Innovation Growth Lab

AI in Science: Emerging evidence from AlphaFold 2

Full Literature Review, Methodology and Results

David Ampudia Vicente and George Richardson

25th November 2025

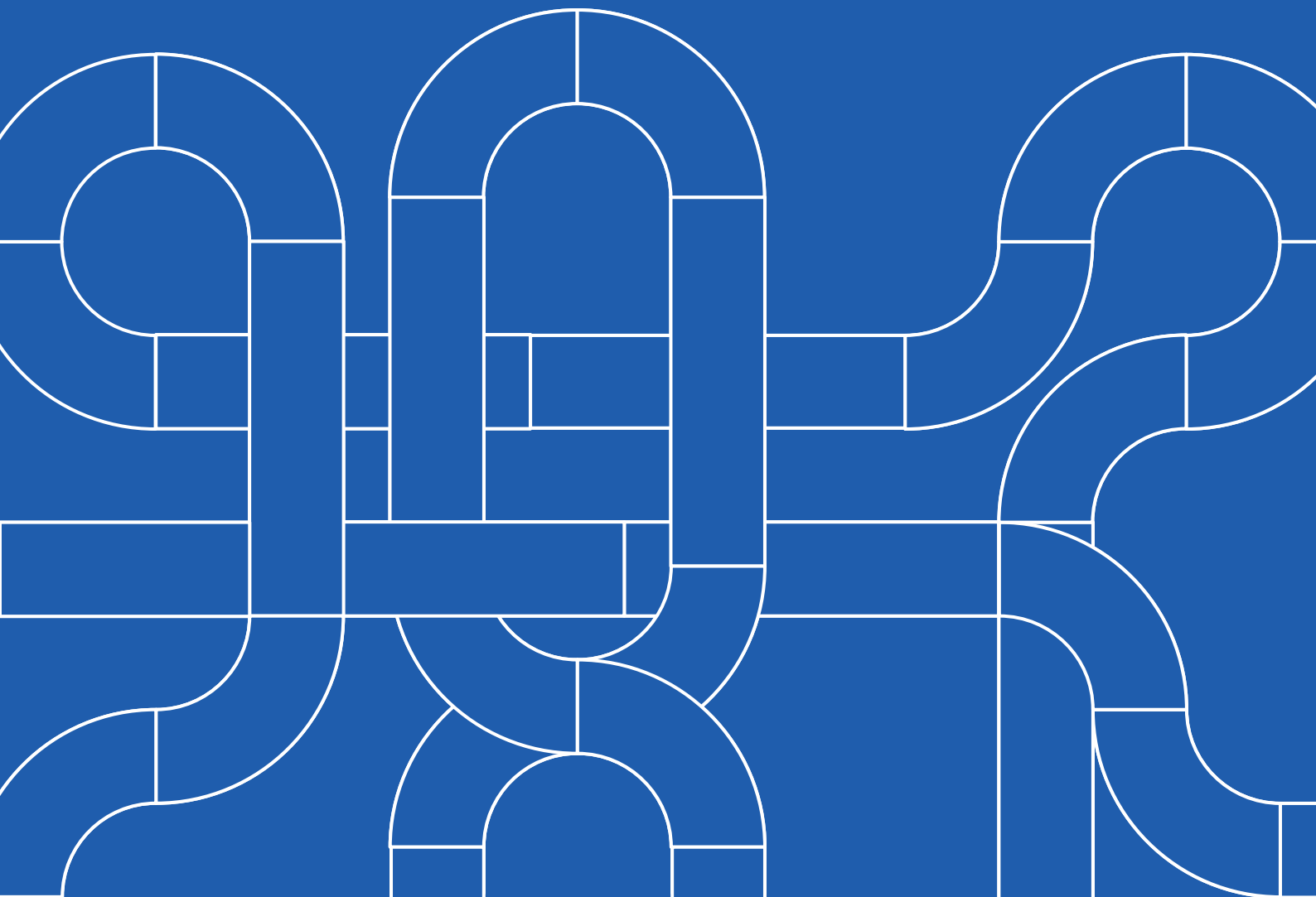

Table of contents

Executive summary

Context: AI in Science

Artificial intelligence is increasingly integrated into scientific research, with the promise of speeding up scientific discovery and widening the scope of exploration. AI tools may make it easier for scientists to explore uncharted areas of knowledge or find solutions in complex and highly combinatorial problem spaces, and accelerate research. Or, they might create "streetlight effects", where scientists focus on areas and disciplines where there is good, existing data, producing only incremental gains and a narrowing of research diversity. While AI tools may speed up particular processes in an area of scientific research, they may not ultimately address the most critical bottlenecks to unlock productivity gains overall, and productivity gains may lead to a deluge of low quality studies.

AI related research has been growing across all fields, however the implications of the development and adoption of advanced AI tools on critical aspects of scientific discovery remain ambiguous. In addition, it is unclear how its effects on influence will vary according to factors such as the experience of researchers and resources available to them. Due to the as yet emergent nature of the technology, the current evidence base concerning its impacts is limited.

AlphaFold 2 represents a notable case. By predicting protein structures with high accuracy, AlphaFold 2 addresses a bottleneck within structural biology that has restricted the rate and diversity of discoveries across the life sciences. It is thought that AlphaFold 2 may improve the allocation of experimental resources, reduce the time required for labour-intensive tasks, and encourage the faster development of new hypotheses, all leading to further applications for human health and beyond.

This report investigates AlphaFold 2's impact on experimental discovery, research productivity, and translational science, comparing its impacts to other AI and non-AI driven frontier developments in structural biology that were published in the same time period.

Data Sources and Methods

To examine the impact of AlphaFold 2 on science, a dataset of academic publications, patents, protein structure submissions, and clinical trials from 2018 to early 2025 was constructed.

To quantify the impact attributable to AlphaFold 2 we compare the performance of papers, researchers and labs who cite AlphaFold 2 with a baseline of general structural biology research activity. We then make an additional comparison of AlphaFold 2 against

three groups of highly cited papers, chosen to represent a 'status quo' counterfactual for frontier structural biology.

The dataset was created by:

- Identifying core AlphaFold 2 publications and all downstream papers in their citation networks.
- Defining counterfactual sets of frontier developments in structural biology, distinguishing between AI-intensive and non-AI driven developments in protein structure prediction, selected based on citation levels and relative impact, referred to as frontier developments.
- Identifying patents, clinical trials, and published protein structures that are linked to all papers in the dataset.
- Enriching the data, including distinguishing between strong and weak citations. This allows for the identification of downstream research that meaningfully builds on AlphaFold 2 methodologically, versus that which references it in passing.

The study then employed a rigorous methodological approach, including:

- Tracking adoption of AlphaFold 2 and counterfactual developments by individual, established researchers and research laboratories led by a principal investigator.
- Examining historical parallel trends and applying Coarsened Exact Matching (CEM) to ensure baseline similarities in publication history, discipline, institutional location, and COVID-19 research activity.
- Using a difference-in-differences framework to estimate the influence of AlphaFold 2, AI-intensive developments, non-AI protein prediction developments, and other frontier structural biology developments. We account for a set of controls and include fixed effects for the quarter and unit of observation across groups.

Findings

Scientific reach

- In total, almost 681,000 publications cite AlphaFold directly or indirectly. Among other frontier methods, we create citation networks numbering 612,000 papers for the AI-intensive frontier; 585,000 for the non-AI protein prediction frontier; and 525,000 from other structural biology frontier papers.
- Looking at the strength of citations in the network, we estimate that some 269,900 research papers are strongly connected - they incorporate AlphaFold 2 or elements of it into their methodology.

- Around 1,957,000 unique scientists have published research connected to AlphaFold directly or indirectly. We estimate around 778,000 of these are scientists who have meaningfully incorporated AlphaFold 2 or aspects of it at least once.
- Papers building on AlphaFold 2 directly tend to focus on biochemistry, while indirect citations tend to be in more applied fields like medicine. This underscores the importance of using indirect citations to track downstream "real world" impacts.

Experimental structural biology:

- Researchers using AlphaFold 2 have exhibited consistent growth in the publication of protein structures, as measured by new submissions to the Protein Data Bank, as well as higher rates of mapping uncharacterised proteins, surpassing both AI-intensive and non-AI counterfactual methods.
- Researchers building on AlphaFold 2 demonstrate 45% to 49% higher rates of protein structure submissions across various measures compared to baseline structural biology researchers. Rates for those researchers are higher than rates for those with work linked to other frontier developments across almost every measure.
- Across a range of metrics for structure similarity, AlphaFold 2 influence is linked with more structurally novel proteins according to several metrics, and is more likely to be involved in the discovery of structures that are practically unique. This increased novelty, however, tends to be associated with lower structure resolution, potentially because these protein structures are harder to determine. In contrast, the association between other frontier developments and structure novelty are smaller, less statistically significant, or non-existent.

Academic productivity

- AlphaFold 2 links show differential effects on publication volume. There is a 2.5% increase for individual researchers and a 5.1% increase for laboratories. This difference may be due to these teams being able to operate at a larger scale, realising productivity gains more easily.
- Experience or capacity for scale appear to amplify publication benefits: Methodological use of AlphaFold 2 is linked to an 11.5% increase in laboratory publication volume. This association is not observed for any other frontier developments.
- Building on AlphaFold 2 corresponds with increased citation counts for researchers and laboratories (8.1% and 10.4% higher respectively) than the structural biology baseline. While AlphaFold 2 is associated with significantly increased citation counts across all units of analysis, and a high number of researchers, AI frontier and protein prediction frontier research publications are also associated with increases in most cases.

Applied research and innovation

Disease-relevant research

- We identify 3,097 papers directly citing AlphaFold 2 that contain disease-related terms (MeSH C), and an additional 76,320 indirectly citing papers.
- Work linked to AlphaFold 2 shows a dual correlation with disease-relevant research: while individual papers show no significant increase in disease focus, researchers building on the tool have a 9.3% higher probability of producing disease-relevant work, an association substantially larger than for any other frontier technique.

Clinical articles

- We identify 287 clinical article citations to direct AlphaFold 2 papers (267 unique articles), with an additional 12,787 clinical citations to papers indirectly citing AlphaFold 2 (5,758 unique). Of these, we identify at least 4,300 as strongly linked.
- At the paper level, AlphaFold 2 is associated with a doubling in the probability of receiving a clinical citation, an effect size roughly double that of other AI and non-AI protein prediction methods. However, no significant association is found for individual researchers or laboratories, where non-AI protein prediction is the only frontier group with a positive link to clinical citations.

Patents

- We identify 100 patents citing papers that directly reference AlphaFold 2, with an additional 3,800 patent citations to papers indirectly citing AlphaFold 2.
- Links to AlphaFold 2 are associated with a substantial increase in patent citations across all units: 36.8% for papers, 34.2% for laboratories, and 22.6% for researchers, suggesting a broad association with work of commercial relevance.
- A positive association with patent quality (measured by subsequent patent citations) is found for AlphaFold 2-related papers, a finding that is distinctive compared to other frontier methods. No significant effect is detected at the researcher or laboratory level.

Implications

AlphaFold 2's accurate predictions and open resources appear to accelerate research on less-explored proteins, enabling more diverse and impactful experimental outputs. Labs and researchers that adopt AlphaFold 2 achieve modest yet significant improvements in publication volume and citation rates. The adoption of insights derived from work building on AlphaFold 2 for commercial and clinical application remains nascent, with a small number of patents referencing AlphaFold 2, indicating potential for future expansion.

Overall, AlphaFold 2 emerges as a notable AI-driven innovation that helps scientists tackle complex protein structures, expands the scope of experimental research, and supports influential publications.

Limitations and Issues for Further Research

- Time frame: The short interval since AlphaFold 2's public release may understate longer-term effects.
- Selection bias: Even with CEM, researchers building on AlphaFold 2 could be more innovative or better resourced.
- Field variance: Observed benefits vary by discipline; further subfield studies could help identify clearer patterns.
- Citation chains: Citation strength is not available for all publications, nor a nuanced measure of methodological use. Improving this data would allow for more fine grained analysis of AlphaFold 2's uses in downstream research..
- Commercial uptake: Patent activity is an imperfect proxy for AlphaFold 2's use in industry and industrial use is rarely published in other ways; evaluating AlphaFold 2's longer-term role in industry will require additional research.
- Data quality: OpenAlex is a valuable source of available publication metadata, but there are known data quality issues associated with publications and authorships.

1. Introduction

"If research and development could be made even 5% more efficient, the returns could be immense." (Wang 2024)

Artificial Intelligence (AI) has experienced fast growth in recent years driven by the arrival of neural networks able to analyse large amounts of complex data. AI systems are being deployed in many sectors including scientific Research and Development, an area where its impact could be particularly consequential.

AI adoption in science has been growing rapidly with AI-engaged research (measured as papers mentioning AI keywords in their abstracts) rising from a share of just over 5% of publications across all fields in 2020 to over 8% in 2022 (Duede 2022). Around half of all AI related research now happens outside of mathematics, statistics and computer science, the disciplines most associated with the development of the field (Gargiulo et al. 2022). Bianchini et al. attribute the expansion of deep learning related work specifically to two 'innovation booms', finding that the life sciences and physical sciences have exhibited the highest levels of growth (Bianchini et al. 2022).

The impacts of AI adoption on scientific outputs and processes are the subject of much discussion (Wang et al. 2023; Agrawal et al. 2024; Bail 2024; Messeri and Crockett 2024; Sumner 2024). Many researchers believe that it could accelerate progress and open new research frontiers, while others are more sceptical and point to the negative ways in which the nature of science and the work of scientists may be impacted (Griffin et al. 2024; Sumner 2024; Messeri and Crockett 2024).

Some of these hypotheses and conjectures are starting to be studied empirically. This study seeks to expand the evidence base with an in-depth analysis of the impact of AlphaFold 2, a protein structure prediction system that was released in 2021 and demonstrated significant advancement over the then state-of-the-art. It has received substantial acclaim including the awarding of the 2024 Nobel Prize in Chemistry for its development to John Jumper and Demis Hassabis.

The rest of this section summarises existing literature about the impact of AI in science and introduces the AlphaFold 2 breakthrough. After outlining our data sources and methods in Section 2, Sections 3-5 present our findings about AlphaFold 2's impacts on experimental structural biology, academic productivity, translational outcomes and their complementarities with skills and other capabilities. Section 6 concludes with implications, limitations and issues for further research.

1.1 What do we know about the impact of AI on science?

Productivity

One core promise of AI has been the ability to transform data, make predictions and reproduce human decision making in a high-throughput, automated fashion, leading to higher efficiency and productivity in the processes where it is applied. In science, this applies to a range of opportunity areas including the design and execution of lab tasks, modelling complex systems, generating and enriching large datasets, identifying solutions in large problem spaces, and creating new modes for the development and communication of scientific knowledge (Griffin et al. 2024). Examples range from mundane micro-tasks, such as coding qualitative data for quantitative analysis in psychological research, through to larger, multi-part tasks, such as automating and optimising chemical flow reactors (Idalski Carcone et al. 2019; Schweidtmann et al. 2018). During the earlier phases of AI adoption within science, these kinds of applications often involved researchers developing their own models using open source software packages (Pedregosa et al. 2011; Abadi et al. 2016; Paszke et al. 2019).

More recently, academic and private research labs have started releasing pre-trained foundation models that can be fine-tuned on new applications. Examples include MetNet GraphCast, for forecasting precipitation, GNoME for predicting material properties, and HLS, for classifying satellite images of the Earth's surface (Sønderby et al. 2020; Andrychowicz et al. 2023; Merchant et al. 2023; Vaswani et al. 2023; Jakubik et al. 2023). In addition, some labs have created model predictions that can be incorporated by researchers in downstream analysis, including the AlphaFold 2 database of protein structures and the PolymerGenome database containing the predictions of functional properties of polymers (Varadi et al. 2022; Doan Tran et al. 2020).

Generative AI models able to analyse and generate text including OpenAI's ChatGPT, Anthropic's Claude and Google's Gemini are also being deployed in scientific R&D - for example, to simulate human behaviour and replicate social science experiments without involving any human participants (Hewitt et al. 2024). Models tailored towards programming languages, such as the one powering GitHub Co-pilot and Codeium, assist in writing code more efficiently, which many scientists use to analyse data and automate routines.

An increasing body of literature analyses the impact of systems like these . Gao and Wang examined the potential benefits of AI to science as a whole, finding that AI related papers (those containing terms describing AI methods) are more likely to be classed as 'hit' papers (in the top 5% cited papers within a field) (Gao and Wang 2024). Another study finds that papers citing other work that are related to deep learning techniques receive 10.32% more citations on average, albeit with a 19.57% increase in the variation of the

citation distribution (Bianchini et al. 2022). One possible explanation for this is AI helping researchers to produce higher quality work, although it may also be the case that the citation impact is due to AI being a trending topic within science.

Although not the focus of this paper, we note that AI can enable other scientific use cases such as searching for information and synthesising knowledge. Applications including Elicit, SciSpace and PaperQA aim to facilitate literature reviews by allowing researchers to search for publications by asking questions, 'chat' with papers, and extract information such as methodological details or findings as structured data. Their creators hope that the often time consuming and challenging tasks of synthesising knowledge and developing hypotheses will be made more productive. The authors of PaperQA report that their model exceeds the performance of humans on scientific literature tasks, including identifying whether contradictions exist between publications (Skarlinski et al. 2024).

Scientific creativity

Going beyond improvements in efficiency, AI could also impact search for new ideas and scientific creativity. Historically, scientists have used their expertise and knowledge of past experiments to plan new experiments. For example, to develop new materials with desired properties, scientists would leverage their experience to generate candidates, synthesise and test them in the lab, and use the results to refine their understanding and improve their ideas. Computational methods, including machine learning, can expedite this process, for example by predicting chemical properties or optimising synthesis conditions (Schweidtmann et al. 2018; Savage et al. 2024). Modern foundation models, coupled with increasingly available computing power, raise the prospect of 'inverse' materials design, in which scientists specify the desired functional properties of a material and let the AI system generate ideas for feasible compounds, predict their characteristics, and prioritise candidates for real-world experimentation. In essence, AI generates ideas and screens them before any real-world experiments take place, allowing scientists to target experimental efforts on the most promising candidates. Going beyond the synthesis of existing knowledge, Deep Science Ventures are using LLMs to power a 'co-pilot' for science, that uses agentic workflows, coupled with large databases of relevant information to simulate R&D processes including decomposing a problem, hypothesising and ranking solutions, searching for relevant collaborators, analysing data, and market analysis for downstream products (Hammond 2023).

All this suggests that AI could act as an emerging general method of innovation due its potential to transform fields not only in terms of scientific outputs, but also through the means in which scientific discovery and innovation happen (Bianchini et al. 2022). This idea draws on Agrawal et al. who suggest that AI can be used in both the 'search' and 'discovery' components of science (Agrawal et al. 2018), identifying the most useful combinations of knowledge in complex search spaces.

One concern surrounding the use of AI in science is the potential for scientific monocultures. Overreliance on AI tools could lead to a narrowing of research focus and the dominance of certain methodologies and perspectives, potentially hindering scientific diversity and creativity (Messeri and Crockett 2024). These concerns are centred on a set of underlying risk factors. One is that science could increasingly focus on areas most suited to AI based discovery. This could lead to a reduction of research in other, potentially valuable domains, or distort research efforts toward quantitative, reductive and predictive avenues of enquiry. Another is that AI will yield solutions to complex problems faced by scientists, but will not help to develop theories and knowledge of underlying mechanisms (Krenn et al. 2022). For example, while AI might aid in the identification of chemical compositions for materials with a set of desired functional properties, the predictions themselves will not further scientific understanding of the relationship between material structure and function. Overreliance on AI might therefore restrict the ability to build generalised models of natural phenomena.

While AI tools are evolving quickly, and efforts may be taken to mitigate some of the potential downsides, it remains that AI has ambiguous implications for scientific creativity. On the one hand, AI models and predictions could make it easier for scientists to explore unknown regions of the knowledge space, thus helping to accelerate scientific discovery. On the other, they could create a "streetlight effect" leading scientists to focus on those areas and disciplines of the knowledge space where there is higher quality data, thus leading to incrementalism and a loss of diversity.

The current evidence base is limited and mixed. Within narrow, well-defined search and discovery problems, AI methods appear to increase the likelihood of traversing less explored parts of the problem space and identifying novel solutions. A comparison of theoretical and data driven genome wide association studies (GWAS) - a method to identify links between genomic attributes and diseases - found that the latter exploited a larger proportion of the genetic landscape and has been more likely to identify neglected genes[1]. Further efforts in the use of AI methods for drug discovery and repurposing, solid state chemistry, flow reactor design have been discussed, with some studies demonstrating the ability of machine to generate solutions to problems that are unintuitive to researchers (Chenthamarakshan et al. 2023; Huang et al. 2024; Zunger 2018; Savage et al. 2024, 2).

Collectively, these studies suggest that within AI suited experiments, where part of the challenge involves exploring a large problem space, its use will not limit innovativeness. Nonetheless, wider issues of potential homogenisation in whole research fields remain. The growth of AI related research appears to be happening despite a lack of integration between AI and non-AI work (Duede et al. 2024a; Fontaine et al. 2024). This dynamic is

[1] While data driven GWAS is not necessarily an AI technique, it exhibits characteristics such as theory-free discovery, high information throughput and scanning large combinatorial spaces. Machine learning and AI methods have been used to further enhance the method (Enoma et al. 2022, 20)

observed alongside AI related papers becoming semantically more similar to computer science papers. The adoption of deep learning methods in the health sciences is also associated with lower recombinatorial novelty of knowledge between domains (Bianchini et al. 2022). It is as yet unclear whether these tendencies should be interpreted as the early signs of a potential cross-fertilisation of concepts between disciplines that will result in diversification, or an imposition of AI-driven mindset that will lead to homogenisation.

Complementarities with other capabilities and domain knowledge

AI is associated with the automation of tasks across a wide range of economic activity, disruptions to the relationship between capital and labour, and changes in the nature of work itself (Eloundou et al. 2023; Brynjolfsson et al. 2023; Brynjolfsson and Mitchell 2017). Science is no exception. AI is likely to affect the demand for certain scientific skills and may lead to some substitution of labour for capital (Bianchini et al. 2022). While some tasks may be automated or deskilled, leading to a decrease in demand for certain skills, other roles requiring a combination of domain knowledge and AI expertise are likely to emerge.

As with scientific creativity, the implications are ambiguous: AI tools could lower the barrier to entry for researchers by automating complex tasks and providing access to new data and advanced methods using only a computer (Gao and Wang 2024). This might allow newcomers (to a field) or low resource labs to more readily participate in frontier research. However, there are also concerns that existing inequalities in the research landscape might be entrenched or exacerbated, with established researchers and higher resource labs more able to access required compute and capitalise on the economies of scale permitted by AI methods.

At present, the development and successful use of AI systems appear to make use of the domain knowledge of researchers. For example, pharmaceutical firms with greater domain knowledge seem to be more discerning and effective when making investments based on the discoveries of new gene-disease associations (Tranchero 2024). Given AI models make predictions based on limited and imperfect information, a scientist's ability to identify 'true positive' solutions returned from high dimensional combinatorial search spaces can be a determining factor in the development of high value downstream innovations. Researchers are used to new breakthroughs disrupting established domains and harnessing new tools in their work. In one survey, 73% of scientists reported that they believed AI tools would improve their work, suggesting enthusiasm for the development of new methods and their diffusion (Van Noorden and Perkel 2023). To capitalise on these developments, it is likely that many scientists will need to develop new skills in AI driven research (Gao and Wang 2024). The generalisability of AI methods and the decreasing relative cost between machine predictions and labour pose the risk of an unusually large disruption.

