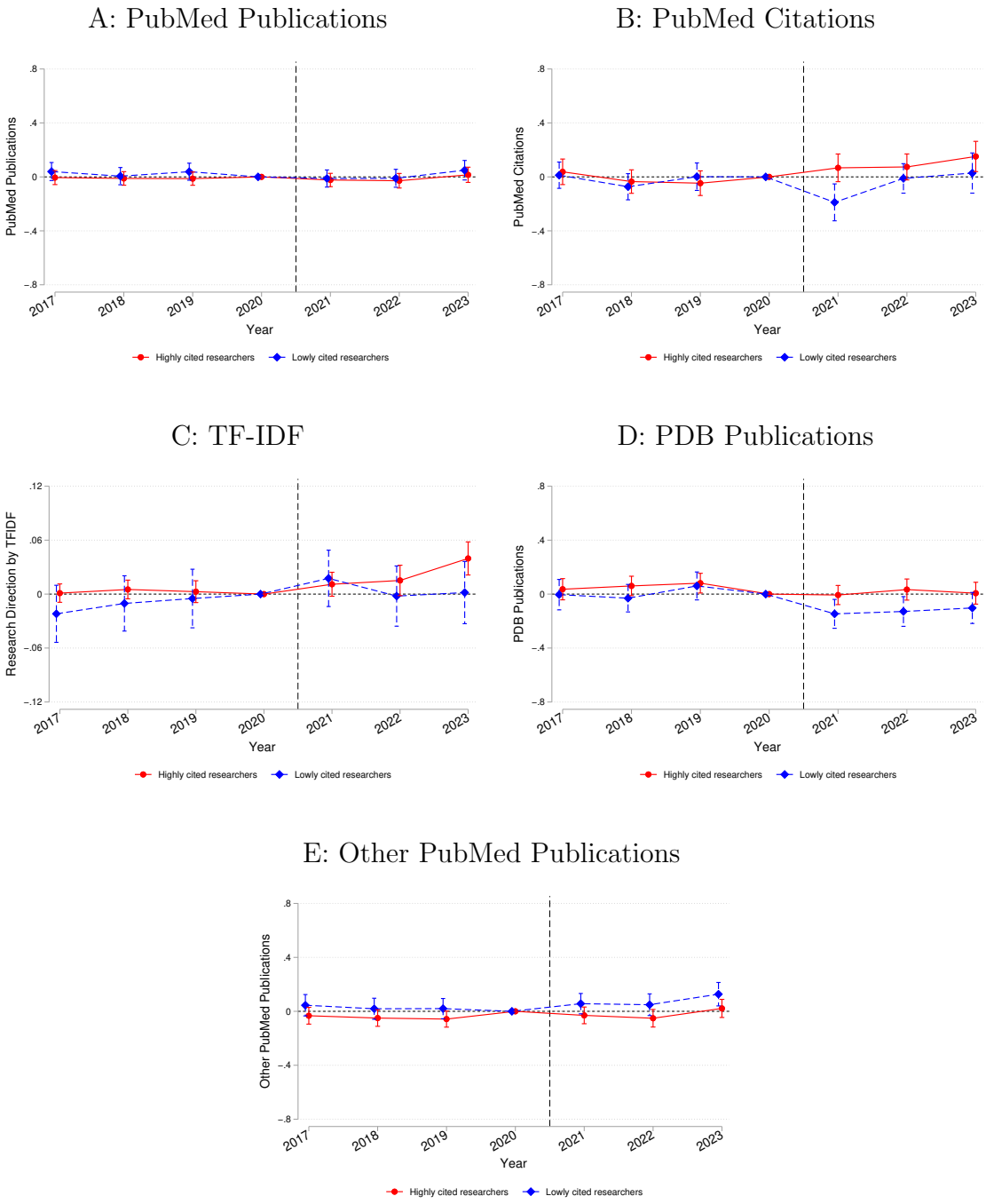
Note: These figures plot the event-study estimates of research direction. The left side illustrates the impacts of AI on structural biologists as a whole, while the right side compares AI users with non-AI users. Panel A presents the results for research novelty, measured by TF-IDF. Panel B shows the estimates for PDB publications, and Panel C reports the results for other PubMed publications. The models in Panel A employ OLS regressions, while the models in Panels B and C use Poisson regressions. All models include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Figure 6: AlphaFold2 Adoption Over Time



Note: This figure depicts the cumulative percentage of structural biologists adopting AI from 2021 to 2023. The red line represents highly cited researchers, while the blue line represents lowly cited researchers.