Going beyond skills, many labs around the world have difficulty in accessing the hardware and stable internet connections to be able to take advantage of these efforts. Those with

easy access to large scale compute will still have an advantage, particularly when it comes to the development of novel large models and the early exploitation of them. For example, in their analysis of AI adoption in science, Bianchini et al. identify enabling factors for participating in AI powered science such as collaboration between AI and non-AI researchers, pointing to the fact that labs with connections to AI labs will find it easier to take advantage of new methods (Bianchini et al. 2024). Researchers who lack these connections may find themselves at a disadvantage. The acceleration of AI development and diffusion has led to calls to democratise data, models and infrastructure, as well as knowledge required for AI development in order to create an AI divide between scientists (Bail 2024; Dessimoz and Thomas 2024).

1.2 AlphaFold 2: A Case Study for AI in Science

With AI related research growing and diffusing, and the continuous development of new tools aimed at accelerating and transforming science, there is an increasing interest in the nature and size of AI impacts . Empirical evidence to understand the effects of AI on the areas covered in the previous section is critical to developing policy responses to harness or steer the direction of AI in science. As noted in the previous section, there are important gaps in the evidence base, partly due to the relatively short timespan since many breakthrough AI tools were released. This study aims to contribute to that evidence base.

We focus on the impact of AlphaFold 2, an AI model developed by Google DeepMind to address the challenge of protein structure prediction, a key challenge in the field of structural biology. AlphaFold 2 is seen as an important milestone as recognised by the 2024 Nobel Prize for Chemistry (Callaway 2020; AlQuraishi 2020; Perrakis and Sixma 2021; Bertoline et al. 2023; Elofsson 2023; Brzezinski et al. 2024). It also offers an interesting opportunity to study the impact of AI in science, presenting somewhat of an exogenous shock to the field of protein structure prediction when it was first demonstrated and again when made openly available for use by scientists.[2]

We seek to contribute to the evidence base with an analysis of publications, patents, clinical trials, and information from protein structure databases. The rest of this section introduces AlphaFold 2 and the questions it raises, as well as existing analyses of its impact.

A brief history of AlphaFold 2

Proteins are chains of amino acids that fold into complex shapes and are present in all life forms. The order of the sequence of amino acids dictates the protein structure, which can be very complex, and determines how it interacts with other proteins and biological

[2] While AlphaFold 2 can be seen as an exogenous shock to the field of structural biology as a whole, we recognise that its adoption and impact will not be strictly randomised. We attempt to control for this using coarsened exact matching (CEM) and our use of dif-in-dif. This is described in Section 2.

molecules. These functional properties give each unique protein a specific role in the body of an organism. Visualising the structures of proteins is an important aspect of understanding biological phenomena that are responsible for life, disease and the effectiveness of drugs and other therapies. It has therefore become an area of significant importance to the biological sciences.

Scientists first began determining protein structures experimentally in the mid-20th Century, using X-ray crystallography. Max Perutz and John Kendrew won the Nobel Prize in Chemistry in 1962 for identifying the structure of haemoglobin, the first time a protein had been mapped. While methods have advanced, determining a protein structure experimentally remains a challenging task, often taking structural biologists several years to complete (Hill and Stein 2020). In an effort to determine protein structures more quickly and efficiently, scientists began to explore computational methods to predict them. It is in computational protein folding that AlphaFold 2 has made a significant contribution (Saplakoglu 2024).

In 1994, the Critical Assessment of Structure Prediction was set up as an exercise for the computational protein folding community to benchmark their results on structures that had recently been determined experimentally, but not yet submitted to the public facing Protein Database (PDB). This community experiment has given researchers a method to objectively compare results and methodological approaches.

One key evaluation metric used at CASP is the Global Distance Test - Test Score (GDT-TS). Between 2006 and 2016, the top scoring algorithms submitted to CASP had plateaued with scores between 30 and 40 (out of 100) (Callaway 2020). While methodological progress had been made since the field first started, the problem remained far from solved.

In 2018, Google DeepMind, a private research lab, entered AlphaFold 2 in the CASP13 competition, achieving a GDT-TS of 68.5 demonstrating superior performance to other competitors and highlighting the promise of deep learning methods.

In 2020, its successor AlphaFold 2 reached levels of accuracy comparable to experimental methods, leading some scientists to suggest that the protein structure prediction problem has been solved. The model attained a GDT-TS of 92.4 at CASP14. In addition to publishing a methods paper about AlphaFold 2 (Jumper et al, 2021), DeepMind also released AlphaFold 2's source code and in partnership with EMBL-EBI (the European Bioinformatics Institute) developed a database that currently contains over 200 million freely-downloadable predicted protein structures. Another innovation of AlphaFold 2 is the ability of the model to predict a confidence score for its predictions, assisting downstream researchers to make their own determination about the accuracy of a computed structure.

AlphaFold 2's Influence

As a novel and accurate AI model targeting a bottleneck in scientific discovery, AlphaFold 2 has the potential to enhance elements of scientific productivity and change the search

and discovery process within structural biology and other fields that draw on it. The creation of the AlphaFold 2 protein database opens up possibilities for biologists without AI skills (Perrakis and Sixma 2021) who can use AlphaFold 2 predictions to accelerate the interpretation of experimental data when solving new protein structures, or develop hypotheses about biological mechanisms that can be tested downstream (Kovalevskiy et al, 2024). The ability to adapt and tweak the methodology provided by AlphaFold 2 has led to a range of models being developed for other applications requiring greater predictive accuracy in the biological sciences and beyond (Saplakoglu 2024). One example is ColabFold which combines optimised code to enable protein folding predictions to be carried out with low hardware requirements (Mirdita et al. 2022). Another is the work of fellow 2024 Nobel Prize in Chemistry winner, David Baker, on novel protein design (Lisanza et al. 2024). It has also been suggested that advancements triggered by the development of AlphaFold 2 will be important for understanding disease mechanisms, drug discovery, and vaccine development (Duran-Frigola et al. 2013; Higgins 2021; Hazra and Patra 2021).

Much of the literature studying the impacts of AlphaFold 2 to date provide qualitative descriptions of the potential impacts, or case studies of individual discoveries and innovations that have enabled the technology. Given the relatively short duration that has passed since its publication and open release, more systemic and quantitative studies have been limited (Yu 2024). This study, which compares publication activity in scientific fields "exposed" to AlphaFold 2 to other fields finds no increase in the volume of papers attributable to AlphaFold 2, but a 8% increase in citations for those authors. The study also finds evidence that AlphaFold 2 is being used to study longer, more complex and more novel proteins. Within structural biology, use of AlphaFold 2 is associated with an attenuation of the citation gap between highly cited and less cited researchers, in line with other studies that show AI assistance tools can equalise the ability of workers to carry out tasks, despite initial differences in skill or experience (Noy and Zhang 2023; Brynjolfsson et al. 2023).

2. Data sources and methodology

Research questions

- Does AlphaFold 2 lead to more impactful research outputs?
- Do these effects differ from those seen with other frontier developments?

Methodological design

Our study examines the impact of AlphaFold 2 on scientific research and innovation. To provide a comprehensive and robust answer to our research questions, we identified a number of methodological design criteria for the analysis that informed data collection and analysis.

1. **Study several units of analysis:** We develop metrics and methods to study impact at the level of individual publications and researchers, distinguishing between established researchers and laboratories led by principal investigators whose research reflects the activities of larger teams / organisations. We also identify early career researchers (ECRs).

2. **Broad coverage of R&D outputs**: We examine AlphaFold 2's effects on a range of research outputs including experimental protein structures in foundational biology, academic publications, clinical trial write-ups, and patent citations.

3. **Track direct and indirect impacts:** We consider both direct and indirect impacts of AlphaFold 2. Direct impacts include research that builds on AlphaFold 2 methodologically or employs it as a research input. Indirect impacts involve works further downstream from AlphaFold 2 that subsequently build on those directly influenced R&D contributions.

4. **Comparative analysis:** We compare AlphaFold 2 impacts with a structural biology baseline matched on observable characteristics to reduce confounding, and with three distinct sets of counterfactual frontier developments in structural biology with similar usage characteristics.

5. **Longitudinal analysis:** In our analysis of researcher and laboratory impacts, we examine changes in trends after the signals of AlphaFold 2 adoption and counterfactual methods, accounting for time-invariant characteristics through fixed effects approaches to help isolate the associations of interest.

In the following sections, we outline the data sources used, our approach to building a sample of relevant works, and the methods applied to our empirical analysis. Several methodological details are mentioned only briefly in the main text, but are expanded in the Appendix.

2.1 Data sources

Consistent with our first criterion ("Broad coverage of R&D outputs") we collect data from a range of sources covering different stages of the R&D process.

Academic publications

We use OpenAlex, an open database of academic papers and authors that provides detailed information such as which papers cite each other, the main topics and concepts of research, details about authors, and measures of impact like yearly citation counts (Priem et al. 2022). OpenAlex is a comprehensive scholarly knowledge graph that succeeded Microsoft Academic Graph, offering structured metadata for over 200 million scholarly documents and their connections. We use OpenAlex to create datasets on papers, authors, and labs, collecting nearly 5 million unique publications.

Recent assessments of OpenAlex's representativeness for bibliometric research underscore its suitability for citation-based analysis; Culbert et al. (2025), for instance, found its reference coverage comparable to Web of Science and Scopus for recent, shared literature. While OpenAlex also demonstrates broad inclusivity, notably in its extensive coverage of open access journals (Maddi et al. 2025), it is noted that, like any large-scale evolving database, specific characteristics such as variations in certain metadata fields (Culbert et al., 2025) or persisting geographical underrepresentation in some areas (Maddi et al., 2025) are documented. The data for our analysis focuses on recent publications, which are less likely affected by any limitations in OpenAlex's coverage and metadata quality.[3]

We complement OpenAlex with Semantic Scholar, another publication database developed by the Allen Institute for AI that provides information about citation intent and identifies particularly influential papers. Semantic Scholar employs natural language processing and machine learning techniques to analyze the semantic content of scientific literature, enabling the classification of citation relationships (Cohan et al. 2019). Intent labels include 'background', 'result', and 'methodology', making it possible to distinguish between references to publications that meaningfully contribute to the creation of new works, rather than simply motivate them. Their labelling algorithm relies on open-access publications, limiting the scope to publications with openly available full-text. Around 37% of citation links in our data contain intent measures.

Structural biology data

As discussed in the previous Section, AlphaFold 2 is being used by structural biologists to interpret experimental data. To examine potential associations between AlphaFold 2

[3] We note that OpenAlex metadata exhibit a pattern of date imputation that disproportionately assigns publication dates to January of each year. To address this, all time-based visualisations in our analysis report data at a quarterly level, using smoothing techniques.

adoption and adoption, and experimental structural biology research, we analyse comprehensive data on protein structure determination activities.

The Protein Data Bank (PDB) serves as the primary global repository for experimentally determined protein structures, containing over 200,000 structures determined through methods such as X-ray crystallography, NMR spectroscopy and cryo-electron microscopy. We complement this with data from UniProtKB, a comprehensive resource for protein sequence and functional information that provides detailed annotations on protein characteristics, evolutionary relationships and biological roles (The UniProt Consortium 2023).

We collect metadata on over 227,000 protein submissions, identifying nearly 79,000 publications linked to these structures in OpenAlex using Digital Object Identifiers (DOIs). For each submission linked to a paper in our sample, we gather UniProt identifiers and merge with UniProtKB to obtain additional metadata, including functional annotations, references to associated diseases, and sequence-level details. These data points allow us to derive variables to assess the novelty and potential impact of new structures from multiple perspectives:

- **Structure rarity**: Our primary analysis of novelty is based on a direct structural comparison of each new PDB submission against all historically prior structures. This is performed using TM-scores and identity values calculated using FoldSeek. This approach provides two key metrics:
    - **Fold-level Novelty:** We use the maximum TM-score (*alntmscore*) to the single closest prior structure for each PDB submission (van Kempen et al. 2024)[4]. A low maximum TM-score indicates a potentially novel fold, as the submission is structurally distant from anything known at the time.
    - **Sequence-level Novelty:** We use the fraction of identity (*fident*) from the same FoldSeek alignment. A low score, particularly for a structure with a high TM-score, signals the discovery of a distant evolutionary homolog (Dennler and Ryan, 2025).
- **Global Shape Novelty:** As a complementary, alignment-free perspective, we assess each structure using the RCSB PDB's similarity method (Guzenko et al., 2020). This generates a 'Structure Match Score' based on global volumetric shape, where a lower score suggests novelty in the protein's overall topology.
- **Organism rarity**: This metric is derived from the inverse of an organism's cumulative appearance frequency within our PDB submission dataset. The level of taxonomic granularity is that of the standard classification of the NCBI taxonomy

[4] The conclusions regarding structural novelty are robust to alternative specifications. For instance, calculating the mean TM-score based on the *n* most similar existing structures (e.g. top 5, 10, or 25 closest neighbors) yields qualitatively similar results.

database (Schoch et al., 2020). Higher values indicate that researchers are studying proteins from less commonly investigated organisms.

- **Disease relevance**: We assess the biomedical significance of protein structures by cross-referencing UniProtKB entries with disease associations documented in the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al., 2005).

These metrics can be used to indicate how AlphaFold 2 might free up experimental resources typically devoted to routine structural determinations, allowing researchers to focus on less-characterised proteins with potentially greater significance.[5]

Translational data

To assess potential associations between AlphaFold 2 adoption and translational research outcomes, we examine data on clinical applications and commercial innovation.

We utilise PubMed data, specifically the database available through iCite, to identify clinical article citations to AlphaFold 2-related publications. The iCite tool classifies publications as clinical based on their study design and content, identifying randomised controlled trials, clinical trials, and observational studies that directly involve human subjects or clinical specimens. This classification enables a measure of how research findings may be influencing clinical practice.

Additionally, we leverage PubMed's Medical Subject Headings (MeSH) tagging system (Lipscomb, 2000), focusing particularly on class C tags which specifically denote terms relevant to human diseases. This classification allows us to estimate the disease relevance of publications in our dataset, offering insights into potential biomedical applications.

For patent analysis, we employ The Lens[6], an open database from which we extract a comprehensive sample of patents filed during our period of interest. Patents represent protected intellectual property with commercial potential and frequently cite academic publications as prior art or foundational knowledge upon which the invention builds.

The Lens dataset contains these citations to academic research, enabling us to trace potential commercial impacts of academic publications. We identify approximately 85,000 patent-paper citation pairs relevant to our analysis. The Lens data also includes forward citations from other patents and Cooperative Patent Classification (CPC) designations, providing additional metrics to assess the commercial significance and technological classification of innovations potentially influenced by AlphaFold 2 and other frontier developments.

[5] OMIM (Online Mendelian Inheritance in Man) a comprehensive, continuously updated database cataloging human genes and genetic disorders, focusing on the relationships between genes and phenotypes.
[6] Accessed 18 June 2025 to collect patents up to and including Q1 2025.

2.2 Methods

To map AlphaFold 2's impact on structural biology and its diffusion into applied research, we collect and process data from the aforementioned sources (see the Appendix A and B for full descriptions of the data and variables we use. Figure 1 summarises the data creation and enrichment pipeline, described further in the rest of this section.

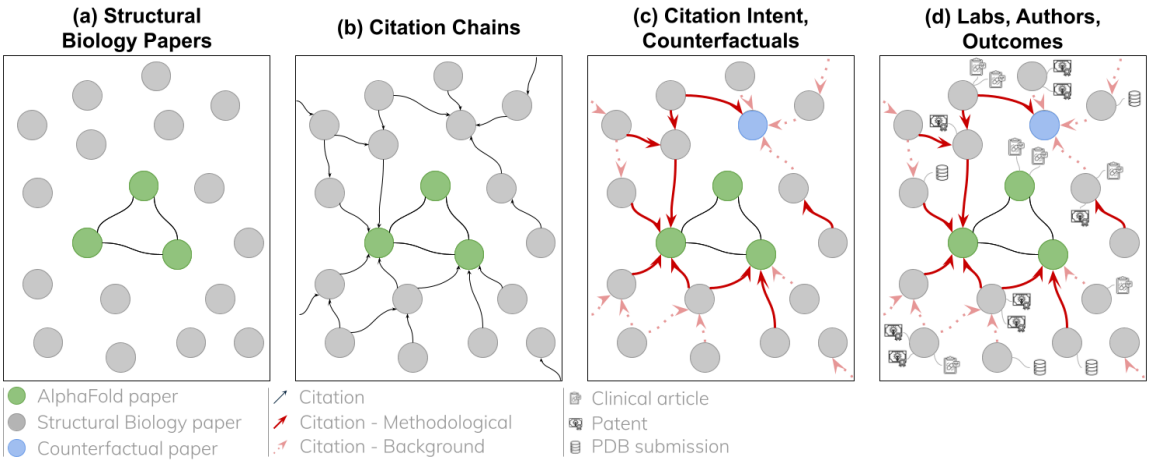

Figure 1. Overview of the dataset construction pipeline. It illustrates the multi-stage process for building our dataset, designed to track AlphaFold 2's impact from its core publications to adjacent and downstream research, as well as to applications outside the immediate research sphere.

a. We identify core AlphaFold 2 publications and collect a comprehensive corpus of papers that includes, but is not limited to, structural biology research. This corpus includes publications across multiple disciplines that engage with AlphaFold 2.

b. We create a citation network to identify adjacent and downstream papers based on their position in citation chains relative to the core publications.

c. Citation chains are labelled according to their properties, noting both direct methodological use and significant influence. This classification allows us to identify appropriate counterfactual papers that share key topics, have at least 50 citations, and exhibit a comparable distribution of both methodological and influential citations as the core AlphaFold 2 papers.

d. The network is enriched with metadata about authors and labs and linked to outcomes, including patents, clinical trials, and protein structure submissions.

Building citation chains

In order to fulfill our second design criterion ("**track direct and indirect impacts**"), we build citation chains from AlphaFold 2 publications to other research and translational outputs.

This process begins with three core AlphaFold 2 publications: Jumper et al. (2021), introducing AlphaFold 2 2; Varadi et al. (2022), formally describing the AlphaFold 2 Protein

Structure Database resource, and Evans et al. (2022), presenting AlphaFold 2 Multimer (Jumper et al. 2021; Evans et al. 2022; Varadi et al. 2022). Using a curated list of structural biology concepts, we identify a structural biology baseline that we collect from OpenAlex[7]. This creates an initial network of 102,457 structural biology papers published since 2018.

We characterise papers based on their relation to AlphaFold 2 core papers. Those which directly reference the core papers are termed 'adjacent' and those which are at least one additional degree of separation further from the core papers are termed 'downstream'. Articles with both a direct and non-direct citation chain to the core papers are assigned by their closest link to ensure non-overlapping datasets.

Citation intent

To account for the varied ways in which AlphaFold 2 is used, each citation link is classified according to its intent and influence, which is a crucial component of our study design. Not all citations are equal; they can serve multiple purposes. Using data from Semantic Scholar, we distinguish between citations that are background references (providing context), cite research results (referencing findings), or, most importantly for our analysis, represent direct methodological uses (employing a technology as a tool in new research). A separate flag also identifies influential citations, where the cited work had a significant impact on the citing publication. This classification allows us to focus on the subset of citations reflecting a direct contribution to the advancement of the field, rather than those merely acknowledging a paper's existence.

This detailed citation data is available for approximately 36% of publications in the AlphaFold 2-connected chains and 32% for those not connected. This incomplete coverage is a known limitation of the method, as the labelling algorithm relies on having access to the full text of open-access publications. The available data reveals a high degree of methodological adoption for AlphaFold 2, with 68% of foundational papers using the technology as a method, compared to 40% of applied, downstream publications. To estimate these figures across our entire dataset, we use a stratified bootstrapping technique (detailed in Appendix G) that resamples from the available data based on a publication's primary field, date, and its proximity to a core paper in the citation chain.

Researchers and their characteristics

Our third design criterion is to **capture different dimensions of analysis**, so in addition to collecting data about individual papers, we compile data on all authors from our citation-chain dataset, yielding 5,774,183 unique researchers. Scientists differ in their readiness and access to resources to adopt new tools so we distinguish three levels of experience:

[7] The structural biology concepts used are available to view in this project's GitHub repository.

- **Established researchers:** Authors with publication records predating 2020 who appear either in our baseline data or in any citation chain. For this group, we consider as their output any publications in which they appear as first-author, reflecting their direct intellectual contribution to the research.
- **Laboratory leads:** Principal investigators who head research groups, identified through their consistent senior authorship positions. We identify PIs by frequent last-author contributions, a standard marker of leadership in scientific fields. This approach relies on an author position factor that weights first, middle, and last-author positions across multiple years. Importantly, we consider as the output of the laboratory all publications with the PI in them, irrespective of position[8].
- **Early Career Researchers (ECRs)**: Authors who first appear in our dataset in 2020 or later. While we collect data on this group, their limited publication history and the recency of their careers restrict statistical power for many analyses. Results for this subgroup are therefore primarily presented in the Appendix rather than the main text.

To ensure data quality, we apply several filtering steps throughout the author processing pipeline. Authors without valid authorship information, ambiguous IDs, or sufficient institutional details are excluded. These steps reduce the initial pool of 5.8 million researchers to 435,265 established researchers authoring 8,051,282 unique publications and 403,098 early career researchers producing 1,488,519 unique papers. At the laboratory level, we identify 33,862 likely principal investigators and collect 5,355,928 publications associated with their teams.

A full breakdown of outputs can be found in Appendix Table 1. See Section D of the Appendix for additional details on these classifications, including how we refine laboratory associations

Counterfactual comparisons

We seek to fulfil our fourth design criterion (**"provide a comparative context"** for AlphaFold 2 impact) by comparing the performance of papers, researchers and labs who cite AlphaFold 2 with a structural biology baseline, and by comparing AlphaFold 2 impacts with three distinct sets of counterfactual high-impact structural biology papers: AI-intensive developments, non-AI protein prediction developments, and other frontier structural biology methods. We choose to distinguish between these innovation pathways in order to assess the value of AlphaFold 2 both as an exogenous shock to "business as usual" structural biology and as an AI tool that sits alongside a range of other, potentially impactful AI and non-AI developments.

[8] There is some overlap between principal investigators and established researchers, as the former may sometimes publish as first authors or co-author with other established researchers.

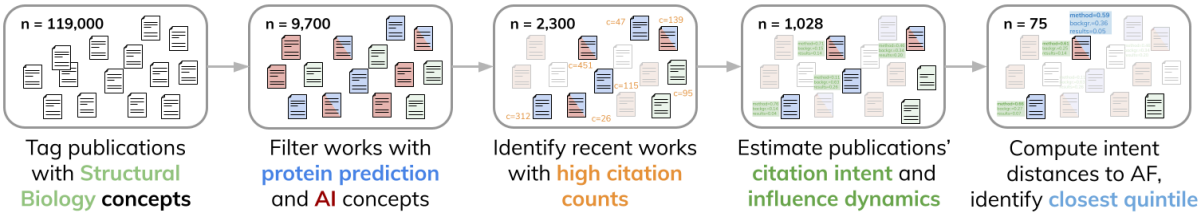

Figure 2. Selection process for identifying relevant papers in structural biology using OpenAlex concepts, CWTS topics, and citation intent metrics from Semantic Scholar.

As shown in Figure 2, the selection process begins by identifying recent works in OpenAlex tagged with OpenAlex concepts related to structural biology. Candidate papers are further filtered using CWTS topics and OpenAlex concepts (tags which proxy for areas of scientific research at various levels of granularity) to focus on those relevant to protein prediction (C18051474, C10010492), artificial intelligence (C154945302), or both.

To increase comparability with AlphaFold 2, we consider the candidates' citation usage patterns. Specifically, we focus on the proportions of methodological and influential citations for each paper relative to AlphaFold 2's citation patterns. We calculate the Euclidean distance between these distributions and select papers in the bottom quantile for their similar usage patterns. From this refined set, we manually curate a list of 90 high-impact papers, evenly divided among the three counterfactual categories. The vast majority of these papers (66 of the 75 frontier developments) are in the top 0.1% of structural biology papers according to citation counts.

Frontier work in: **Protein Prediction & AI**	A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core
	Improved protein structure prediction using predicted interresidue orientations
	HH-suite3 for fast remote homology detection and deep protein annotation
	Unified rational protein engineering with sequence-based deep representation learning
	QMEANDisCo—distance constraints applied on model quality estimation
Frontier work in: **Protein Prediction**	IUPred2A: context-dependent prediction of protein disorder as a function of redox state
	DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability
	An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking
	Structure-based design of prefusion-stabilized SARS-CoV-2 spikes
	PANTHER: Making genome-scale phylogenetics accessible to all
Frontier work in: **Structural Biology**	Coactivator condensation at super-enhancers links phase separation and gene control
	mRNA structure regulates protein expression through changes in functional half-life
	The SARS-CoV-2 spike protein alters barrier function in 2D static and 3D microfluidic in-vitro mo.
	Structural and Chemical Biology of Terpenoid Cyclases
	Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV2

Figure 3. Titles of randomly selected publication examples in citation chain samples, including AI and non-AI protein prediction frontier developments, as well as other structural biology innovations.

Regression analysis

Our final design criterion seeks to ensure that, as much as possible, we identify associations between building on AlphaFold 2 and research outcomes whilst accounting for systematic differences between studies or researchers that cite it and those that do not. We compare the relationships that using AlphaFold 2, or alternative AI and non-AI frontier developments, has with researchers' outputs and downstream work compared to the baseline of other structural biology research publications.

In our publication-level analysis, we account for potential confounding factors by employing a stringent specification that includes time and author fixed effects for every publication in our dataset. This approach helps control for unobserved time-invariant characteristics of individual researchers and temporal trends that might influence research outcomes independently of AlphaFold 2 adoption. Additionally, we include controls for all primary fields a publication may cover, capturing distinct differences in outputs as they relate to key disciplinary contexts.

For researcher and laboratory analyses, we employ a difference-in-differences framework to track how outcomes change for those who make use of AlphaFold 2 compared to those who do not, before and after their first citation to the tool. We collect data on a quarterly basis and consider as key variables the indicator variables of having previously cited one of four groups (AlphaFold 2, other AI frontier developments, non-AI protein prediction frontier developments, and other structural biology frontier research). We also track strong citations, and for all specifications that only consider intent data, we include only researchers and laboratories who have a sizable portion of their data with intent information. This approach helps minimise issues stemming from wrongful inclusion of data for which we have no actual information of intent, though it substantially reduces our sample size.

Recognising that adoption of AlphaFold 2 is not randomly assigned, our empirical strategy includes several validation steps to maximise the robustness of this approach. To increase the probability that measured associations stem from adoption of or links to AlphaFold 2 rather than pre-existing differences, we match researchers and laboratories on institutional location, prior publication and citation patterns, discipline and field specialisation. We implement this matching with Coarsened Exact Matching (CEM), a method that groups and pairs researchers with similar profiles before comparing the outcomes of those who build on AlphaFold 2 to those who do not.

After balancing on levels, we test for the common trends assumption through multiple methods. Visual inspection of annual trend plots (Appendix Figures A.1 and A.2) confirms that pre-treatment trajectories for outcomes are largely parallel across groups. We further substantiate this with placebo tests, which displace treatment dates to earlier periods and only show significant pre-existing trends for publication counts. Finally, we implement staggered event study designs (Callaway and Sant'Anna, 2021), which show statistically

insignificant coefficients for all pre-treatment periods, providing strong evidence against anticipatory effects (Appendix Figures A.14 and A.15).

As with our publication-level analysis, we control for different dynamics across academic fields and introduce stringent researcher/laboratory and time fixed effects across all specifications. This approach helps account for any remaining unobserved time-invariant characteristics and temporal trends that might influence research outcomes independently of AlphaFold 2 exposure.

Appendix Table 2 shows that after matching on baseline characteristics, the groups were reasonably comparable at the researcher level, although laboratory-level differences were more pronounced, with those influenced by AlphaFold 2 reporting higher average PDB submissions and patent activity. This means we cannot rule out that our findings may be driven by systematic differences across groups, although our difference-in-differences approach, which focuses on changes in pre-exposure trends, seeks to account for this.

After applying these validation and matching procedures, our main analysis compares outcomes for researchers developing on top of AlphaFold 2 against three distinct counterfactual groups: those building on other AI-intensive protein prediction developments, non-AI protein prediction methods, and other frontier structural biology innovations. This four-way comparison allows us to distinguish whether observed associations are specific to AlphaFold 2 or reflect broader patterns related to AI adoption or methodological innovation in general. We examine various outcomes including publication rates, citations, PDB submissions, and translational metrics across all groups.

Our models include linear regressions (often log-transformed) and Poisson regressions for count data, both of which help us interpret coefficients as approximate percentage changes in our primary outputs. Full technical details of the empirical design are available in Section B of the Appendix. For each outcome, we present results in three complementary ways:.

- First, we compare works and researchers with links to AlphaFold 2 with the structural biology baseline to establish fundamental patterns.
- Second, we contrast those with links to AlphaFold 2 with those building on other frontier techniques to determine whether observed effects are specific to AlphaFold 2 or reflect broader trends in methodological innovation.
- Third, we often explore complementarities with researcher experience and citation intent, examining whether associations differ between experienced and less experienced researchers, and between those using AlphaFold 2 methodologically versus citing it as background knowledge. Additionally, we occasionally report field-specific analyses for biochemistry and medicine-focused researchers and labs to identify potential disciplinary variations in AlphaFold 2's relationship with research outcomes.

3. Findings: Scientific reach

- AlphaFold 2's overall reach including direct and indirect citations (up to 3 degrees of separation from AlphaFold 2 focal papers) is almost half a million papers.
- We estimate that 68.1% of adjacent research outputs methodologically cite AlphaFold 2. A further 38.1% of downstream research also relies on these AlphaFold 2-powered papers, amounting to a total of over 218,000 papers built strongly on the technology.
- The number of publications citing AlphaFold 2 continues to show strong growth over time. This contrasts with the less steep or even linear growth rates of other frontier developments.
- Adjacent papers tend to concentrate in biochemistry and molecular biology, while downstream works show a clear expansion into more applied disciplines, particularly the medical sciences and engineering.
- We estimate that over 778,000 scientists have worked on research that builds strongly on AlphaFold 2. The number of scientists that have made use of AlphaFold 2 directly and indirectly continues to grow very rapidly.
- Most biochemistry papers directly cite AlphaFold 2, while research in applied areas like medicine and agriculture tends to cite papers that build on AlphaFold 2. This suggests its use in foundational research enables later applications.
- AlphaFold 2 uniquely bridges the gap between foundational research and practical medical applications, with its research being 37% more likely to be primarily focused on medicine compared to other AI-intensive frontier techniques.

A transformative tool is ultimately measured by the new science it enables. This section, therefore, seeks to characterise its broader scientific reach, quantifying how, where, and how quickly it has been integrated into the research ecosystem. This provides a detailed portrait of how one technological leap can spur innovation across the scientific landscape.

The network associated with AlphaFold 2 alone comprises 40,808 adjacent papers and 639,458 downstream publications. When combined with the citation chains from our counterfactual innovations, the total network includes 3,731,130 unique publications. This covers any citations made between the time of publication and September 2025. This layered network tracks AlphaFold 2's reach within research, with adjacent papers tending to capture work concentrated in structural biology, and derivative works exhibiting expansion into other disciplines, including further downstream studies in the medical sciences.

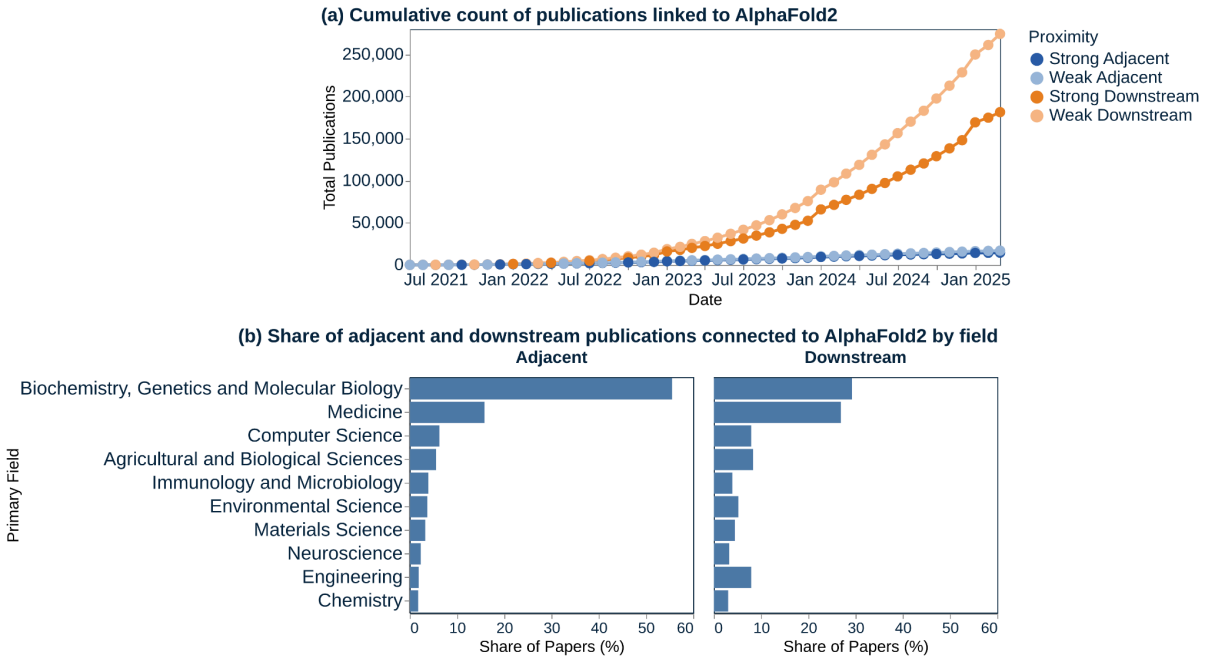

Figure 4. Papers are classified as adjacent and downstream the core AlphaFold 2 papers, with the strength of their citation chains based on citation intent. a) The total number of AlphaFold 2 related publications. Counts are estimated using a bootstrapping method. b) The share of adjacent and downstream publications for the top 10 primary fields of papers linked to AlphaFold 2 by count. All figures include papers up to and including March 2025.

We trace the reach of AlphaFold 2 both in terms of proximity and influence in Figure 4 which shows how a steady increase in the total number of papers with direct links to AlphaFold 2 has been accompanied by a diffusion of links to a wider body of downstream publications. Among adjacent papers, we estimate that 68.1% (95% CI [67.4, 68.8]) have at least one outgoing methodological (strong) citation to AlphaFold 2 up to March 2025, which serves as our baseline for defining its use as a tool or method. A further 38.1% (95% CI [37.0, 39.2] of downstream research works also rely methodologically on these AlphaFold 2-powered papers, highlighting the extent to which researchers are creating new research outputs from subsequent methodological developments built on AlphaFold 2.

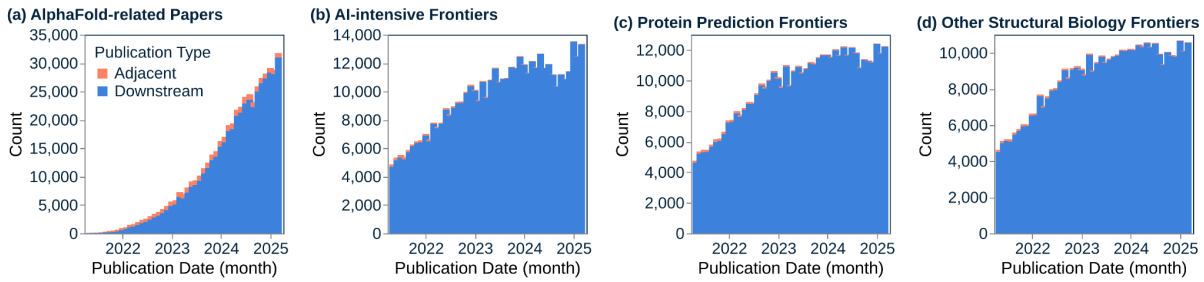

Figure 5. Monthly counts of new publications across three groups: (a) publications stemming from AlphaFold 2, (b) AI-intensive protein prediction frontier developments, (c) non-AI intensive protein prediction frontier developments, and (d) other structural biology frontier developments.

Citation chains from AlphaFold 2 and counterfactual papers capture a wide view of structural biology and span multiple adjacent fields. The downstream influence of AlphaFold 2 appears to be steadily increasing, with the number of monthly publications still growing strongly, as seen in Figure 5. Downstream publications have grown rapidly and make up 94% of our sample to date. In contrast, while downstream papers are similarly represented, other frontier developments or structural biology citation chains show less steep or even linear growth in downstream influence. By the end of our sample, nearly half of the monthly publications were associated with AlphaFold 2, which underscores the technology's profound and growing impact.

3.1 Mapping AlphaFold 2's Influence

Although citation chains help trace AlphaFold 2's immediate influence on structural biology, they may understate the full scope of its adoption. Researchers who cite AlphaFold 2 once may not consistently cite it in subsequent work, even if they continue using its predictions. To address this, we track adoption at the individual level, measuring how researchers' work shifts following their first AlphaFold 2 citation. This allows us to capture outcomes such as ongoing publications, patent activity, and experimental work that may rely on AlphaFold 2's tools, even if citations become less explicit over time.

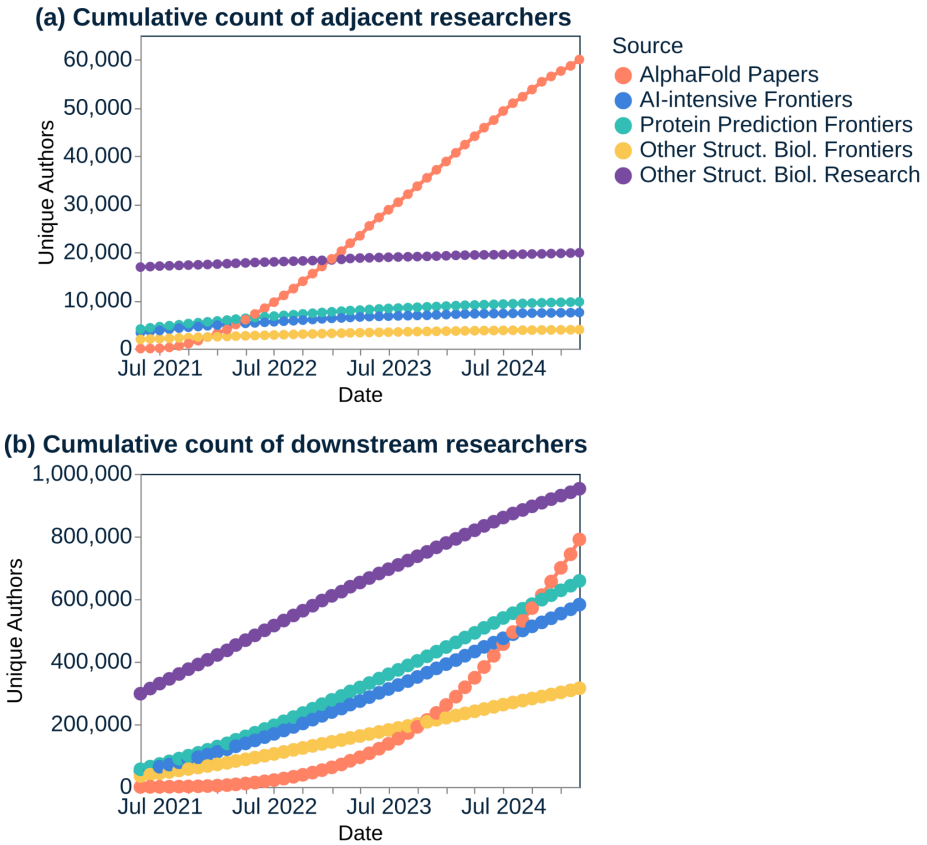

Figure 6. The estimated number of unique researchers that are associated with papers that have strong methodological links back to their respective core papers. Counts are estimates based on the mean value after bootstrapping the sample of data with available citation intent information.

AlphaFold 2's spread has extended its methodological influence to over 778,000 researchers working on adjacent (62,501, with 95% CI [60,063, 64,938]) and downstream (726,278, with 95% CI [708,996, 743,560]) researchers, directly or indirectly. By early 2025, the influence of AlphaFold 2 on adjacent publications continued to grow, as seen in Figure 6, while growth in direct influence of other frontier developments appear to have somewhat stabilised. AlphaFold 2's downstream influence shows a pattern of fast growth, with its influence continuing to expand through applications. This trajectory seems to indicate sustained diffusion across structural biology and closely related fields and the enduring relevance of AlphaFold 2 compared to other frontier works. It should be noted that these numbers do not necessarily represent currently active researchers, nor are they indicative of ongoing use of AlphaFold 2.

3.2 Disciplinary focus

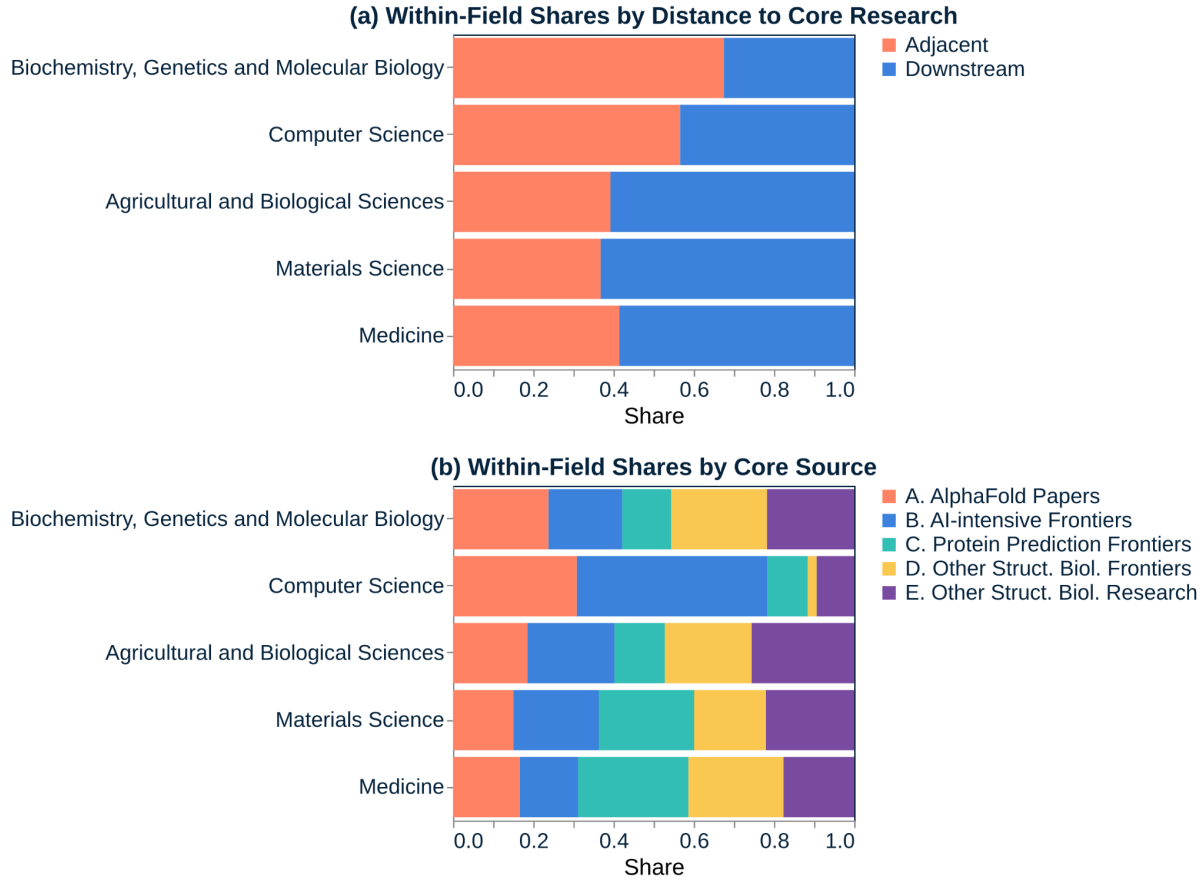

Figure 7. Share within fields of papers for AlphaFold 2 and counterfactual papers: (a) of 'adjacent' vs 'downstream' relative to core research, (b) within-field shares by source types.

AlphaFold 2's use is distributed unevenly across disciplines. In adjacent papers, the primary field of the majority of publications is Biochemistry, Genetics and Molecular Biology (53.4%), with Medicine accounting for the next largest share (15.2%). Downstream papers are less concentrated, with those two fields accounting for 26.5% and 24.6% of

papers respectively. While the numbers are comparable across other fields, this highlights how downstream papers are generally more applied in nature. There are also some notable differences, such as Engineering whose share of papers almost triples (from 1.8% to 7.3%) in downstream papers, pointing to the translatable aspects of AlphaFold 2. These differences in research focus highlight the importance of tracing downstream citation chains, which capture the broader and more diverse impacts of frontier developments like AlphaFold 2, particularly in clinical and translational applications.

The disciplinary focus of research citing AlphaFold 2 differs from the patterns observed in papers citing our counterfactual frontier developments. As Figure 7 shows, the majority of biochemistry citations are adjacent to AlphaFold 2 and other frontier developments (i.e. cites them directly). By contrast, research in potential application areas such as Agricultural research or Medicine are further downstream. This is consistent with the idea that researchers are using AlphaFold 2 or other innovations for adjacent research that might be enabling downstream research applied to health and other domains.

Unlike traditional AI-intensive innovations, which are primarily cited in computational research, AlphaFold 2 demonstrates broader applicability. As shown in part (b) of Figure 7, its influence extends into experimental and applied domains such as medicine, where adoption often depends on innovations achieving a high degree of maturity (McNamee and Ledley 2017). Notably, AlphaFold 2-derived research is 37% more likely to have medicine as its primary topic compared to other AI-intensive frontier innovations. By facilitating advancements in protein structure prediction, AlphaFold 2 seems to provide a direct basis for foundational research that applied studies build on, potentially helping bridge the gap between theoretical breakthroughs and real-world challenges. This seems to suggest that AlphaFold 2 may sufficiently solve a problem that matters significantly to structural biologists, rather than simply be a high performing advancement in AI techniques that is more relevant to computer science. We explore this result in further detail in Section 5.

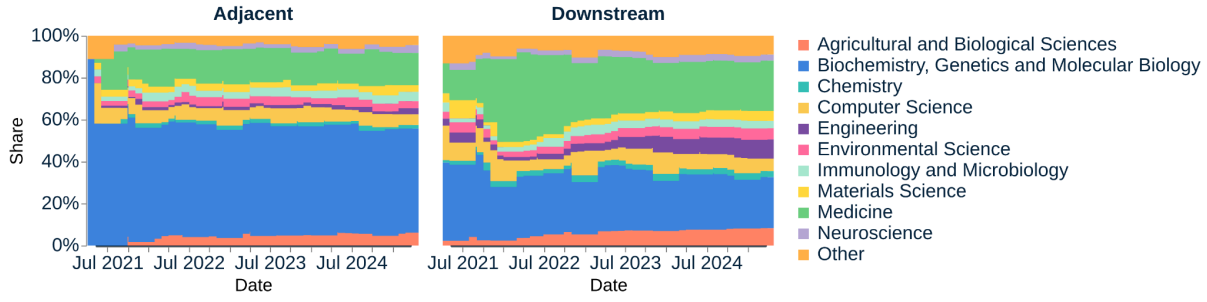

Figure 8. Monthly distribution of AlphaFold 2-related publications by primary topic (July 2021 - March 2025). The plots are faceted to show trends for adjacent papers, which directly cite core AlphaFold 2 research, and downstream papers, which are at least one degree of separation further.