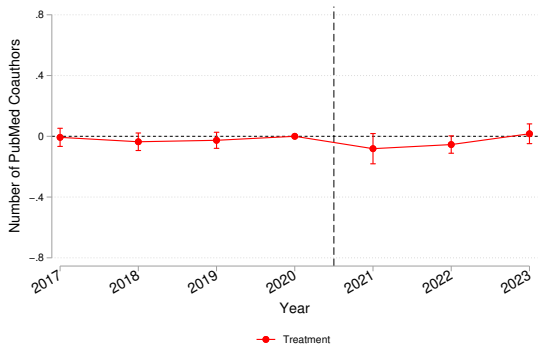
Figure 7: Heterogeneous Impacts: Highly Cited vs Lowly Cited Researchers



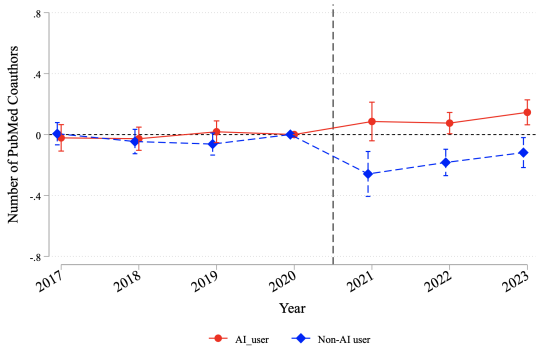
Note: These figures compare the impacts of AI on academic performance and research direction between highly cited and lowly cited researchers. Panels A–E present the outcomes for PubMed publications, PubMed citations, TF-IDF measures of novelty, PDB publications, and other PubMed publications, respectively. The models in Panel C employ OLS regressions, while the models in other panels use Poisson regressions. All models include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Figure 8: Impacts on Authorship Size

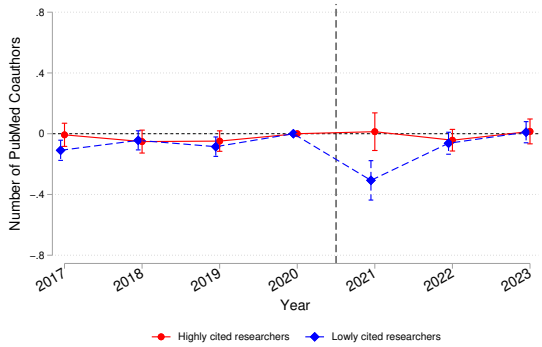
A: Full Sample



B: AI vs. Non-AI Users



C: Highly Cited vs. Lowly Cited



Note: These figures illustrate the impact of AI on the number of coauthors among structural biologists. Panel A shows the impact on the full sample, Panel B compares AI users with non-AI users, and Panel C contrasts highly cited structural biologists with lowly cited ones. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Table 1: Summary Statistics

	Structural Biologists		Control Biologists		T − C
Baseline (2017-2020)	Mean	St. Dev.	Mean	St. Dev.	Difference
PubMed Publications	4.8881	5.3106	4.4039	5.1588	0.4842***
PubMed Citations	205.3112	379.5406	213.5378	402.4663	−8.2266
PubMed TF-IDF	0.8547	0.1676	0.8091	0.2014	0.0457***
Number of Coauthors	46.2756	75.1454	45.4257	132.6364	0.8499
Academic Age	27.8314	9.3861	28.0113	8.6513	−0.1799
University	0.7779	0.4158	0.8824	0.3222	−0.1045***
Covid Citation	6.7301	57.7111	1.6205	28.0056	5.1095***
AlphaFold2 User (post-2020)	0.3787	0.4852	0.0289	0.1675	0.3498***
Number of obs	1756		10865		-9109

Note: This table summarizes the baseline (2017–2020) characteristics, including PubMed publications, PubMed citations, TF-IDF novelty measures, number of coauthors, academic age, and university affiliation. The last row reports the percentage of researchers who cited an AlphaFold2 paper after 2020. All values represent the averages of the raw data. Columns (1) and (2) provide the annual averages and standard deviations for structural biologists, while Columns (3) and (4) present the corresponding values for control biologists. Column (5) reports the differences between structural biologists and the control group. Statistical significance: $^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

Table 2: Impacts on PubMed Publications and Citations

	PubMed Publication				PubMed Citation			
	(1) Total	(2) High Impact	(3) Low Impact	(4) Continuous	(5) Total	(6) High Impact	(7) Low Impact	(8) Continuous
Treatment × Post	−0.0065	−0.0020	−0.0086	−0.0158	0.0802**	0.1129**	0.0651*	0.1908**
	(0.0154)	(0.0232)	(0.0201)	(0.0370)	(0.0326)	(0.0550)	(0.0389)	(0.0850)
Log likelihood	−137872.29	−74362.71	−80617.17	−137872.30	−2355548.81	−1237515.38	−1493638.46	−2355579.37
Number of obs	75559	41904	42848	75559	75531	41885	42835	75531
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CEM Matching	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table presents the DiD regression estimates on the impact of AI on structural biologists' academic performance. Column (1) to (4) report the estimates for PubMed publications using various treatment definitions: a dummy variable for all structural biologists, a dummy for high-impact researchers, a dummy for low-impact researchers, and a continuous treatment based on the percentage of PDB publications at baseline. Columns (5) to (8) present the estimates for PubMed citations using the same treatment definitions as in Columns (1) to (4). All estimates exclude observations from 2021. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Statistical significance: $^{***}p < 0.01,^{**}p < 0.05,^{*}p < 0.1$.

Table 3: Correlations between Protein Characteristics and AI Usage

	Protein Characteristics			Protein Quality		
	(1) Novelty	(2) Length	(3) Weight	(4) Resolution (Å)	(5) Rfree	(6) Citation
AlphaFold2	0.2203***	0.1705***	0.1728***	0.0123	0.0016	0.2899**
	(0.0230)	(0.0313)	(0.0313)	(0.0138)	(0.0108)	(0.1275)
R^2	0.3095	0.4657	0.4652	0.5832	0.3245	
Log likelihood						-42435.99
Characteristics	No	No	No	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Number of obs	5390	5390	5390	5390	4072	5388

Note: This table presents the correlation between protein characteristics/quality and the usage of AlphaFold2. Column (1) reports the correlation results for a dummy variable indicating whether the protein is novel. Columns (2) and (3) show results for the log-transformed length and weight of the proteins, respectively. Columns (4) and (5) present quality metrics using the log-transformed resolution and R-free values, while Column (6) examines citation counts. Columns (1) to (5) use OLS regressions, and Column (6) employs a Poisson regression. Columns (4) to (6) also control for the characteristics from Columns (1) to (3). All regressions include both individual fixed effects and year fixed effects. Statistical significance: $^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

Table 4: Impacts on Research Direction

	(1) TF-IDF	(2) PDB Pub	(3) Other Pub
Treatment × Post	0.0180^{***}	-0.0670^{***}	0.0282
	(0.0069)	(0.0234)	(0.0178)
R^2	0.4892		
Log likelihood		−129564.22	−135050.00
Number of obs	75726	74554	75320
Individual FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
CEM Matching	Yes	Yes	Yes

Note: This table presents the DiD regression estimates of the impact of AI on structural biologists' research direction. Columns (1) to (3) report the estimates for research novelty measured by TF-IDF, PDB publications, and other PubMed publications, respectively. All estimates exclude observations from 2021. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Statistical significance: $^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

Table 5: Characteristics of AlphaFold2 Users

	AlphaFold2 User							
Baseline Characteristics	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High Citation	0.2559***							0.2093***
	(0.0224)							(0.0230)
High Novelty		0.1871***						0.1361***
		(0.0228)						(0.0228)
High Length			0.1269***					−0.0466
			(0.0230)					(0.1153)
High Weight				0.1315***				0.1027
				(0.0230)				(0.1154)
Cryo EM					0.2397***			0.1069***
					(0.0280)			(0.0302)
High Age						−0.0223		−0.0168
						(0.0233)		(0.0221)
University							−0.0238	−0.0150
							(0.0300)	(0.0285)
High Rank							0.0742***	0.0115
							(0.0255)	(0.0246)
Constant	0.2608***	0.2925***	0.3103***	0.3078***	0.3290***	0.3909***	0.3680***	0.1836***
	(0.0152)	(0.0155)	(0.0169)	(0.0169)	(0.0127)	(0.0172)	(0.0246)	(0.0311)
R^2	0.0691	0.0370	0.0170	0.0183	0.0401	0.0005	0.0049	0.1104
Number of obs	1756	1756	1756	1756	1756	1756	1756	1756