4. Findings: Experimental protein structure determination

- researchers building on AlphaFold 2 demonstrate 45% to 49% higher rates of protein structure submissions across various measures compared to baseline structural biology researchers.
- Laboratory adoption of AlphaFold 2 corresponds to significant increases in experimental output, with associations of a 27% increase for PDB submissions and a 44% increase for novel structure submissions.
- Building on AlphaFold 2 is linked with the production of structures for more structurally novel proteins across multiple metrics (fold, sequence, and global shape), though this is accompanied by a potential trade-off in lower experimental resolution.
- A strong association with disease-relevant research is observed, with researchers influenced by AlphaFold 2 more than doubling the number of disease-linked structures at the researcher level and increasing them by 39% among labs.
- The positive associations between building on AlphaFold 2 and these productivity and novelty metrics consistently distinguish it from other frontier techniques, whose correlations are smaller, less statistically significant, or absent entirely.

AlphaFold 2's ability to predict protein structures with high accuracy can significantly impact experimental research, primarily by providing reliable templates for interpreting experimental data. This fundamental contribution may be helping lower barriers to structure determination, particularly for novel proteins where the absence of homologous structures previously hindered experimental work. This association might manifest in two key dimensions.

First, by providing a large repository of predicted structures, AlphaFold 2 can promote more efficient resource allocation toward experimental research. These computational templates may assist researchers in interpreting crystallographic or cryo-EM data more effectively, suggesting potential for improvements in the success rate (or economic returns) on research funding and the pace of structural determination.

Secondly, the availability of reliable structural predictions, along with clear indications of prediction uncertainty that AlphaFold 2 also provides, may be associated with researchers

exploring beyond well-characterised protein families. This expansion of feasible research targets potentially helps mitigate the 'streetlight effect', wherein scientists predominantly investigate areas with readily available structural data (Tranchero et al. 2022). This broader exploration may, in turn, increase the diversity of projects undertaken and diminish the risk of overlooking valuable leads. AlphaFold 2's predictions may also facilitate the study of complex or poorly understood structures, improving the chances of identifying unique proteins. We measure this novelty through multiple lenses: using Template Modeling Scores (TM Scores)[9] to assess fold-level novelty, the fraction of sequence identity[10] to identify distant homologs, and RCSB's PDB's shape similarity scores[11] to capture novelty in global topology. Such precision could lead to meaningful discoveries, particularly when these newly characterised structures involve proteins implicated in human diseases.

4.1 Experimental Activity

Descriptive analysis

We begin by examining the relationship between building on AlphaFold 2 and experimental work in structural biology at the publication and researcher level. Figure 9 presents three measures of experimental activity associated with publications, shown as 4-quarter rolling averages: the number of linked UniProt structures, the number of submissions to the PDB, and the number of primary submissions (the first time a structure is mapped and identified in UniProtKB, which we refer to as novel submissions). These averages are calculated as the share of submissions associated with each type of frontier development relative to their respective research output: AlphaFold 2-related research, other AI-intensive frontier developments, non-AI frontier developments, and the broader structure biology baseline. The descriptives are presented at both the citation chain level and the researcher level, which includes both early-career and established researchers.

At the publications level, the share of submissions associated with AlphaFold 2-related research grows substantially across all three metrics: these proportions increase beyond those observed in other methods and thus represent the highest relative share of UniProt structures, PDB submissions, and novel structure submissions. Although these proportions begin to decline after this peak, this pattern may reflect two factors. First, the emergence of new methods not covered in our data could affect these shares. Second, and perhaps

[9] The TM Score is a metric used to evaluate the structural similarity between two protein models. Higher TM Scores indicate greater structural similarity to known proteins, while lower scores suggest more novel structures with limited similarity to previously characterised proteins.
[10] The fraction of identity measures the proportion of identical amino acids in the aligned regions of two protein structures. It is a metric of sequence similarity within a structural alignment. A high score indicates a close evolutionary relationship, while a low score, particularly when paired with a high TM-score, suggests the discovery of a distant homolog.
[11] The RCSB PDB Shape Score is a metric that evaluates the similarity of two proteins based on their global 3D shape, independent of their sequence or specific fold topology. This alignment-free method compares volumetric properties; high scores indicate a strong match in overall shape.

more significantly, the downstream adoption of these methods in non-structural biology research may be expanding the denominator of AlphaFold 2-related publications without corresponding experimental submissions.

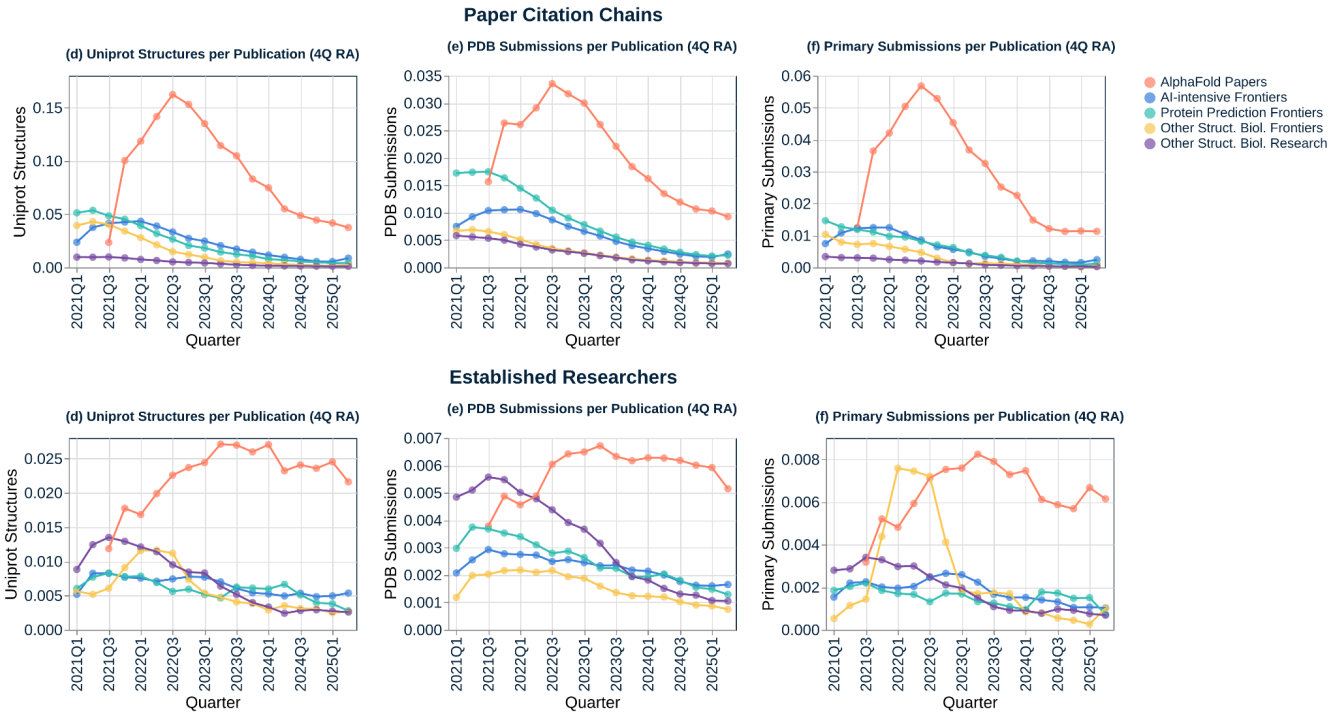

Figure 9. Numbers of UniProt structures, PDB submissions, and primary structure submissions per publication in the citation chains or researcher (both early-career and established) data. Values are presented as 4-quarter rolling averages.

The researcher and lab-level data reveals a complementary pattern. While publication-level data show a decline in submission shares over time, researchers who adopt or have used AlphaFold 2 and its downstream applications continue to exhibit steady growth in their proportion of experimental structures. These researchers see sustained increases in their shares of UniProt structures, total PDB submissions, and especially novel submissions. These patterns suggest that using AlphaFold 2 is associated with strong growth in experimental activity, particularly for novel proteins.

Regression analysis

While the descriptive patterns are informative, they may reflect underlying differences in researcher characteristics, field-specific trends, or temporal changes in structure determination activity. Our regression framework, introduced earlier, allows us to examine these relationships while accounting for these potential sources of bias.

We organise our findings in three parts: first, comparing frontier developments against the structural biology baseline; second, examining differences across frontier development types; and third, exploring complementarities with researcher expertise and the robustness of our results to alternative specifications of methodological adoption.

Comparison with structural biology baseline

Our regression results confirm systematic differences in experimental structure determination between AlphaFold 2 influence and the broader structural biology research community.

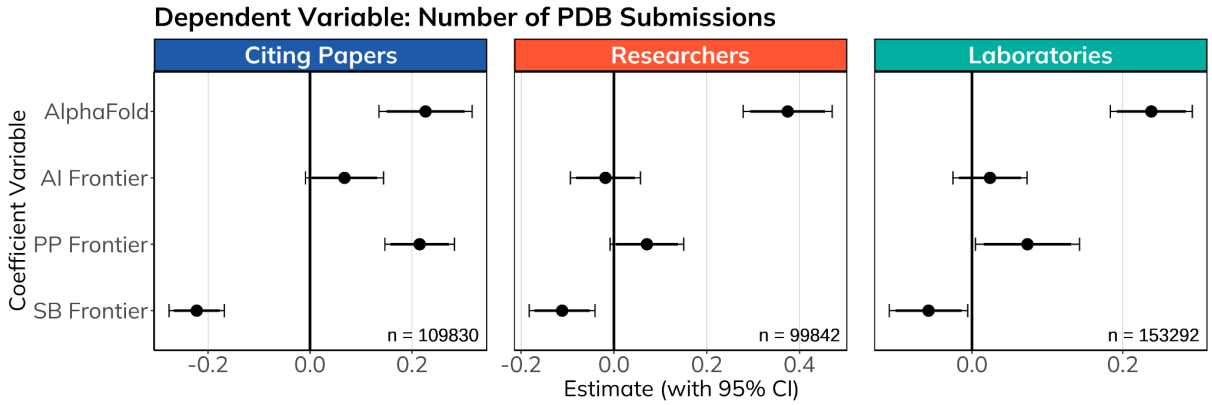

Figure 10. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

Figure 10 presents results for PDB submissions, where we observe a fairly consistent pattern across research units. Citation chains show a significant association comparable to that of non-AI protein prediction frontier methods. Individual researchers employing AlphaFold 2 exhibit substantial differences in submission patterns, with increases of 45.3% compared to baseline structural biology researchers[12]. At the laboratory level, teams using AlphaFold 2 demonstrate a 26.8% higher submission rate. The attenuated laboratory level associations may reflect the complexity of team adoption dynamics, where not all members actively employ these tools, compared to individual researchers who are identified through direct usage. In the latter two groups, alternative protein prediction algorithms show muted or minor positive associations, while other structural biology frontier research is negatively associated with submissions to the PDB.

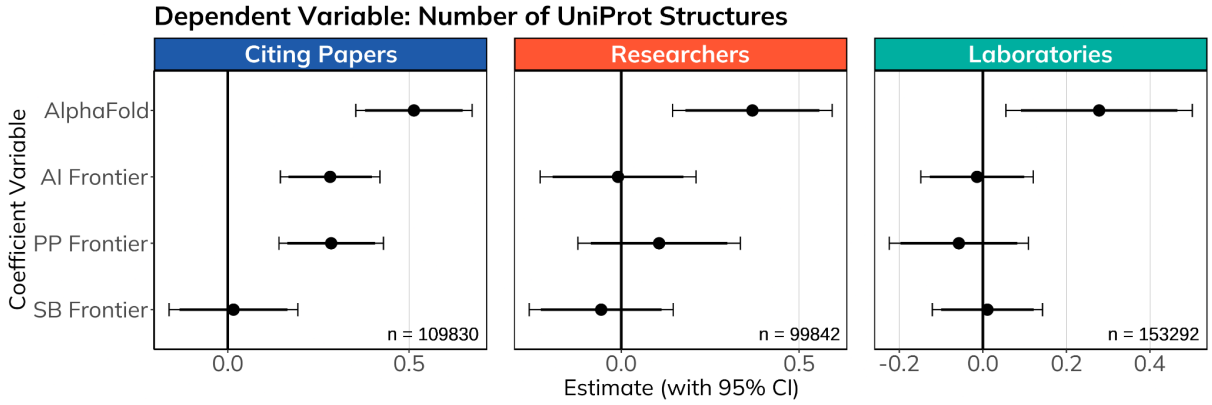

Figure 11. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

[12] In Poisson regression, coefficients represent the log of the expected count. To interpret them in terms of percentage change, the coefficient is exponentiated, ie. exp(0.374) ≈ 1.453.

If we consider submissions at the protein chain level, as defined by distinct UniProt identifiers, our analysis reveals similarly strong associations with building on AlphaFold 2 (Figure 11). In this context, we measure the number of distinct UniProt sequence entities for which a researcher or lab submits any experimental structure during our study period. This metric is designed to capture all contributions to the structural knowledge base, including both the first-ever structures for a given protein and subsequent structures that expand upon what is known.

Using this measure, we find that researchers employing AlphaFold 2 show substantially higher rates of structure determination, with an 44.6% higher output of distinct protein structures compared to baseline researchers. At the laboratory level, building on AlphaFold 2 corresponds to a 32.1% difference for the same outcome.

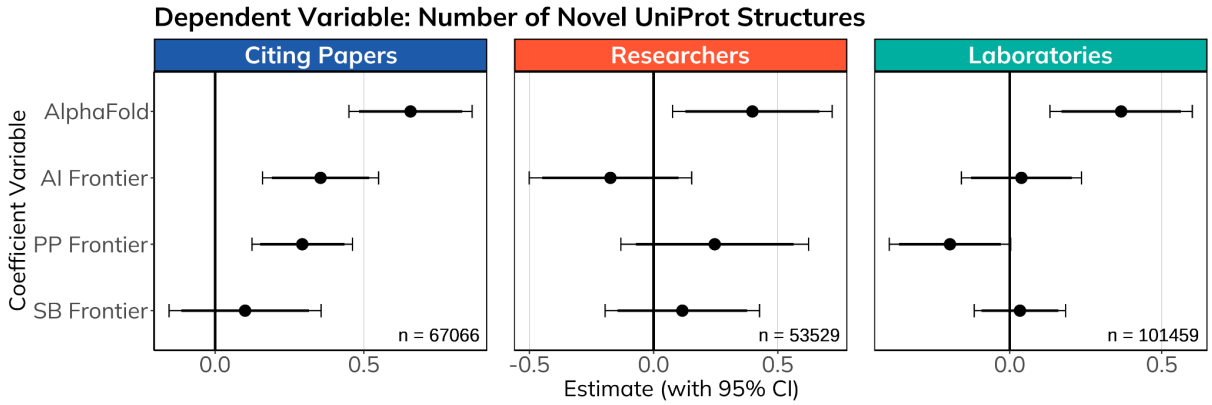

Figure 12. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

Figure 12 examines primary or novel submissions, which represent first time appearances of structure chains in UniProt's Knowledge Base. Researcher level analysis shows substantial positive associations with building on AlphaFold 2 (48.8%), whilst laboratory level analysis indicates an 44.3% increase for AlphaFold 2 influence.

These findings suggest systematic differences in experimental structure determination patterns between AlphaFold 2 influence and baseline structural biology researchers. The variation between individual and laboratory level results may reflect differences in adoption intensity and the complexities of team based research practices.

Comparison with other frontier techniques

The comparison of building on AlphaFold 2 with other frontier developments helps to address potential selection bias in our previous analysis. Researchers and laboratories that adopt novel methods may systematically differ from the broader research community in ways that affect experimental output. By examining differences across types of frontier development, we can better contextualise the associations reported above.

Papers relying on other protein prediction methods show an association with PDB submissions similar to that of papers using AlphaFold 2, but the correlation is roughly half

for chains identified in the UniProt Knowledge Base, both existing and new. Among other AI frontier methods, we observe only a weak association with PDB submissions and similarly subdued associations with UniProt Knowledge Base chains. For other frontier structural biology research, there is a negative association with PDB submissions and non significant associations with UniProt Knowledge Base chains.

At the researcher level, the picture for other frontier methods is considerably different from that of AlphaFold 2. For most of these alternative methods, both AI and non AI, we find no statistically significant association with experimental outputs. A similar pattern is observed at the laboratory level, where most correlations are either non significant or weak.

These distinctions at both user levels, which persist after controlling for individual or team fixed effects, suggest our findings are not simply driven by a selection bias where successful researchers and teams are more likely to adopt any frontier technology. Were that the primary driver, we would expect to see comparable positive associations from the other AI and non AI innovations, which is not what the data show. Nonetheless, we interpret these associations with caution, as our model may not fully account for all potential time varying unobserved confounders.

Complementarities with methodological adoption

Subgroup analysis by methodological usage allows us to explore whether the nature of engagement with AlphaFold 2 is associated with different outcomes. However, it is important to interpret these findings with caution. Our classification of methodological use relies on citation intent algorithms, which can introduce noise and serve only as a proxy for the actual use of the technology in a research workflow. Complementarities with researcher experience are also available in Appendix Figures A.10 and A.13.

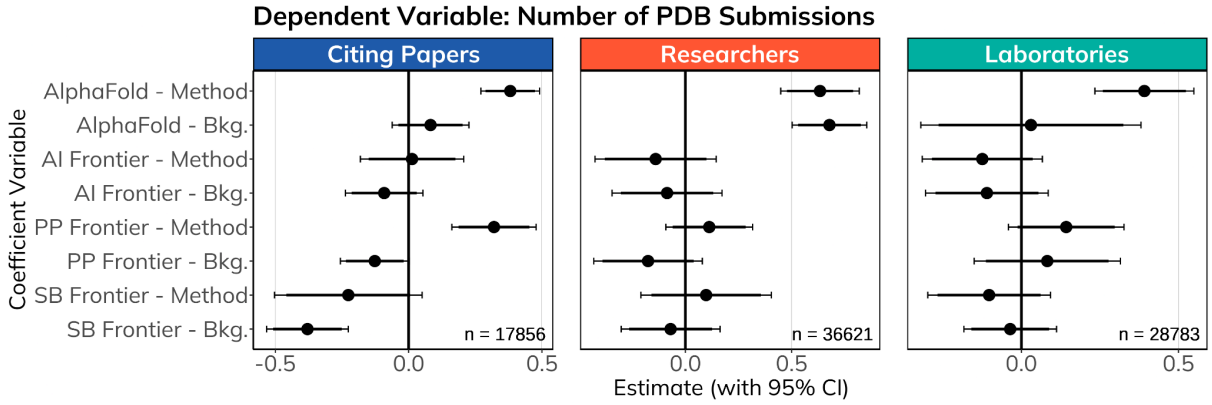

Figure 13. Coefficient estimates from Poisson regressions showing links for highly-experienced biologists. Bars represent 95% confidence intervals.

For PDB submissions, our analysis at the paper level shows a clear divergence based on usage type (Figure 13). The methodological use of AlphaFold 2 is associated with a large, positive, but statistically insignificant correlation with submissions. This stands in contrast

to the non-significant association observed for papers that cite it as background. Alternative frontier research groups show either no association (AI) or a positive association for methodological use only (non AI). At the researcher level, both methodological and background use of AlphaFold 2 are associated with strong positive correlations, while no other group displays any significant association.

The results for laboratories show a strong link between methodological use of the technology and no positive association for background use, while other frontier methods have no significant correlations. Additional specifications, available in the project's code repository, suggest this pattern is related to opposing trends: for foundational (adjacent) research, methodological adoption is associated with higher submission rates compared to background use. This association is largest when considering experimental labs, which we define based on their history of PDB submissions.

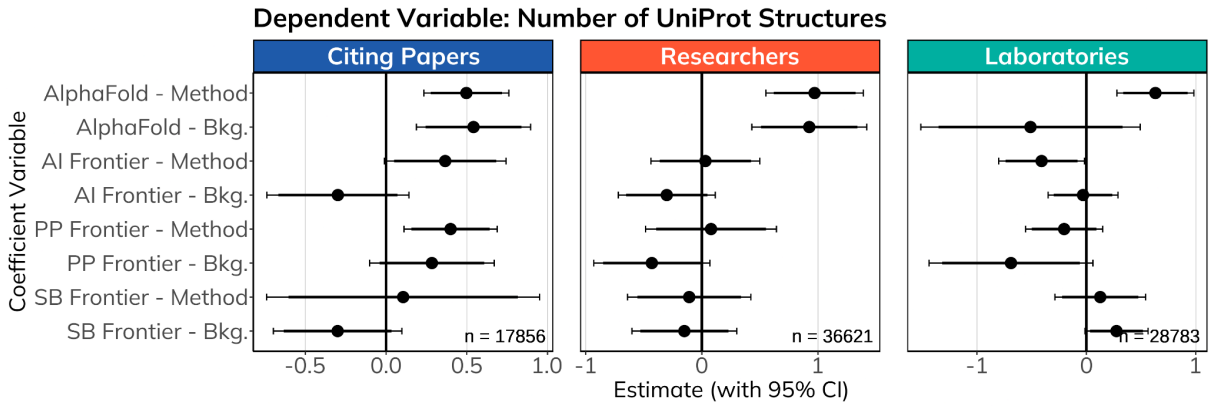

Figure 14. Coefficient estimates from Poisson regressions showing links for highly-experienced biologists. Bars represent 95% confidence intervals.

For UniProt submissions (Figure 14), positive associations at the publication level are seen for AlphaFold 2 use (both types) and the methodological use of other frontier methods. This association narrows at the researcher level to only AlphaFold 2 use, and further at the laboratory level to only the methodological use of AlphaFold 2.

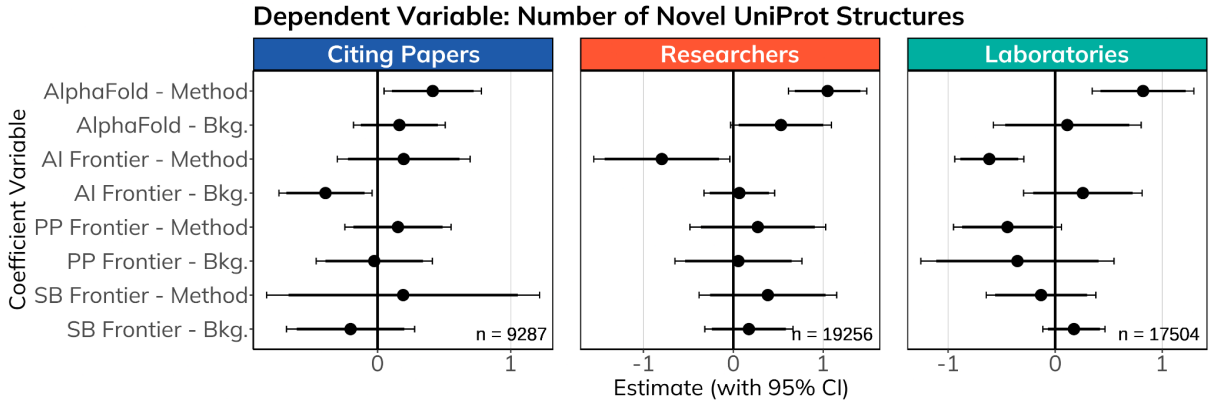

Figure 15. Coefficient estimates from Poisson regressions showing links for highly-experienced biologists. Bars represent 95% confidence intervals.

For novel protein structure submissions (Figure 15), a positive association is found almost exclusively with the methodological use of AlphaFold 2. This association becomes particularly large and strong for researchers and laboratories, whereas background use shows a much weaker, or non-significant, link. In contrast, other frontier methods show either negative or non-significant correlations.