This table presents the OLS correlation between the usage of AlphaFold2 and the baseline characteristics of structural biologists. The outcome variable equals 1 if the structural biologist has cited AlphaFold2. Statistical significance: $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

Table 6: Heterogeneous Impacts: Highly Cited vs. Lowly Cited Researchers

	Publication		Citation		TF-IDF		PDB Pub		Other Pub	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Highly Cited	Lowly Cited	Highly Cited	Lowly Cited	Highly Cited	Lowly Cited	Highly Cited	Lowly Cited	Highly Cited	Lowly Cited
Treatment × Post	−0.0020	−0.0052	0.1035***	0.0139	0.0252***	0.0091	−0.0242	−0.1240***	0.0160	0.0621**
	(0.0197)	(0.0252)	(0.0399)	(0.0425)	(0.0065)	(0.0115)	(0.0297)	(0.0385)	(0.0222)	(0.0313)
Log likelihood	−67552.87	−66413.46	−1660007.26	−624027.39			−63634.90	−62434.83	−66423.18	−64844.22
R^2					0.44	0.45				
Number of obs	30069	43702	30069	43657	30096	43890	29909	42880	30065	43472
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CEM Matching	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table presents the heterogeneous impact of AI by skill level on structural biologists' academic performance, measured by publications and citations, and research direction, measured by TF-IDF, PDB publications, and other PubMed publications. All estimates exclude observations from 2021. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Statistical significance: $^{***}p < 0.01,^{**}p < 0.05,^{*}p < 0.1$.

Table 7: Impacts on Authorship Size

	(1) Full Sample	(2) Highly Cited	(3) Lowly Cited
Treatment × Post	−0.0272	−0.0099	−0.0489
	(0.0254)	(0.0331)	(0.0345)
Log likelihood	−694230.28	−389029.46	−280934.74
Number of obs	75559	30069	43702
Individual FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
CEM Matching	Yes	Yes	Yes

Note: This table presents the impact of AI on structural biologists' team size. Column (1) reports the impacts on all structural biologists, Column (2) focuses on highly cited researchers, and Column (3) examines lowly cited researchers. All estimates exclude observations from 2021. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Statistical significance: $^{***}p < 0.01$,** $p < 0.05$,* $p < 0.1$.

Appendix A Additional Figures and Tables

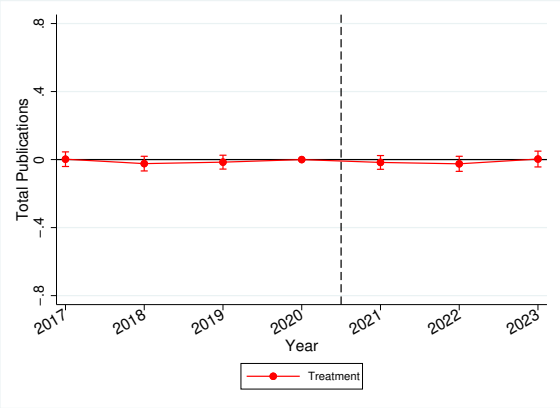
Table A1: Distribution of Treatment Intensity

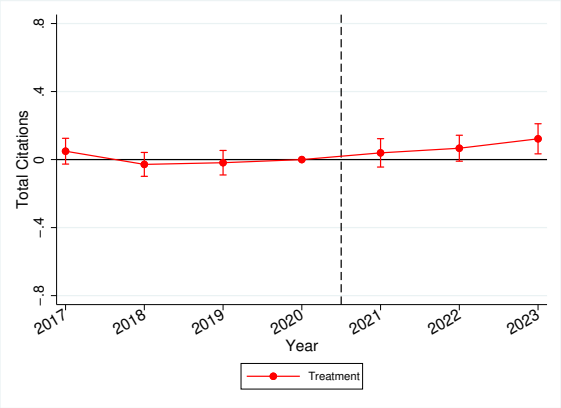
Percentile	Pct. of PDB Papers
1	0.0484
25	0.2143
50	0.3636
75	0.5417
99	1
Number of obs	1756

Note: This table presents the distribution of the percentage of PDB publications for structural biologists during 2013–2016.

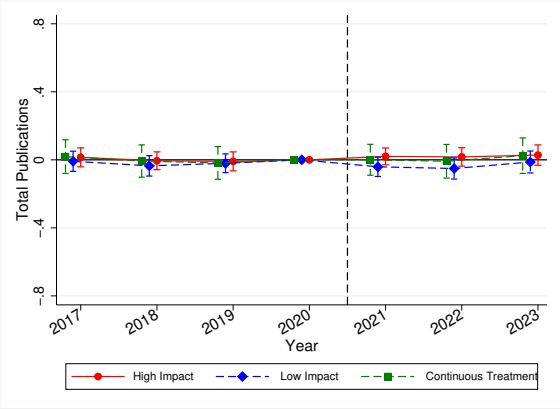
Figure A1: Impacts on Total Publications and Citations

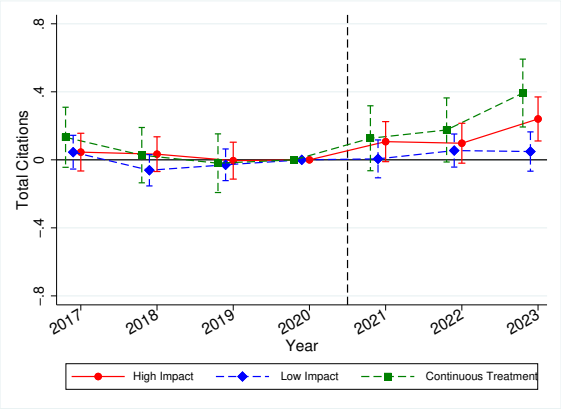
A: Total Publications





B: Total Citations





Note: These figures plot the event-study estimates of AI's impact on structural biologists' academic performance. The left side of the figure presents results for total publications, while the right side shows results for total citations. Panel A reports estimates using a dummy treatment variable representing all structural biologists. Panel B provides results based on various treatment definitions: high impact, low impact, and continuous treatment. All models are Poisson regressions that include individual fixed effects, year fixed effects, and CEM. Robust standard errors are clustered at the author level. Each vertical bar represents the 95% confidence interval of the estimate.

Table A2: Characteristics of Highly Cited Researchers

	(1) High Novelty	(2) High Length	(3) High Weight	(4) Cryo EM
High Citation	0.1111***	0.1429***	0.1432***	0.2139***
	(0.0235)	(0.0236)	(0.0236)	(0.0187)
R^2	0.0125	0.0204	0.0205	0.0692
Number of obs	1756	1756	1756	1756

Note: This table presents the correlation between a structural biologist's baseline citation level and their expertise in novel, long, and heavy proteins, as well as their use of Cryo-EM in protein research. Statistical significance: $^{***}p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$.

Appendix B Coarsened Exact Matching

To ensure that control biologists are comparable to structural biologists in academic background and exhibit similar publication trends prior to the treatment year (i.e., the debut of AlphaFold2), I employ the Coarsened Exact Matching (CEM) procedure. First, I perform an exact match based on authors' affiliations as of 2020, using the affiliation listed in their most recent paper by that year. For authors with multiple affiliations in a single paper, I select the first listed affiliation. Affiliations supporting experiments in protein structure determination can differ significantly from those without structural biologists, and this step helps account for potential confounding effects related to institutional differences.

Next, I select several baseline covariates: publication age, COVID citations, and baseline citations, which are matched coarsely. Accounting for potential confounding effects caused by differences in career stages is a standard procedure in the economics of science literature. Publication age is defined as the difference between the year of an author's first publication and 2020. Baseline COVID citations account for potential confounding effects from researchers' focus on COVID-related topics. These are measured by the number of citations received for COVID-related studies, as identified by MeSH terms, during 2017–2020. Baseline citations ensure that treated and control groups are at comparable citation levels, calculated as the average citations received during 2017–2020.

For the matching process, I divide publication age into nine bins: 1–5, 6–10, 11–15, 16–20, 21–25, 26–30, 31–40, 41–50, and > 50 years. For COVID citations, I classify researchers into five categories: 0, 1–100, 100–250, 250–500, and > 500 citations. For baseline citations, I group researchers into two categories: 0–150 and > 150 citations. In this way I successfully matched 72% of the structural biologists. The CEM procedure relies heavily on human judgment, especially when choosing relevant covariates and deciding the bin size. To ensure robustness, I tested various cutoff points for the covariate bins and found that minor adjustments did not significantly alter the main results.