Notably, when comparing these outcomes with our proxy for experimental expertise (Figures A.10, A.13), we find no evidence of additionality from being a high PDB submitter. This suggests the benefits of methodological adoption for these outcomes are not confined to experienced experimentalists, a finding that contrasts with other studies suggesting that such users may be best poised to leverage new AI technologies (Yu, 2024).

4.2 Protein structures

Descriptive analysis

Building upon our analysis of AlphaFold 2's association with increased experimental structure determination, we now examine the qualitative characteristics of these structures. If AlphaFold 2 is genuinely expanding research capabilities rather than merely accelerating existing workflows, we would expect to observe a shift in the types of protein structures being investigated. We assess this by exploring several key dimensions: structural novelty (measured at the fold, sequence, and global shape levels), the taxonomic diversity of source organisms, relevance to human disease, and any potential trade-offs in experimental precision.

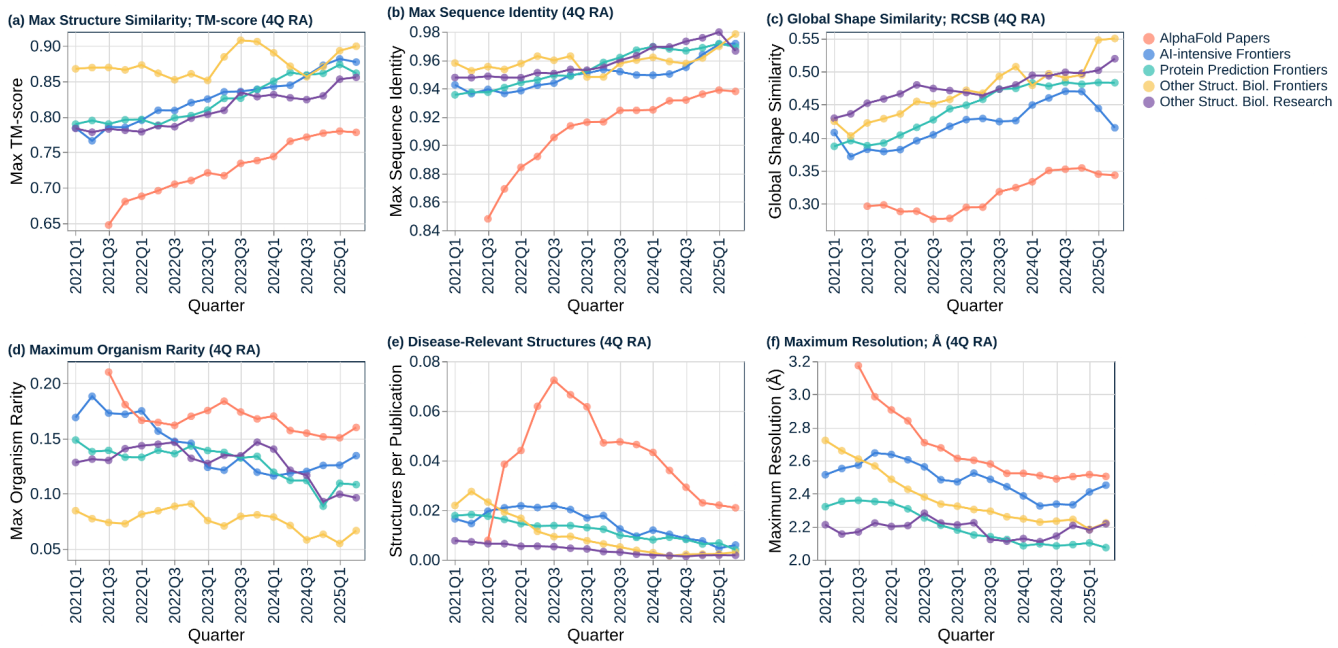

Figure 16. Citation chain trends. Organism rarity is the inverse frequency of UniProtKB organism IDs. TM Scores measure structural similarity, with higher values indicating greater resemblance. Disease relevance reflects the likelihood of PDB submissions being linked to human diseases.

Our analysis of quarterly trends in Figure 16 reveals distinctive patterns in the structures submitted by those building on AlphaFold 2. This group consistently produces structures with lower TM-scores, lower fractional identity, and lower RCSB shape scores, indicating a sustained focus on proteins that are novel across multiple dimensions. Complementing this, we observe a clear trend towards targeting proteins from rarer organisms[13] and a notable spike in the submission of disease-relevant structures[14] that remains elevated throughout the sample period, suggesting that researchers are leveraging AlphaFold 2's capabilities to address proteins with direct biomedical significance. However, this pursuit of novelty appears to involve a trade-off, as these structures also report consistently worse (higher) resolution values, suggesting a potential cost in experimental precision.

The normalised distributions in Figure 17 reinforce this focus on novelty, showing a distinct leftward skew for the AlphaFold 2 group across all three structural similarity metrics when compared to the counterfactual and baseline groups. It is important to recall that these similarity scores are always calculated against historically prior structures, thus capturing novelty at the time of publication. The distributions for AlphaFold 2 influenced work are distinctive: the TM-scores show a bimodal pattern with a second, well-defined mode near a score of 0.2, suggesting AlphaFold 2 is enabling the discovery of a specific class of highly novel folds. Similarly, the distribution for fractional identity has a notably long tail at low identity values, while the RCSB shape scores are heavily concentrated near zero.

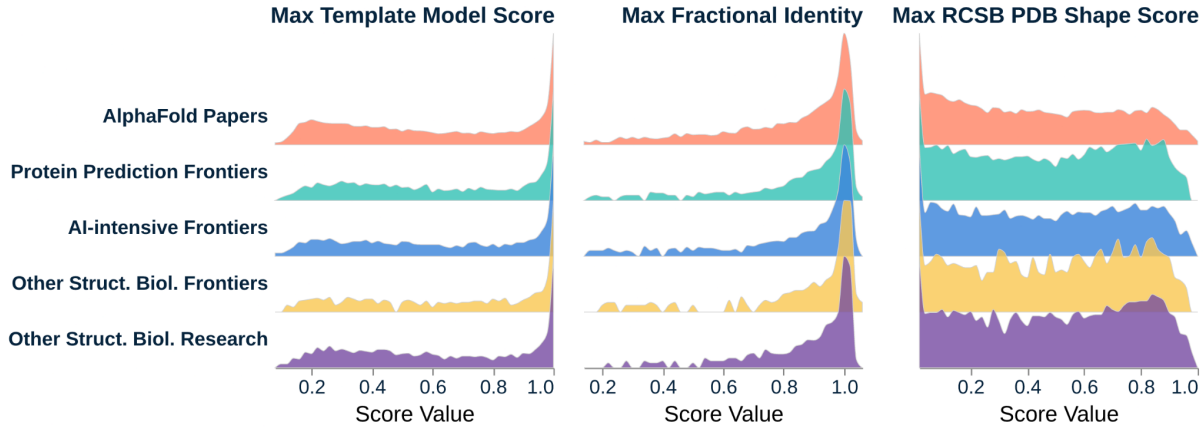

Figure 17. Distribution of the max TM-Score, fractional identity, and RCSB shape in AlphaFold 2- and frontier-linked publications. The density has been normalised within each source and subsequently square-root transformed, reducing the visual dominance of extreme peaks.

Taken together, these findings paint a consistent picture of a qualitative shift in research focus associated with building on AlphaFold 2. We observe a simultaneous trend towards structurally unique folds, evolutionarily distant sequences, diverse organisms, and disease-relevant targets. This pattern suggests an expansion into more challenging

[13] Our measure of organism rarity is based on the inverse frequency of a protein's source organism within our PDB submission dataset, using standardised taxonomic classifications from the NCBI Taxonomy database provided via UniProtKB. We acknowledge this metric does not account for all levels of taxonomic granularity (ie. specific strains) but serves as a consistent proxy for diversity.
[14] Disease relevance is determined using data from UniProt's "Involvement in disease" section

research areas, rather than a simple acceleration of existing workflows, and corresponds with a potential trade-off in experimental precision.

Regression analysis

Comparison with structural biology baseline

Our regression analysis reaffirms and quantifies the differences in the structural characteristics of proteins submitted by those works and academics building on AlphaFold 2 compared to the broader structural biology community.

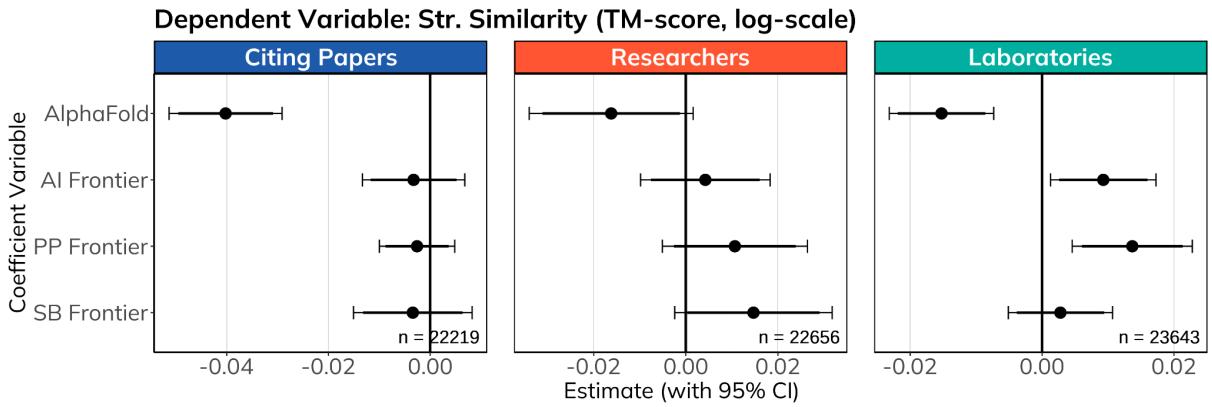

Figure 18. Coefficient estimates from linear regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

The analysis of the logarithmic TM-score (Figure 18) confirms the patterns observed in the descriptive data. Papers associated with AlphaFold 2 have mean TM-scores that are, on average, 4% lower than the baseline and counterfactuals, indicating higher structural novelty. A similar, though less statistically significant, novelty association is found for individual researchers, while at the laboratory level, building on AlphaFold 2 corresponds to a 1.7% lower average TM-score. This correlation is primarily concentrated among adjacent users, with downstream users showing no significant benefit.

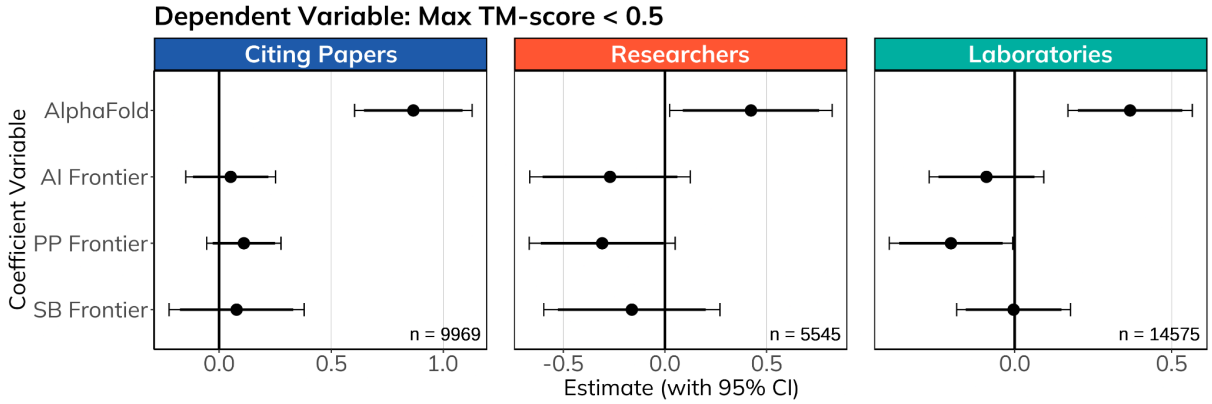

Figure 19. Log-odds ratios from logistic regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

To explore the upper bounds of novelty, we next consider the likelihood of discovering a truly distinct structure using a logistic regression (Figure 19). For this, we model the probability of a submission having a maximum TM-score below 0.5, a commonly accepted threshold indicating that two proteins possess different folds (Xu and Zhang, 2010).[15] Building on AlphaFold 2 is associated with a large and statistically significant 52.7% increase in the odds of producing a novel fold at the researcher level. At the laboratory level adoption corresponds to a 44.5% increase in the odds. This helps contextualise the small magnitude of the mean TM-score difference; while most structures from all groups are non-novel, building on AlphaFold 2 is strongly associated with an increased likelihood of producing a structure from the highly novel mode of the distribution seen previously.

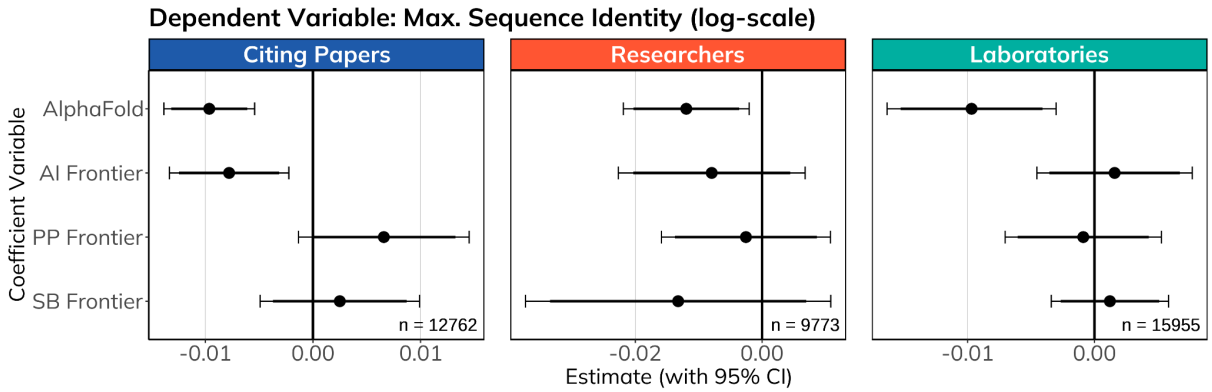

Figure 20. Coefficient estimates from linear regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

Our analysis of sequence-level novelty, which uses Foldseek-calculated fractional sequence identities, shows a small yet statistically significant negative association of approximately 1% for AlphaFold 2 across papers, individual researchers, and laboratories. This indicates that building on AlphaFold 2 is consistently linked to the submission of proteins that are more evolutionarily distant from previously characterised structures. While other AI frontier developments display a comparable correlation at the paper level, this association becomes insignificant for authors and labs.

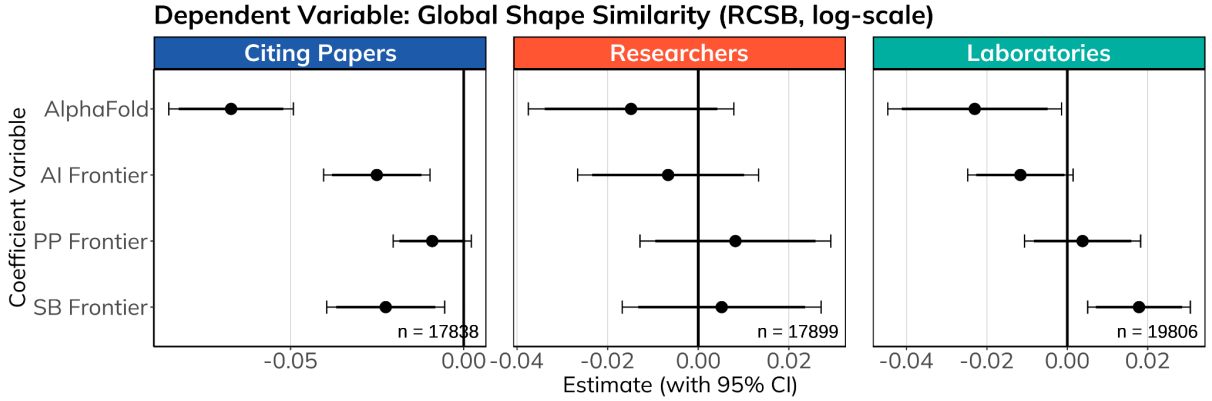

Figure 21. Coefficient estimates from linear regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

[15] Our sample size decreases as some labs and authors never have TM-scores higher than 0.5.

We also consider novelty through the lens of global topology using the alignment-free RCSB Shape Score (Figure 21). Although not our preferred metric for novelty, as alignment-free comparisons can overlook the finer details of fold-level and evolutionary relationships, the results provide further evidence consistent with our main findings. Building on AlphaFold 2 is associated with a significant 6.9% decrease in shape similarity scores at the paper level. For authors and labs, the association is considerably smaller, and only significant at the 95% level for the latter group. AI frontier methods also show a slight negative association for laboratories but have muted correlations among researchers, while both non-AI frontier groups have insignificant associations.

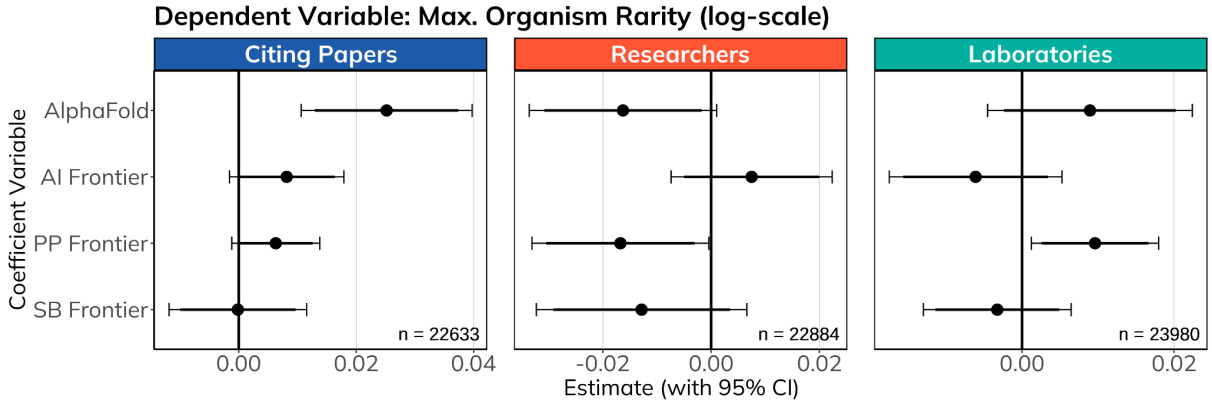

Figure 22. Coefficient estimates from linear regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

When examining organism rarity, the results are modest and vary by the unit of analysis. At the paper level, AlphaFold 2 is associated with a small but significant 2.6% increase in the rarity of source organisms, while other AI developments have a more muted association and other structural biology frontiers show a negative association. This pattern disappears at the researcher level, where a weakly significant negative association between Alphafold2 and organism rarity is observed. At the laboratory level, both Alphafold2 and non-AI methods are linked to small positive increases in organism rarity, but the result is non-significant for the former group.

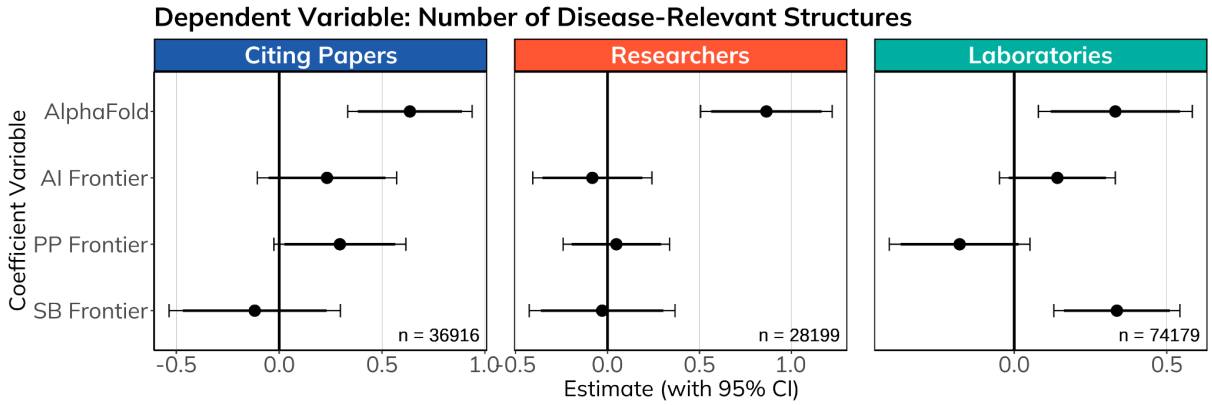

Figure 23. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

Our analysis of disease-relevant structures reveals large positive associations for AlphaFold 2 across all units of observation. The number of structures with a disease link more than doubles for individual researchers building on AlphaFold 2, while at the laboratory level, the increase is 39.2%. This strong association with biomedically relevant research is not consistently observed for other frontier developments.

These findings suggest that building on AlphaFold 2 is associated with a qualitative shift towards more novel and biomedically relevant research, although the specific dynamics differ between individual researchers and laboratories. This complex pattern could be interpreted in several ways:

- The consistent association with lower similarity scores across multiple metrics suggests that AlphaFold 2 is linked to an exploration of higher-risk, less characterised areas of protein space. This pursuit may come with a trade-off, as reflected by the association with lower-resolution experimental structures.
- The varied results between individual researchers and laboratories may reflect the complex nature of team-based science. The pooling of resources and expertise within a lab could enable the pursuit of ambitious projects, while increases in experimental productivity may be freeing up resources to target novel areas.

Complementarities with methodological adoption

To understand how methodological use of AlphaFold 2 shapes its scientific impact, we briefly consider the different approaches in which AlphaFold 2 and other methods can be used. Notably, data availability is a problem as publications relevant for this section appear underrepresented in our sample of intent-enriched papers, researchers, and labs.

Where evidence is available, the association between AlphaFold 2 and higher structural novelty is most pronounced for researchers that use it as a core methodological tool. Among users, methodological applications are linked to a statistically significant 4.2% reduction in TM-scores, indicating the submission of more novel structures. Similarly, researchers using AlphaFold 2 methodologically see a 1.5% decrease in sequence identity scores of submitted protein structures, suggesting that methodological adoption encourages exploration of proteins with greater evolutionary distance from known structures. At the paper and laboratory levels, these specific associations are not statistically significant.

Overall, our findings suggest that building on AlphaFold 2 is associated with a qualitative shift in research priorities towards more novel and disease-relevant structures, particularly when used as a methodological tool. This pattern, which corresponds with a trade-off in experimental precision, points to an expansion of research frontiers rather than a simple acceleration of existing work.[16]

[16] It is worth noting that the impact of AlphaFold 2 on disease-relevant submissions estimated at the level of individual publications, researchers and laboratories is being realised across a

5. Findings: Academic output

- Building on AlphaFold 2 shows varied correlations with publication volume: a small but statistically significant 2.5% increase for individual researchers and a weakly significant 5.1% increase for laboratories.
- Methodological use of AlphaFold 2 is linked to a significant 11.5% increase in lab-level publication volume, an association not observed for other AI frontier tools and driven primarily by downstream, applied research.
- Building on AlphaFold 2 is associated with an increase in citation counts of 8.1% for individual researchers and 10.4% for laboratories, with the impact being strong for both methodological and background use.
- The association with field-weighted citation impact (FWCI) is also consistently positive, and notably stronger for researchers building on AlphaFold 2 than for those using any other frontier technique.
- The dynamics of building on AlphaFold 2 appear to be split: modest increases in publication productivity are primarily observed at the laboratory level, while researchers see the largest and most consistent increases in citation impact.

AlphaFold 2's ability to predict protein structures with high accuracy may reduce the time and effort required for key stages of structural biology research, such as modeling and speeding up the interpretation of experimental results via molecular replacement[17].

This efficiency could enable researchers to pursue more ambitious projects or increase the overall throughput of their work, potentially boosting publication volume. Moreover, AlphaFold 2's availability as an open source model, implementations such as Colabfold and the AlphaFold 2 structure database helps facilitate access to structural data (Varadi et al. 2022; Mirdita et al. 2022). These resources not only benefit applied labs, but also enhance the capabilities of researchers engaged in basic research, fostering new opportunities to test hypotheses and generate experimental insights.

AlphaFold 2's broad accessibility sets it apart from other high-impact AI developments in structural biology, which, as seen in Figure 7, tend to receive the majority of their recognition within computer science or AI-focused venues. Non-AI frontier approaches, on the other hand, often focus on established techniques that see use in medical or applied

substantially larger user base than other frontier techniques, potentially contributing to greater aggregate impacts on biomedical research.

[17] Molecular replacement is "a simple trial-and-error method of solving crystal structures when a suitable related model is available." (Evans and McCoy 2008)

contexts, but may have limited scope for increasing publication output. In contrast, AlphaFold 2 has been adopted by both computationally focused groups and researchers more centered on biochemistry and molecular biology, reflecting its broader applicability.

The wide adoption and applicability of AlphaFold 2 across diverse research groups provide an opportunity to examine its influence on publication activity. To assess this impact, we analyse publication volumes and citation counts, comparing researchers who have adopted AlphaFold 2 with those using other frontier developments.

5.1 Academic throughput

Regression analysis

As in the previous section, we use our regression framework to examine the relationship between building on AlphaFold 2 and academic output while accounting for potential sources of bias. When examining publication volumes, however, we exert particular caution. As noted in earlier sections and in our Appendix, this is the one outcome where our analysis suggested potential, albeit only moderate, violations of the parallel trends assumption, limiting our ability to interpret these results.

Comparison with structural biology baseline

With the above caveat in mind, our regression analysis reveals differing patterns between the researcher and laboratory levels. For individual researchers, we find a small but statistically significant association between building on AlphaFold 2 and a 2.5% increase in publication output. At the laboratory level, the association is larger, corresponding to a weakly significant 5.1% increase in publication volume.

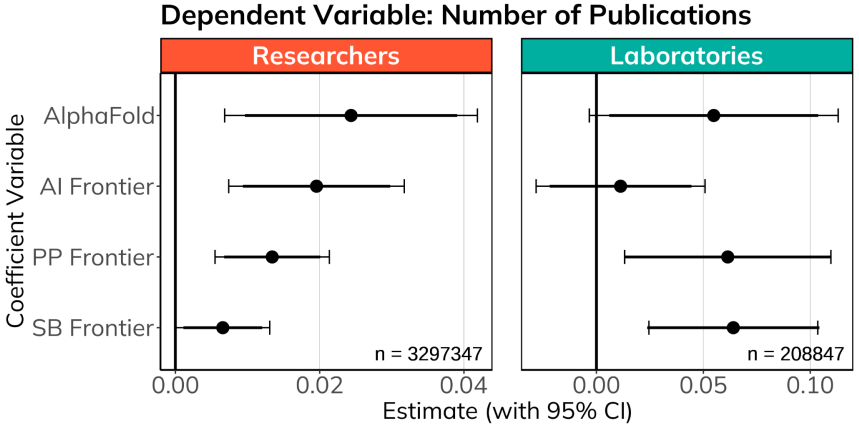

Figure 24. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

This larger association at the laboratory level, compared to the individual level, can be interpreted in two non-mutually exclusive ways. It could reflect a selection effect, where more productive laboratories are more likely to adopt new tools. Alternatively, it may suggest that the collective capabilities of a laboratory, such as diverse expertise and

collaborative workflows, are important for translating the use of such tools into higher academic throughput. Our analysis cannot definitively distinguish between these two potential mechanisms.

Comparison with other frontier techniques

When placing these findings in the context of other frontier techniques, a clearer picture emerges. At the individual researcher level, the association for AlphaFold 2 influence is comparable to that of other frontier groups; while the point estimates for other methods are smaller, the differences are not statistically significant.

At the laboratory level, however, a distinction becomes apparent. The modest increase in publication volume associated with building on AlphaFold 2 is similar in size to that linked to both non-AI protein prediction and other structural biology frontiers. This contrasts with adopters of other AI-intensive frontiers, for whom we find no significant association with publication output. The observation that the association for AlphaFold 2 aligns with non-AI methods, but differs from other AI tools, may indicate that AlphaFold 2 is integrated into experimental workflows in a manner more akin to a traditional, broadly applicable scientific method.

Complementarities with methodological adoption

We next consider how the association with publication volume differs when a technology is used as a core method versus cited as background knowledge. Since this analysis relies on the subset of our data with available citation intent information, the sample size is substantially reduced, and the results should be interpreted with this in mind.

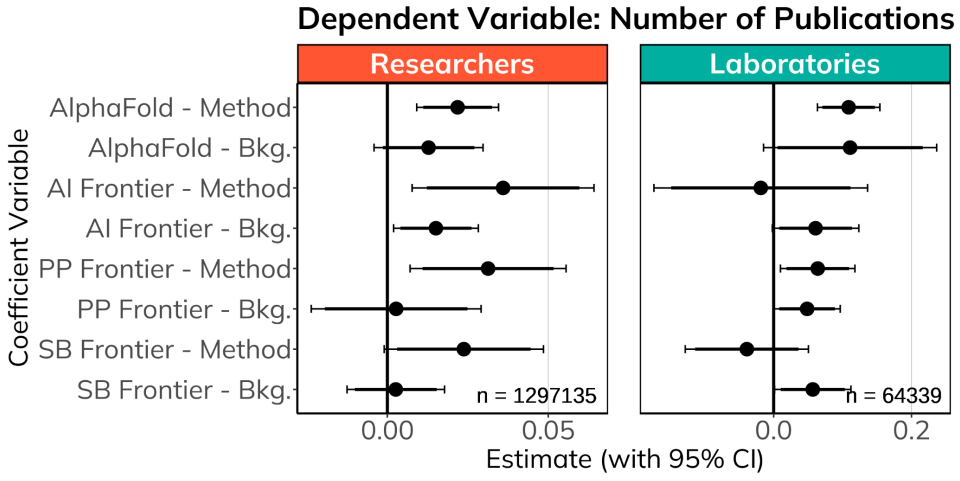

Figure 25. Coefficient estimates from Poisson regressions showing links for structural biologists. Bars represent 95% confidence intervals.

At the individual researcher level, this subgroup analysis helps to clarify the small association observed in the full sample. Figure 25 shows that the positive link with publication output is driven specifically by the methodological use of both AlphaFold 2, a trend which holds broadly for other frontier developments. In contrast, a weaker

relationship is generally found for the background use of these tools and methods, although this effect is more consistent for researchers.

At the laboratory level, the distinction for AlphaFold 2 adoption is clearer. Methodological use is associated with a significant 11.5% increase in publications, whereas the association for background use is not statistically significant. This association is also larger than the modest increases linked to the non-AI frontiers. Furthermore, this link is primarily associated with downstream applications; additional specifications show a strong association for applied labs, while the link for adjacent labs is more muted. This dynamic differs from the non-AI frontiers, which see positive associations exclusively at the adjacent level.

These patterns suggest that the nature of engagement with a technology is a key factor. The significant increase in publication output is primarily linked to the methodological use of AlphaFold 2, an association which, while present for individual researchers, becomes more pronounced at the laboratory level. This distinguishes it from other AI frontiers, where a similar link is not observed for labs, and from non-AI methods, where the associated increases are smaller. This may indicate that while individual researchers can see benefits, the full translation of methodological adoption into higher publication rates is more evident within the collaborative structure of a laboratory.

5.2 Citation impact

Descriptive analysis

Beyond its scale, the influence of AlphaFold 2 is also reflected in the quality, influence and visibility of associated publications. To assess this dimension, we consider both raw citation counts and field-weighted citation impact (FWCI), a standardised metric that accounts for differences in publication type, subfield, and publication year to provide a normalised measure of relative impact.

An initial comparison suggests that AlphaFold 2 research seems to accrue more citations, with publications in AlphaFold 2 citation chains 2.47 times more likely to rank among the top 10% of year- and field-normalised citations compared to the average of all other research groups in our sample. In particular, AlphaFold 2 publications are 1.89 times more likely to be in the top 10% compared to other AI-related publications, 2.37 times more likely compared to non-AI publications, and 3.39 times more likely compared to other publications. In absolute terms, 39.11% of AlphaFold 2 publications rank in the top 10% compared to 20.74% for other AI publications, 16.50% for non-AI counterfactuals, and just 11.54% for other publications. These differences could be explained by systematic differences between researchers who use AlphaFold 2 and the baseline. Our regression analysis tries to adjust for these differences.

Regression analysis

Comparison with structural biology baseline

Our regression analysis reveals positive associations between building on AlphaFold 2 and citation impact across multiple units of analysis. For raw citation counts (Figure 26), the association is strongest among papers directly citing AlphaFold 2, which see a 28.9% average increase in citations. Publications from Alphafold-using individual researchers see a 8.1% increase, while the association for laboratories is 10.4%.

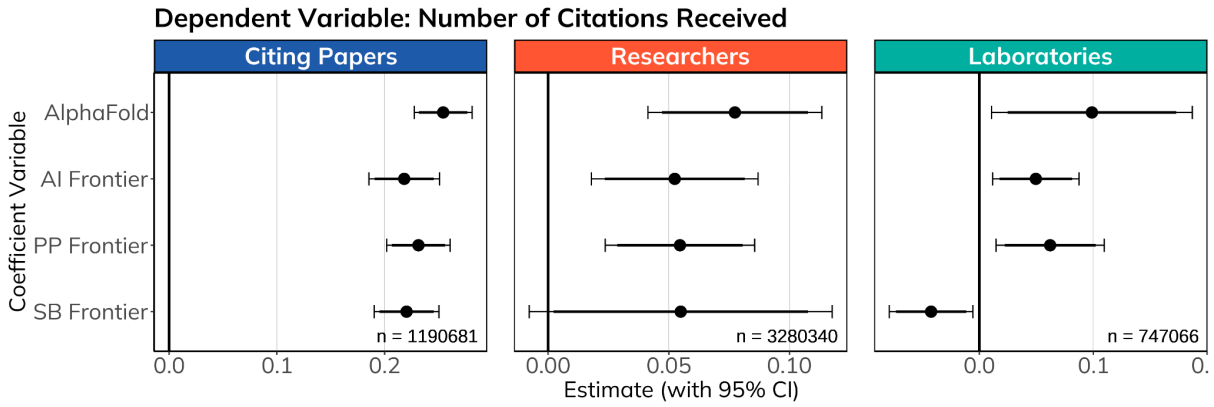

Figure 26. Coefficient estimates from Poisson regressions showing links for highly-experienced biologists. Bars represent 95% confidence intervals.

The patterns are similar, though more modest, for field-weighted citation impact (FWCI), which accounts for field- and year-specific citation dynamics (Figure 27). This suggests the larger associations with raw citations are partially driven by the increased prominence of AlphaFold 2-related fields. A positive association remains across the board: we observe a 5.2% increase in FWCI for researchers, 2.7% for laboratories, and 3.9% for linked papers.

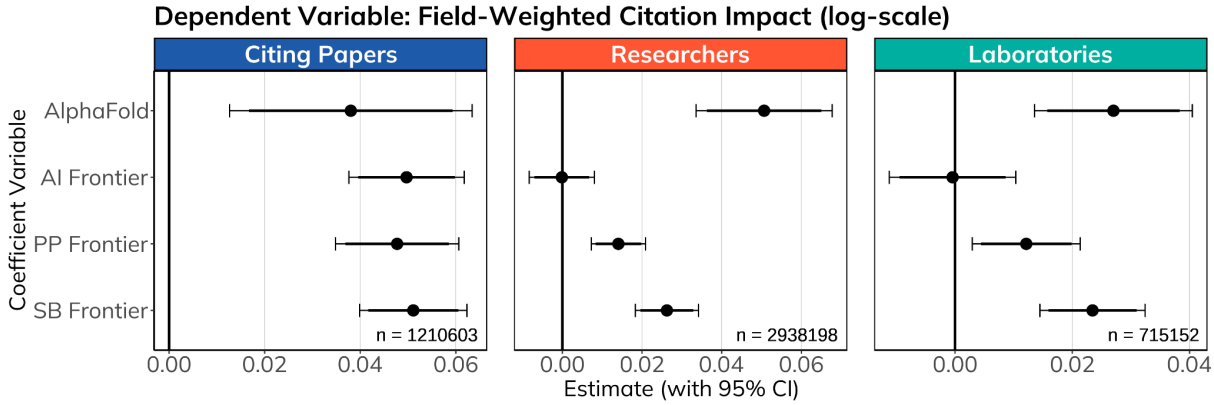

Figure 27. Coefficient estimates from log-linear regressions showing links with key developments. Error bars represent 95% confidence intervals.

The consistency of these positive findings across different citation metrics points to a robust association between building on AlphaFold 2 and increased research impact. This relationship allows for two interpretations: it may be that AlphaFold 2 directly enhances

the quality and visibility of research, leading to higher impact, or that leading researchers and labs already on a trajectory towards high impact are more likely to adopt cutting-edge tools. While our analysis controls for observable characteristics and the lack of strong pre-treatment trends, it cannot fully disentangle potential selection and treatment links,

Comparison with other frontier techniques

When comparing AlphaFold 2 with other frontier developments, we find distinct patterns for raw citation counts versus field-weighted impact. For raw citation counts, we first note that at the paper level, the associations are nearly identical across all frontier groups. This is expected by the design of our counterfactual selection, which relies on citation patterns, and provides reassurance that we selected publications of comparable baseline impact.

At the researcher and laboratory levels, some differences in the magnitude of the point estimates emerge. For individual researchers, whilst all frontier groups are associated with a positive increase in citations, the association for AlphaFold 2 (8.1%) is larger in magnitude than for the counterfactuals (5.1-5.2%). A similar pattern is observed at the laboratory level, where the association for AlphaFold 2 is approximately double the size of that for other frontier methods. However, at both levels , the confidence intervals for these estimates overlap, meaning the observed differences are not statistically significant.

The analysis of field-weighted citation impact (FWCI) reveals a different pattern of associations. At the researcher level, building on AlphaFold 2 is linked to a positive and significant association with FWCI. In contrast, other AI methods show no statistically significant association. While non-AI protein prediction and other structural biology frontier methods also show positive associations, their magnitude is smaller than that observed for AlphaFold 2. At the laboratory level, the association for AlphaFold 2 is comparable in size to that for laboratories building on other structural biology methods. For laboratories using other AI-intensive methods, we again find no significant association with FWCI.

Taken together, these results suggest that while adopting any high-impact frontier technology is associated with positive citation outcomes, AlphaFold 2's impact profile is distinct. Its link with citation impact for researchers is potentially stronger than other methods, and among labs, it shows a robust link that other contemporary AI tools do not.

Complementarities with methodological adoption

To understand how the nature of engagement with new technologies relates to citation impact, we explore complementarities with methodological adoption. As this analysis relies on the subset of our data with available citation intent information, the sample size is substantially reduced, and the results should be interpreted with this caveat in mind.

At both the paper and individual researcher levels, we find a pattern where the association with citation impact is larger for the background use of AlphaFold 2 than for its methodological use. At both the paper and individual researcher levels, whilst both usage

types are linked to a significant increase in citations, the association is consistently stronger when AlphaFold 2 is cited as background context. This may indicate that background use is more common in broader, interdisciplinary studies that appeal to a wider readership.

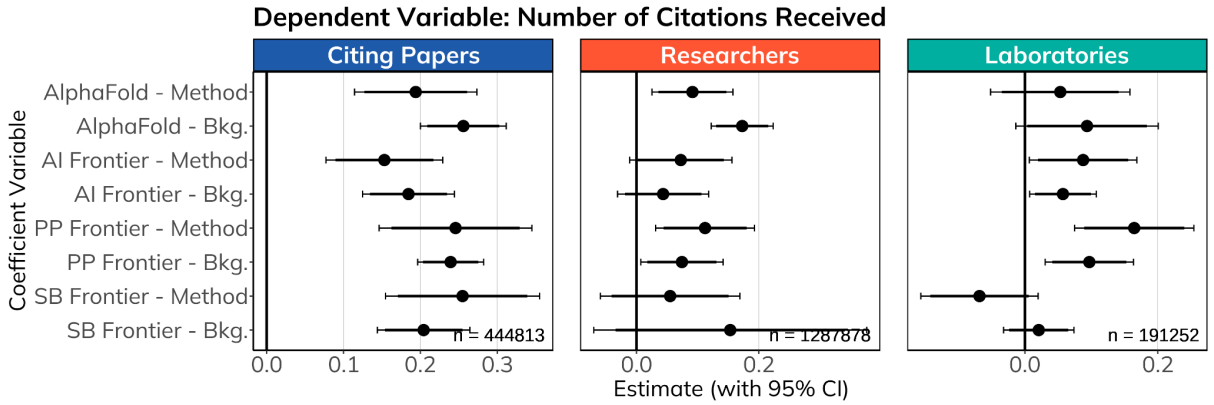

Figure 28. Coefficient estimates from Poisson regressions showing links for highly-experienced biologists. Bars represent 95% confidence intervals.

For the field-weighted citation impact (FWCI) results (Figure 29), the distinction between methodological and background use is less pronounced. At both the researcher and laboratory level, AlphaFold 2 is linked to a strong positive association with citation impact. This contrasts with other AI methods, for which the association is muted, and with non-AI protein prediction and other structural biology frontiers, which show weakly positive associations.

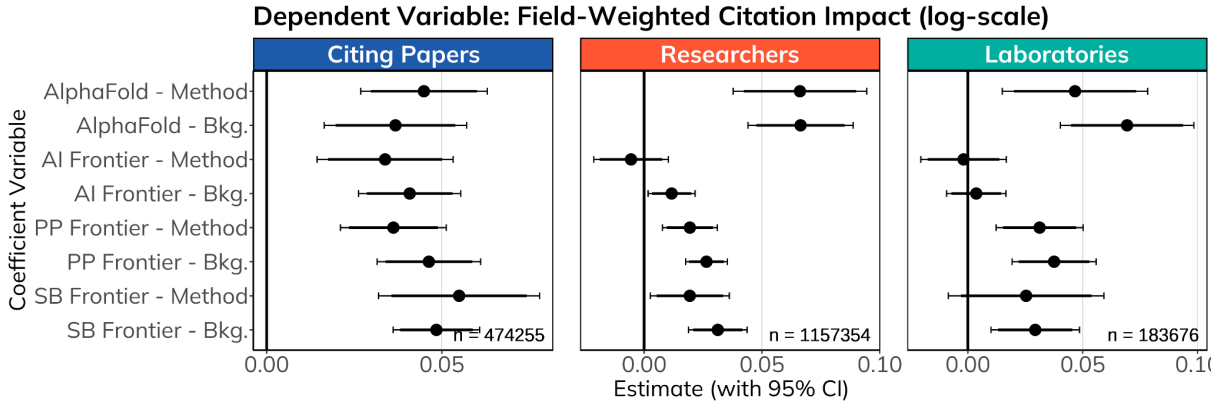

Figure 29. Coefficient estimates from log-linear regressions showing links with key developments. Error bars represent 95% confidence intervals.

All in all, results in this section point to the fact that all these methods have a clear positive association with citations, as one would expect through the design of our study. Laboratories seem less reactive to these, possibly because of their associated team dynamics or the fact that these are by and large specialised on benchside research (as seen in the previous section). Conversely, researchers have strong positive associations, dominated by AlphaFold 2 over other frontier developments when normalising by field.

6. Findings: Applied research and innovation

- AlphaFold 2 influence shows a dual correlation with disease-relevant research: while individual papers have a comparable disease focus to the baseline, researchers (9.3%) and laboratories (5.0%) that adopt it show a significantly higher probability of producing disease-relevant work.
- At the paper level, AlphaFold 2 is associated with a doubling in the probability of receiving a clinical citation, an effect size that is also double that of other frontier AI and protein-prediction methods. However, no significant association is found for individual researchers or laboratories.
- Adoption is linked to a substantial increase in patent citations across all units: 36.8% for papers, 34.2% for laboratories, and 22.6% for researchers, suggesting a broad association with work of commercial relevance.
- A positive association with patent quality (measured by subsequent patent citations) is found for AlphaFold 2-related papers, a finding that is distinctive when compared to the largely non-significant results for other frontier methods. No association is detected at the researcher or laboratory level.

AlphaFold 2's structural predictions have the potential to support translational research by improving understanding of disease-related proteins and informing drug discovery efforts. Accurate protein structure models can aid in identifying drug targets and clarifying protein-disease relationships. While these advancements hold promise for clinical and applied research, their integration into translational applications remains in its early stages. In commercial contexts, AlphaFold 2's open-access database provides a resource for early-stage innovation, but translating structural insights into patentable products requires further development.

This section examines AlphaFold 2's role in translational research by assessing its associations with disease-focused academic research outputs (measured via MeSH terms), and the associations between those outputs and clinical article submissions, and related clinical trials. We also explore its impact on patenting activity, including citations to AlphaFold 2-linked research and the degree to which these patents receive further citations, a proxy for commercial relevance.

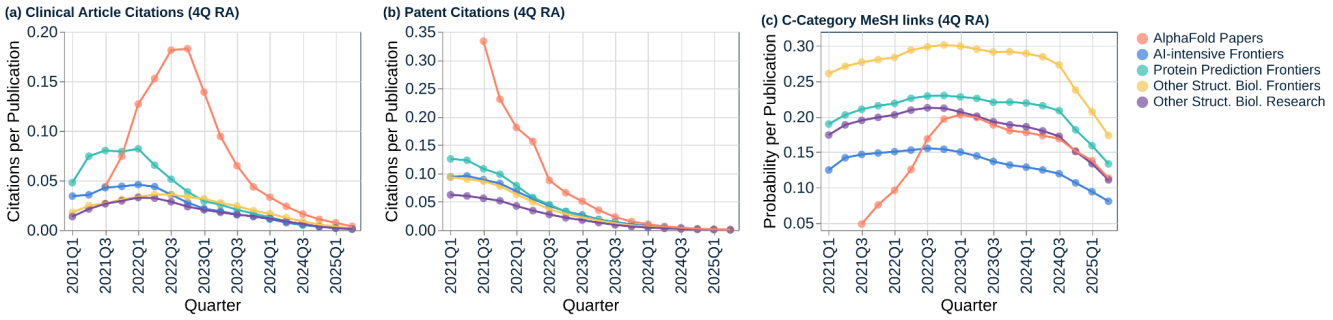

Figure 30. Rolling averages of clinical and translational outputs for established researchers building on frontier methods. These plots show the per-researcher mean number of clinical and patent citations, as well as the prevalence of disease-related MeSH terms in their research. Trends help track relative exposure to translational applications before and after adoption.

Figure 30 presents overall trends using a 4-quarter rolling average of clinical article citations, patent citations, and disease-relevant papers (proxied by C-category MeSH keywords) associated with researchers.

Non-AI protein prediction frontier research is associated with the highest number of clinical article citations and disease-related outputs, which aligns with the descriptive results. While these non-AI methods have traditionally represented the most clinically-oriented innovations, AlphaFold 2-related work has seen a marked increase in this area, receiving more clinical citations than other AI methods. The cited works are often not the original foundational papers but rather downstream applications from late 2022 onwards, suggesting a dynamic where applied work is driving clinical interest.

A distinct pattern is observed for patent citations, where publications related to AlphaFold 2 have received the majority of citations among the frontier groups since mid-2021. In contrast to the pattern for clinical citations, it is the earliest, foundational AlphaFold 2 publications that are most frequently cited in patent filings. While this partly reflects natural citation lags, it also suggests that the core innovation is viewed as directly applicable for intellectual property purposes without necessarily requiring further downstream development.

Finally, when examining the production of disease-relevant papers, the output for AlphaFold 2 is lower than that of the non-AI driven frontiers, which is consistent with their differing research focuses. However, when compared to its AI-based peers, the output for AlphaFold 2 is higher and has been since mid-2022. This suggests a stronger link between this particular AI technique and research on human diseases compared to other contemporary AI innovations.

It is important to note that the observed decline in clinical article and patent citations towards the end of the time series likely reflects citation lag rather than diminishing impact. More recent publications have not yet had sufficient time to accumulate downstream citations. This temporal constraint does not affect our MeSH term analysis, which directly examines disease-related content within the research papers themselves.

6.1 Disease research

Descriptive analysis

Following the introduction of AlphaFold 2's potential role in translational research, we begin by assessing the prevalence of disease-focused research across different methodological approaches. We utilise Medical Subject Headings (MeSH) C-class keywords[18], which form a dedicated branch of the hierarchical taxonomy specifically for disease-related terms. This classification provides a structured way to identify research outputs with direct relevance to human health.

Among papers directly citing AlphaFold 2 (adjacent links), 9.4% contain disease-related MeSH terms, compared to 10.6% for other AI frontier developments, 28.6% for other protein prediction approaches, and 22.2% for other structural biology innovations. These dynamics persist on papers with downstream connections, where AlphaFold 2-linked papers show a disease-related prevalence of 14.8%, compared to 12.2% for other AI methods, and 24.2% for both protein prediction and structural biology frontiers.

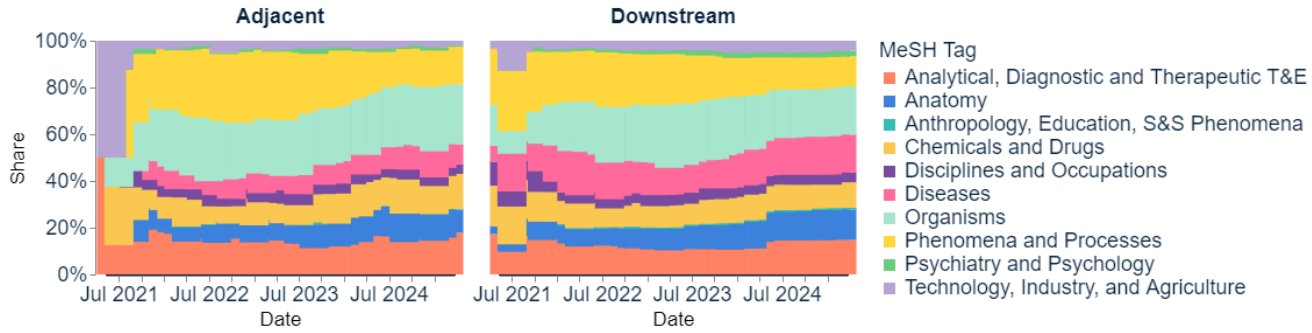

Figure 31. Monthly distribution of AlphaFold 2-related publications by MeSH category (July 2021 - March 2025). The plots are faceted to show trends for adjacent papers, which directly cite core AlphaFold 2 research, and downstream papers, which are at least one degree of separation further.

When focusing on papers with methodological citations, the pattern becomes slightly more pronounced. Among papers directly citing AlphaFold 2, 8.5% demonstrate strong disease relevance, compared to 8.3% for other AI frontier developments and 16.1% for non-AI protein prediction approaches. For papers with downstream strong connections, AlphaFold 2-linked research shows a strong disease relevance rate of 11.6%, compared to 10.2% for other AI methods, and 21.1% for non-AI protein prediction approaches.

These statistics suggest that while AlphaFold 2-linked research demonstrates greater disease relevance than other AI frontier developments, both categories show substantially

[18] Disclaimer: MeSH-C levels are defined in different ways across the datasets, which distorts the association sizes, but not the sign / significance. In particular, papers capture the individual use of MeSH-relevant tags. Researchers and laboratories instead reflect the cumulated proportion of papers authors or teams produce with the key MeSH-C levels, which suffers from increasing large denominators driving most values to zero. This does not affect the sign and significance of results.

lower disease focus than non-AI methodological approaches. In absolute terms, we identify 3,097 papers directly citing AlphaFold 2 that contain disease-related MeSH terms, with an additional 76,320 papers in the broader citation network.

Regression analysis

To account for potential confounding factors, we again employ our regression framework to study the relationship between AlphaFold 2 use and disease-related research outputs.

Comparison with structural biology baseline

At the paper level, Figure 32 shows that AlphaFold 2-linked research has comparable prevalence of disease-relevant content compared to baseline structural biology papers. However, a different picture emerges at the user level. For individual researchers, we observe a significant positive association, with AlphaFold 2 influence corresponding to a 9.3% higher probability of producing disease-relevant research. A similar, though smaller, positive association of 5.0% is observed for laboratories.

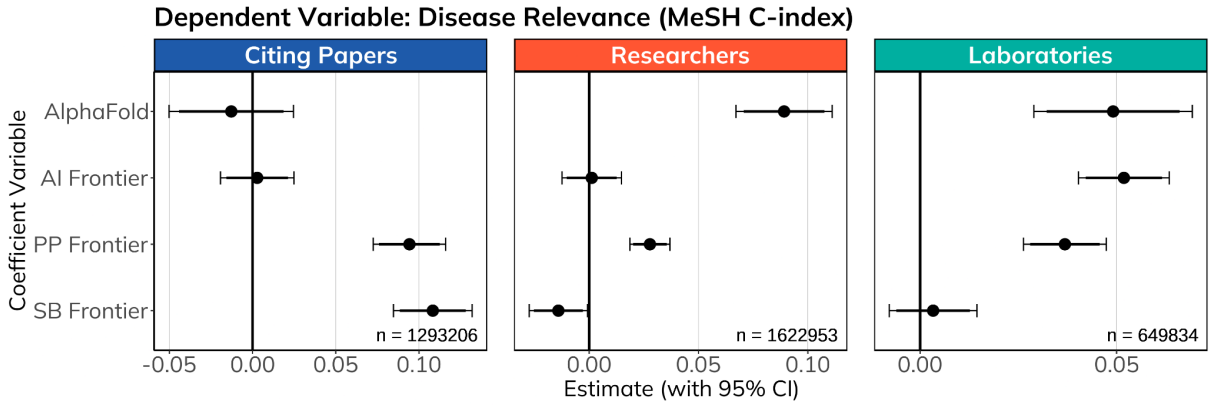

Figure 32. Coefficient estimates from log-linear regressions showing links with key developments. Error bars represent 95% confidence intervals.

Comparison with other frontier techniques

At the paper level, publications linked to non-AI methods have a higher probability of being disease-relevant, whilst AI-based methods, including AlphaFold 2, show no significant association.

This pattern shifts at the researcher level, where AlphaFold 2 adoption is associated with a substantially higher probability of producing disease-relevant work than any of the other frontier groups, which show much smaller or non-significant associations.

At the laboratory level, the positive association for AlphaFold 2 is comparable in size to that for other AI-adopting labs. The link is slightly smaller for non-AI protein prediction labs and non-significant for other structural biology frontiers.

6.2 Clinical research

Descriptive analysis

To assess the translational impact of AlphaFold 2 and other frontier developments, we examine clinical article citations using PubMed's iCite data[19], which captures the extent to which research is cited in clinical literature. The three core papers we use to define our AlphaFold 2 sample have four unique clinical article citations, listed in Figure 33.

Our broader descriptive analysis (Table 1) reveals a distinction between AI-based and non-AI approaches. Publications linked to frontier structural biology garner the highest rate of clinical citations by a significant margin among adjacent research. In contrast, both AlphaFold 2 and other AI frontiers see substantially increased baseline rates in downstream research, above (AlphaFold 2) and on par with (other AI) that of frontier protein prediction and other structural biology.

	Adjacent		Downstream	
	All	**Method**	**All**	**Method**
AlphaFold 2	0.009	0.015	0.025	0.059
AI Frontiers	0.005	0.002	0.014	0.023
Protein Prediction	0.009	0.003	0.021	0.039
Other Struct. Biology	0.021	0.021	0.016	0.021

Table 1. Clinical article citations per paper by frontier group and citation sequence distance.

However, this overall pattern masks important dynamics based on usage type and proximity to the core research. When we consider only papers with strong methodological links, the clinical citation rate for AlphaFold 2-linked work improves and surpasses that of other AI innovations. Furthermore, the relevance of AlphaFold 2 to research that has impact in clinical works appears larger for downstream works, with citation rates for methodologically-linked downstream papers being nearly four times higher than for adjacent ones.

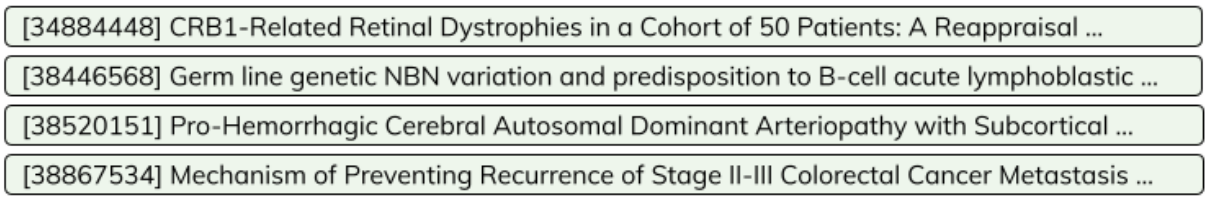

Figure 33. Clinical articles citing 'Highly accurate protein structure prediction with AlphaFold'

[19] Accessed on June 12, 2025.

While AI-based approaches currently show lower overall clinical citation rates than established non-AI methods, the pattern for AlphaFold 2 is noteworthy. Its performance relative to other AI frontiers, particularly when used as a method in downstream research, may suggest its high accuracy is helping to bridge the gap between computational prediction and clinically relevant applications more effectively than previous AI tools.

In absolute terms, our wider dataset contains 287 clinical article citations (267 unique clinical articles) to adjacent AlphaFold 2 papers and 12,787 citations to downstream papers (5,758 unique articles). When focusing only on chains with methodological links, we identify 69 clinical article links for adjacent papers and 4,244 for downstream ones.

Regression analysis

Comparison with structural biology baseline

When adjusting for field differences, authors, and time fixed effects, our regression analysis reveals a strong positive association between AlphaFold 2 and clinical research impact, but only at the paper level. Figure 34 shows that publications in an AlphaFold 2 citation chain are associated with a doubling in the probability of receiving a clinical article citation.

Conversely, at the researcher and laboratory levels, we find no statistically significant association with clinical citations for those building on AlphaFold 2. This may reflect the composition of these groups in our data; it is possible that our method identifies researchers and laboratories with a primary focus on fundamental or experimental structural biology, whose work is less likely to generate direct clinical citations.

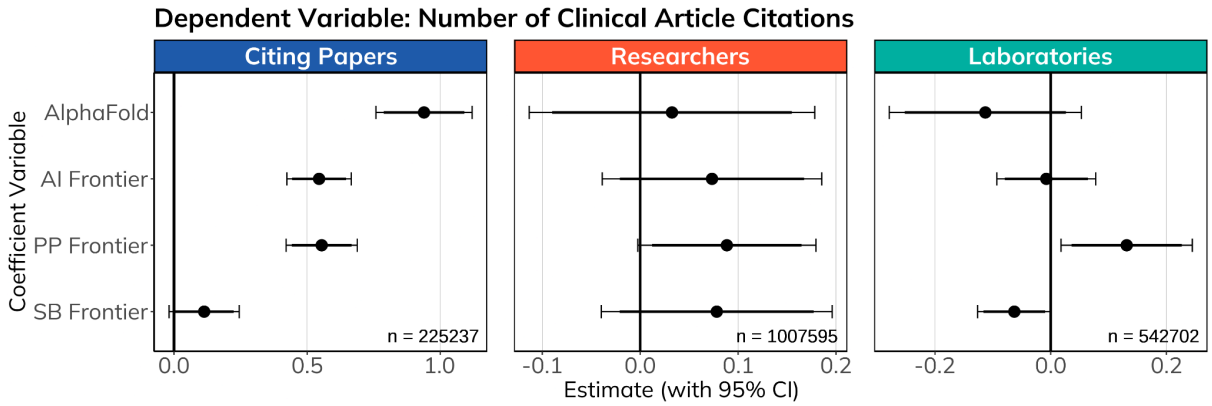

Figure 34. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

In line with the idea that applied research is more relevant for clinical trials, additional results available in the codebase show that, at the paper level, the positive association with clinical impact is driven entirely by downstream, applied research.

Comparison with other frontier techniques

When compared with other frontier developments, the strong association for AlphaFold 2 at the paper level is particularly notable. The probability of receiving a clinical citation is approximately double that associated with other AI and non-AI protein prediction methods. For other structural biology frontiers, the association is non-significant and close to zero.

At the researcher and laboratory levels, however, the pattern is different. Here, AlphaFold 2 influence shows no significant association, a finding that is consistent across most frontier groups. The sole exception is for non-AI protein prediction techniques, which are linked to a positive association with clinical citations.

These findings suggest a clear divergence. While AlphaFold 2 is linked to a uniquely strong association with clinical impact at the publication level, this does not translate to a significant increase for the individual researchers or laboratories in our sample. In contrast, the more established, translational focus of non-AI protein prediction methods appears to be reflected in a more consistent link to clinical citations at the user and team level.

6.3 Patent Citations

Descriptive analysis

Our analysis of patent citations, a proxy for commercial relevance, shows that on a per-publication basis, research linked to AlphaFold 2 receives fewer patent citations than work from the other frontier groups (Table 2). This result, however, should be interpreted with considerable caution. Patent citations have significant time lags, and the literature stemming from the more recent AlphaFold 2 innovation has had less time to accrue these citations compared to the more mature corpuses of the counterfactuals.

	Adjacent		Downstream	
	All	**Method**	**All**	**Method**
AlphaFold 2	0.024	0.028	0.007	0.012
AI Frontiers	0.143	0.191	0.026	0.025
Protein Prediction	0.117	0.042	0.034	0.076
Other Struct. Biology	0.075	0.319	0.028	0.037

Table 2. Patent citations per paper by frontier group and citation sequence distance.

While the citation rate is still maturing, a qualitative analysis of the types of patents citing AlphaFold 2-related work reveals a clear translational pathway (Figure 35). Patents that cite adjacent, foundational research are concentrated in core technology areas like machine learning and genetic engineering. In contrast, patents that cite downstream research are dominated by applied pharmaceutical and therapeutic categories, such as

anti-infectives and protein-based drugs. This seems to suggest a clear diffusion of impact from foundational methods to applied commercial products

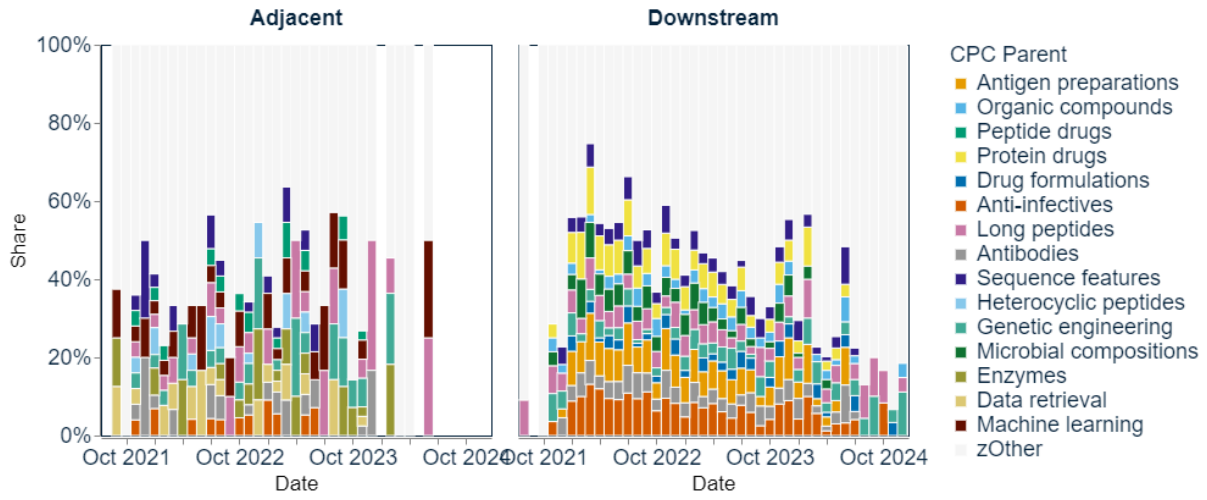

Figure 35. Composition of CPC parent codes over time for adjacent and downstream AlphaFold 2 publications between June 2021 and January 2025. Each stacked bar represents the proportional share of CPC parent codes by publication month. Only the ten most prevalent CPC codes per location are shown explicitly; all others are grouped under "zOther,"

In absolute terms, the scale of this impact is already notable. The three core papers are cited by 138 unique patents. Adjacent articles are linked to 607 patents, while downstream research is cited in 2,537 unique patents[20]. This suggests that AlphaFold 2 is acting as a dual-use innovation: it is a source of foundational methods itself, but also a tool that supports the early-stage research that may impact the development of downstream, commercially relevant applications in medicine and biotechnology.

Regression analysis

Comparison with structural biology baseline

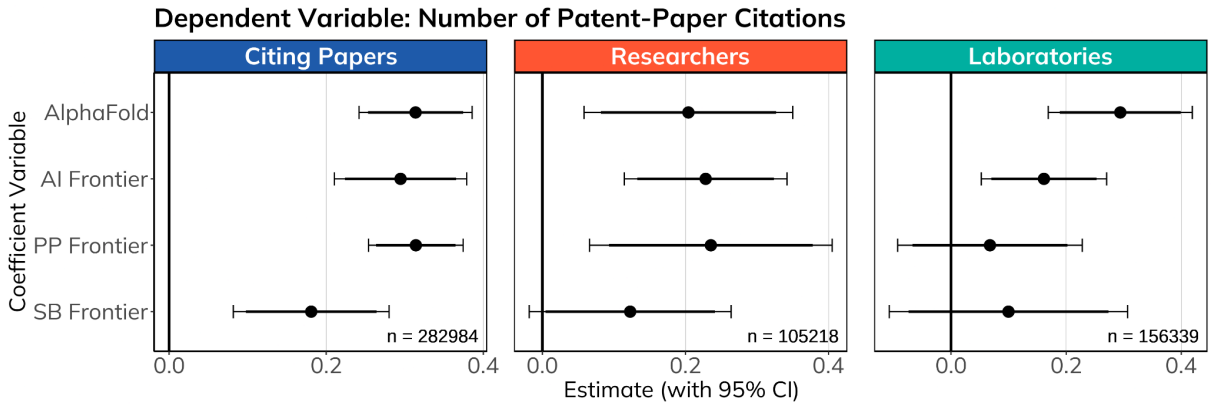

Figure 36. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

[20] Our sample is largely focused on the EU, US and WO patents, due to limitations with the data.

Our regression analysis reveals positive associations between building on AlphaFold 2 and patent citations across all units of analysis. Figure 37 shows that articles in an AlphaFold 2 citation chain are associated with a 36.8% increase in patent citations. Similarly, positive associations are found for laboratories (34.2%) and individual researchers (22.6%).

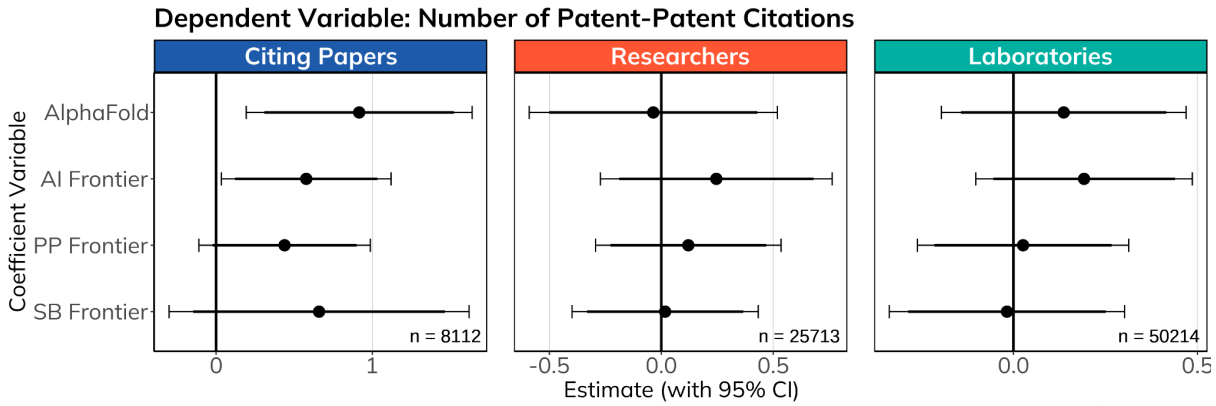

Figure 37. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

Figure 37 shows our proxy for patent quality: the number of citations patents linked to any papers in our data received from other patents published or submitted at a later date.[21] A positive and significant association is found only for the direct paper citation chains, with no significant effect detected at the researcher or laboratory level.

Comparison with frontier techniques

When comparing AlphaFold 2 with other frontier methods, its association with patent citations is often comparable to that of other innovative techniques. At the paper and researcher levels, the magnitude of the association for AlphaFold 2 is similar to that for other AI and non-AI protein prediction methods, and larger than the association for other structural biology frontiers.

At the laboratory level, however, a distinction appears: the association for AlphaFold 2 is notably larger than for other AI methods, while non-AI and other structural biology frontiers show no positive association.

Regarding patent-to-patent citations, the results for other frontier developments are largely non-significant, suggesting the link between AlphaFold 2-related papers and this particular metric is distinctive.

[21] The additional lag bias that this citation-on-citation dynamic introduces warrants caution when interpreting results. Time fixed effects should capture this, but near-present values remain unstable.

Complementarities with experience and methodological adoption

Figure 38 shows that, at the paper level, the nature of engagement with a technology appears to be a distinguishing factor for non-AI innovations. For both non-AI protein prediction and other structural biology frontiers, the positive association with patent citations is primarily driven by methodological use. In contrast, for AlphaFold 2 and other AI-based methods, no significant distinction between methodological and background use is observed. At the researcher level, the results are less conclusive.

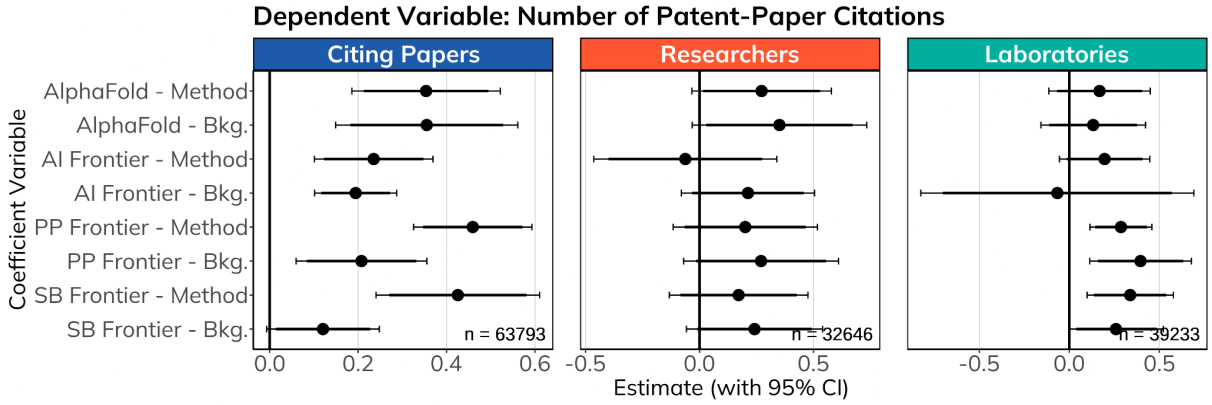

Figure 38. Coefficient estimates from Poisson regressions showing links with frontier developments. Error bars represent 95% confidence intervals.

7. Discussion

Our study has examined the impact of AlphaFold 2 on experimental structural biology, academic outputs and translational results, revealing its ability to accelerate research and enhance scientific impact across a diversity of fields. AlphaFold 2's impact over the time frame covered by this investigation is characterised by:

- A significant boost in experimental structure submissions and a diversification in the range of proteins explored. Here, AlphaFold 2 seems to be making a unique contribution by comparison to other frontier techniques which for example are linked to fewer new protein structure submissions and that are less novel structurally.
- A modest increase in the academic publishing productivity of those who build on it accompanied by significant growth in the citations that this research receives. The effects of counterfactual frontier techniques are similar, although realised over a much larger user population.
- An increase in clinical research and patent citations to laboratories that build on AlphaFold 2, showing impact between other AI techniques and more applied non-AI frontier techniques. AlphaFold 2 is not linked to an increase in disease-related research for researchers or laboratories, and papers citing AlphaFold are less likely to be disease-related.

This supports the notion of AI as a valuable research tool and tentatively hints at the idea of AI as a method of invention. In this final section of the report, we discuss the implications of our study for the issues raised in the Introduction, followed by additional considerations, as well as suggestions for further study.

7.1 AI and Scientific Exploration

AlphaFold 2 influences the nature of the protein structures that researchers prioritise. The consistent reduction in TM Scores across AlphaFold 2-associated PDB submissions suggests a notable shift toward the experimental validation of less-explored, structurally unique proteins. This diversification reflects a move away from types of structures that are well-characterised, potentially unlocking insights in previously underexplored areas. This is reinforced by AlphaFold 2 supporting an increased focus on structures linked to rarer organisms compared to non-AI counterfactual developments, and with disease relevance compared to other AI counterfactuals. We observe these changes alongside an increased rate of PDB submissions associated with publications and academics linked to AlphaFold 2, particularly established researchers. These dynamics indicate AlphaFold 2's capacity to widen the scope of research that scientists consider carrying out, while also continuing to generate recognised research outputs. This supports the idea of AI as a tool for enhancing search and discovery and is in line with other similar studies on the use of AlphaFold 2 (Yu 2024).

It may be that AlphaFold 2 lowers the barrier and reduces the risks associated with research portfolio diversification, tipping the balance of their attention away from safer, well-understood areas where there are diminishing returns to further exploration. The patterns we observe align with the notion that predictive AI tools can facilitate more efficient resource allocation, by screening large knowledge spaces computationally and potentially allowing researchers to direct experimental efforts and applied research toward higher value targets (Agrawal et al. 2024). For AlphaFold 2, this appears to manifest as integration between greater exploration in sparser parts of the protein universe, and an increase in the volume of experimental work carried out. It is important to note, that these results are often equivalent in magnitude between established researchers and laboratories, suggesting that both individuals (with collaborators) and teams are able to reorganise resources to take advantage of AlphaFold 2.

These results are consistent with the idea that AlphaFold 2 is a complement to experimental and domain-specific work rather than a substitute. This adds texture to the "oil and water" phenomenon by suggesting that the lack of integration between AI and existing research within a discipline is not a fixed property, but potentially an attribute of specific AI tools and the problems they seek to solve (Duede et al. 2024b).

Our findings suggest that AI can make it easier for scientists to explore new areas in a specific knowledge space, providing evidence to counter the notion of problem of

'streetlight effects', in which scientists become more drawn to data-rich parts of a problem area via AI solutions (Tranchero et al. 2022). The exploration of unchartered problems in a field, coupled with facilitation of increased experimental activity points toward AI being a tool that could open up new research questions within a field. This demonstrates the feasibility for a system in which, at the meso level, AI can be used for high quality exploration in large and defined knowledge spaces, yielding results that can be exploited downstream to generate knowledge. It is important to note that this does not address other concerns that the uptake of AI will lead to a feedback loop in which scientists overly focus on problems that are suited to AI, or that overreliance on AI will diminish the ability of science to understand mechanisms that drive phenomena observed in the lab (Messeri and Crockett 2024). While it is possible that AI-enhanced exploration may itself open up surprising new research avenues, the policy decisions of funders and other agencies can also help guard against monocultures at the macro level, by continuing to encourage the pursuit of novel, path breaking research that might lean less heavily on AI.

7.2 AI and Research Productivity

AlphaFold 2's success supports the idea that AI-based predictive models can significantly advance scientific discovery when applied to problems with a defined objective, vast search spaces, and ample data or simulation capabilities, given the relative ineffectiveness of other modelling methods in that domain (Agrawal et al. 2024). The complex and unsolved challenge of protein folding prediction satisfies these criteria, and our analysis provides evidence of the consistent breadth and depth of impact AlphaFold 2 has had on that problem. As well as encouraging exploration leading to the experimental discovery of new protein structures, AlphaFold 2's ability to provide highly accurate protein structure predictions has enabled researchers to produce high impact work, as evidenced by an enhancement in citation impact across disciplines, and particularly within experimental structural biology. Other observed increases in citation impact, such as those in medical research, indicate AlphaFold 2's ability to not only advance fundamental insights in biochemistry, but also to assist the wider development of impactful, pointing to spillover effects that advance science as a whole.

One caveat to these findings is that the benefits of using AlphaFold 2 are not always equivalent for laboratories and established researchers. For example, in observations of publication output volume, laboratories appear to gain a modest comparative increase over researchers. Perhaps because laboratories are able to leverage economies of scale and flexibility in order to adopt new methods and incorporate them into larger scientific processes. Other factors that may play a role include the experience contained within labs, access to knowledge through larger collaboration networks, and resources and infrastructure, which may make it easier to leverage AI methods. However, researchers see a larger comparable increase in citation impacts, including field weighted-citation counts, suggesting they are targeting research avenues perceived as more valuable.

It is also important to consider the comparison to other counterfactual developments. The increase in publication production for laboratories using AlphaFold 2 is on a scale comparable to that achieved by laboratories using other non-AI and non-protein prediction frontier developments. This points to the dynamics of AlphaFold 2's adoption being comparable to developments built on more established techniques. It also underpins the idea that the ingredients required to take advantage of AI are not necessarily dissimilar to those already required to researchers to take advantage of building on the latest research.

Besides being high performing in protein structure prediction, AlphaFold 2 distinguished itself from many other innovations within structural biology by being based on a generalised, deep learning architecture. Some of our findings point to impacts where other AI innovations have struggled to make inroads in comparison. This, coupled with the fact that the reach of AlphaFold 2 also stretches beyond biochemistry, suggests that it presents a significant advancement on the state of AI in the field, while also pushing the state-of-the-art for AI in general. This highlights the potential compounding effect of AI advancements across science if adopted as a general method of invention.

AlphaFold 2 has made an impact in structural biology research, including experimentation, but does not excel in all areas of impact in comparison to the counterfactual frontier developments we have chosen. In particular, we see less evidence around translational impacts, which could be linked to the short elapsed timeframe available to study the impact and lags in transforming novel scientific discoveries into applications. Our results are indicative of the direction of travel for AI in science, but do not comprehensively cover the full breadth of impacts, nor the longer term potential benefits or drawbacks of its wider adoption. New improvements to AlphaFold 2 are already further enhancing its performance and addressing its limitations, widening the pool of researchers who will find it useful and highlighting the potentially accelerating effect that AI research could have on science (Abramson et al. 2024). Other recent developments in the ability to develop large language models with greatly reduced inputs suggest nonlinear progress in AI development may reduce some of the barriers required for researchers to build and use new tools (DeepSeek-AI et al. 2025).

7.3. AI and Invention

AlphaFold 2 demonstrates some contributions to translational research outputs in the form of clinical trials and patenting activity. In terms of citation impact, AlphaFold 2 linked papers exhibit increased clinical citation rates, significantly higher than those achieved by counterfactual developments, but this does not translate into higher likelihoods of researchers or laboratories being involved in clinical work. For patenting activity, AlphaFold 2 is associated with higher levels of activity for researchers and laboratories, but at a level comparable to other frontier developments in almost all cases.

AlphaFold 2 has a dual association with disease related research outputs. Although we observe the growing influence of AlphaFold 2 research across all fields, it is cited at a lower rate in disease relevant research when compared to baseline structural biology work. However, researchers and laboratories who adopt AlphaFold 2 see an increase in their likelihood of producing disease-relevant work. For laboratories, this is matched by comparable positive associations with AI and non-AI frontier developments. This points to a phenomenon where conventional methods still appear to be used more in disease related research at a community level, but the uptake of AlphaFold 2 by individual researchers or labs is associated with a shift towards disease relevant topics. These results do not provide a clear view of impact beyond the frontier.

Together, these findings lead us to speculate that AlphaFold 2 is having influence in applied areas of R&D alongside more established methods. Downstream researchers are potentially less ready to adopt new methods or leverage research that builds on them. This may be due to issues of trust, or because workflows for applied innovations consist of many more steps beyond protein structure prediction, into which new findings must be integrated. Together, our findings support the idea that AlphaFold 2 sits in a unique position in the biologist toolkit, leveraging AI techniques to generate clinically and commercially relevant outputs, although still in its infancy in terms of reaching full potential in more applied research.

7.4 Limitations

The findings presented in this report should be interpreted in light of several data and methodological limitations.

First, while we employ CEM and two-way interacted fixed effects to compare AlphaFold 2 influence with their peers, we cannot entirely rule out endogeneity. The decision by authors and laboratories to adopt AlphaFold 2 is likely not random. It may be influenced by unobserved strategic priorities, funding landscapes, or internal resource availability that our models cannot fully capture. We anticipate that our comparative design, which benchmarks the technology against other frontier methods, helps to negate some of these unobservable characteristics shared by early adopters. Nevertheless, despite these controls and comparisons, we cannot claim to have isolated a causal link in the absence of an experimental design.

Second, the comparative nature of the study is complicated by the varying release dates of the frontier developments. Counterfactual technologies published years before AlphaFold 2's announcement have had significantly longer to accrue usage and citations compared to our technology of interest. As impact metrics are partially a function of exposure time, older developments may appear more impactful simply due to their longevity. Furthermore, relying on patent activity serves as an imperfect proxy for commercial relevance. It captures only the subset of innovation that is protected intellectually and explicitly cites

academic work, likely missing a significant volume of industrial application that remains proprietary or unpublished. Consequently, our analysis captures only realised rather than latent value, and likely underestimates the total impact of frontier research in translational fields.

Third, our definition of adoption (or influence) relies on a methodological assumption that overestimates the intensity of use. By suggesting that researchers and laboratories have built on frontier developments based on past citations, we assume a persistent association across all subsequent quarters. While we interpret this not as a continuous link but rather as a signal of technological exposure, it implies a level of consistent reliance that may not reflect actual day-to-day usage of AlphaFold 2. This limitation affects all frontier groups in our sample but serves as an important caveat when interpreting the magnitude of researcher- and lab-level associations.

Fourth, scientific adoption is rarely mutually exclusive. Many researchers and laboratories in our sample integrate many frontier technologies rather than selecting a single method among our identified groups. However, to maintain interpretability and keep the number of reported coefficients manageable, our model treats the adoption of each frontier group as an independent variable rather than interacting them. This prevents us from isolating potential complementarities or substitution effects. Consequently, we cannot determine whether using AlphaFold 2 alongside other AI or non-AI methods is associated with synergistic benefits or redundant efforts.

Fifth, quality and coverage impose constraints on the granularity of our analysis. Our primary bibliometric source, OpenAlex, is a valuable resource but contains known noise regarding authorship disambiguation and institutional affiliations. Additionally, the classification of citation intent, used to distinguish between methodological use and background mention, relies on Semantic Scholar data which is not universally available. Attempts to limit our regression analysis to only those observations with confirmed intent often resulted in a loss of statistical power due to sample attrition. Consequently, even where effect signs remained consistent with our main findings, results for these subgroups occasionally lacked statistical significance.

Finally, our study presents aggregate and average effects that may mask important distributional impacts and that do not shed light on underlying mechanisms. This limits our ability to identify why particular results occur, how important subgroups might be affected, and how to interpret apparently conflicting findings. While this work is a contribution to the evidence base, which highlights some important dynamics in relation to the impact of AlphaFold 2 and AI in science, the interpretations and conclusions presented are broadly speculative.

7.5 Outlook

This report contributes to the evidence base on the patterns of use and impacts of AI in science. Besides a repetition of this analysis in the coming years to monitor progress on the trends we have explored, there are many further questions, some of which could be answered through an extension of this analysis[22]. In this outlook, we provide a summary of some of the possible future research directions.

Economic impacts

It has been estimated that the financial cost of protein structure discovery through experimental methods is around $100,000 and takes several years of researcher time (Hill and Stein 2020). Our research suggests that AlphaFold 2 might be expediting the process of protein structure determination, including for more novel structures that are more challenging to characterise. It is therefore fair to assume that this direct cost could be reduced through the use of AI, however the size of that reduction has not yet been estimated.

An attempt to approach this from an economic perspective would be to consider the association between AlphaFold 2, counterfactuals, and broader structural biology to submission redundancy (fourth quartile submission as seen in the UniProt structure). The difference in first submissions, relative to these redundant submissions, could be interpreted as efficiency gains, which can then be interpreted economically using figures from Hill and Stein. A further extension would be to examine any displacement effects on funding that AlphaFold 2 has had, to identify whether the efficiency gains are reinvested to accelerate structural biology, shifted to other topics or reduced overall. A preliminary exploration of NIH funding data in this study suggested that there has been a gradual increase in AI-related funding for structural biology research, aligned with wider AI trends, and no substitution away from experimental structural biology research. This type of research received record funding in 2023.

This study has also been limited in its study of labour market dynamics. It is reasonable to consider that AlphaFold 2 may cause displacement of scientific workers through the efficiency gains in protein structure prediction, and by rendering some research activities less relevant. It has been reported that some researchers have shifted away from protein structure prediction to other avenues of research in the wake of AlphaFold 2 (Saplakoglu 2024). Our overall findings suggest that there is limited evidence for displacement happening on a large scale, as we observe an increase in the number of researchers participating in research stemming from AlphaFold 2, with no accompanying decrease in researchers leveraging other advances in structural biology. Nonetheless, our work does not track individual researchers and therefore cannot account for inflows and outflows that may still be occurring underneath these high level trends.

[22] The code to do this can be found in the GitHub repository for this work.

Access to resources and networks

Our study has treated researchers as individual units of analysis. This neglects the potential impacts of collaboration on both AI adoption and the ability of researchers to translate new findings into downstream applications. A future study might investigate the relationship between the dimensions we have explored and the collaboration networks that researchers are part of.

In a similar vein, some of our findings support the idea that more established researchers are better positioned to take advantage of frontier developments including AI methods such as AlphaFold 2. It would be useful to characterise what factors beyond experience and being well established have an impact on AI adoption. For example, examining the impact of laboratory funding, access to physical and cloud infrastructure and the geographical distribution of AI related research activity could shine a light on this dynamic.

Transformation

While we have observed early signals of increased exploration, research productivity gains, and influence on downstream applications, our study has not accessed information about the qualitative process of science using AI, nor whether it is changing the overall trajectory of scientific research. Additional studies that took advantage of text data from publications, as well as domain knowledge and qualitative information gathered directly from researchers about the research process would help to understand how AI is changing the process of science, the extent to which it is influencing how researchers think and whether it is changing the direction of science at the macro scale as well as its pace. and increased exploration complement narrow task oriented benchmarks, helping to understand whether AI is having a significant transformative effect in science.

Acknowledgements

The Innovation Growth Lab acknowledges support from Google DeepMind for enabling the development of this work.

We also thank Oleg Kovalevskiy, Julian Jacobs, Zanna Iscenko, and Juan Mateos-Garcia for their comments on this work.

Disclaimer

No experiments or analyses were carried out by or at Google DeepMind. All points of view and conclusions expressed are those of the Innovation Growth Lab and do not necessarily reflect the position or endorsement of Google DeepMind. Control of the methodology, research and editorial decisions were retained by the Innovation Growth Lab.

Attribution

Code

The code for this analysis can be found in the GitHub project repository.

Bibliography

Abadi, Martín, Paul Barham, Jianmin Chen, et al. 2016. 'Tensorflow: A System for Large-Scale Machine Learning'. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–83.

Abramson, Josh, Jonas Adler, Jack Dunger, et al. 2024. 'Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3'. *Nature* 630 (8016): 493–500. https://doi.org/10.1038/s41586-024-07487-w.

Agrawal, Ajay, John McHale, and Alexander Oettl. 2018. 'Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth'. *NBER*, April.

Agrawal, Ajay, John McHale, and Alexander Oettl. 2024. 'Artificial Intelligence and Scientific Discovery: A Model of Prioritized Search'. *Research Policy* 53 (5): 104989. https://doi.org/10.1016/j.respol.2024.104989.

AlQuraishi, Mohammed. 2020. 'AlphaFold 2 @ CASP14: "It Feels like One's Child Has Left Home."' *Some Thoughts on a Mysterious Universe*, December 8. https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/.

Andrychowicz, Marcin, Lasse Espeholt, Di Li, et al. 2023. 'Deep Learning for Day Forecasts from Sparse Observations'. arXiv:2306.06079. Preprint, arXiv, July 6. https://doi.org/10.48550/arXiv.2306.06079.

Bail, Christopher A. 2024. 'Can Generative AI Improve Social Science?' *Proceedings of the National Academy of Sciences* 121 (21): e2314021121. https://doi.org/10.1073/pnas.2314021121.

Bertoline, Letícia M. F., Angélica N. Lima, Jose E. Krieger, and Samantha K. Teixeira. 2023. 'Before and after AlphaFold 2: An Overview of Protein Structure Prediction'. *Frontiers in Bioinformatics* 3 (February). https://doi.org/10.3389/fbinf.2023.1120370.

Bianchini, Stefano, Moritz Müller, and Pierre Pelletier. 2022. 'Artificial Intelligence in Science: An Emerging General Method of Invention'. *Research Policy* 51 (10): 104604. https://doi.org/10.1016/j.respol.2022.104604.

Bianchini, Stefano, Moritz Müller, and Pierre Pelletier. 2024. 'Drivers and Barriers of AI Adoption and Use in Scientific Research'. arXiv:2312.09843. Preprint, arXiv, February 22. https://doi.org/10.48550/arXiv.2312.09843.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2023. 'Generative AI at Work'. arXiv:2304.11771. Preprint, arXiv, April 23. https://doi.org/10.48550/arXiv.2304.11771.

Brynjolfsson, Erik, and Tom Mitchell. 2017. 'What Can Machine Learning Do? Workforce Implications'. *Science* 358 (6370): 1530–34.

Brzezinski, Peter, Heiner Linke, Johan Åqvist, Ann Fernholm, and Vincent von Sydow. 2024. 'The Nobel Prize in Chemistry 2024'. NobelPrize.Org, November 22. https://www.nobelprize.org/prizes/chemistry/2024/popular-information/.

Callaway, Ewen. 2020. '"It Will Change Everything": DeepMind's AI Makes Gigantic Leap in Solving Protein Structures'. *Nature* 588 (7837): 203–4. https://doi.org/10.1038/d41586-020-03348-4.

Chenthamarakshan, Vijil, Samuel C. Hoffman, C. David Owen, et al. 2023. 'Accelerating Drug Target Inhibitor Discovery with a Deep Generative Foundation Model'. *Science Advances* 9 (25): eadg7865. https://doi.org/10.1126/sciadv.adg7865.

Cohan, Arman, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. 'Structural

Scaffolds for Citation Intent Classification in Scientific Publications'. *Proceedings of the 2019 Conference of the North*, 3586–96. https://doi.org/10.18653/v1/N19-1361.

Dana, Jose M, Aleksandras Gutmanas, Nidhi Tyagi, et al. 2019. 'SIFTS: Updated Structure Integration with Function, Taxonomy and Sequences Resource Allows 40-Fold Increase in Coverage of Structure-Based Annotations for Proteins'. *Nucleic Acids Research* 47 (D1): D482–89. https://doi.org/10.1093/nar/gky1114.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. 2025. 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning'. arXiv:2501.12948. Preprint, arXiv, January 22. https://doi.org/10.48550/arXiv.2501.12948.

Dessimoz, Christophe, and Paul D. Thomas. 2024. 'AI and the Democratization of Knowledge'. *Scientific Data* 11 (1): 268. https://doi.org/10.1038/s41597-024-03099-1.

Doan Tran, Huan, Chiho Kim, Lihua Chen, et al. 2020. 'Machine-Learning Predictions of Polymer Properties with Polymer Genome'. *Journal of Applied Physics* 128 (17): 171104. https://doi.org/10.1063/5.0023759.

Duede, Eamon. 2022. 'Instruments, Agents, and Artificial Intelligence: Novel Epistemic Categories of Reliability'. Preprint. November 6. https://philsci-archive.pitt.edu/21352/.

Duede, Eamon, William Dolan, André Bauer, Ian Foster, and Karim Lakhani. 2024a. 'Oil & Water? Diffusion of AI Within and Across Scientific Fields'. arXiv:2405.15828. Preprint, arXiv, May 23. http://arxiv.org/abs/2405.15828.

Duede, Eamon, William Dolan, André Bauer, Ian Foster, and Karim Lakhani. 2024b. 'Oil & Water? Diffusion of AI Within and Across Scientific Fields'. arXiv:2405.15828. Preprint, arXiv, May 23. http://arxiv.org/abs/2405.15828.

Duran-Frigola, Miquel, Roberto Mosca, and Patrick Aloy. 2013. 'Structural Systems Pharmacology: The Role of 3D Structures in Next-Generation Drug Development'. *Chemistry & Biology* 20 (5): 674–84. https://doi.org/10.1016/j.chembiol.2013.03.004.

Elofsson, Arne. 2023. 'Progress at Protein Structure Prediction, as Seen in CASP15'. *Current Opinion in Structural Biology* 80 (June): 102594. https://doi.org/10.1016/j.sbi.2023.102594.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. 'GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models'. arXiv.Org, March 17. https://arxiv.org/abs/2303.10130v5.

Enoma, David O., Janet Bishung, Theresa Abiodun, Olubanke Ogunlana, and Victor Chukwudi Osamor. 2022. 'Machine Learning Approaches to Genome-Wide Association Studies'. *Journal of King Saud University - Science* 34 (4): 101847. https://doi.org/10.1016/j.jksus.2022.101847.

Evans, Philip, and Airlie McCoy. 2008. 'An Introduction to Molecular Replacement'. *Acta Crystallographica Section D: Biological Crystallography* 64 (Pt 1): 1–10. https://doi.org/10.1107/S0907444907051554.

Evans, Richard, Michael O'Neill, Alexander Pritzel, et al. 2022. 'Protein Complex Prediction with AlphaFold-Multimer'. Preprint, bioRxiv, March 10. https://doi.org/10.1101/2021.10.04.463034.

Fontaine, Sylvain, Floriana Gargiulo, Michel Dubois, and Paola Tubaro. 2024. 'Epistemic Integration and Social Segregation of AI in Neuroscience'. arXiv:2310.01046. Preprint, arXiv, March 6. http://arxiv.org/abs/2310.01046.

Gao, Jian, and Dashun Wang. 2024. 'Quantifying the Benefit of Artificial Intelligence for Scientific Research'. arXiv:2304.10578. Preprint, arXiv, May 31.

https://doi.org/10.48550/arXiv.2304.10578.
Gargiulo, Floriana, Sylvain Fontaine, Michel Dubois, and Paola Tubaro. 2022. 'A Meso-Scale Cartography of the AI Ecosystem'. arXiv:2212.12263. Preprint, arXiv, December 23. http://arxiv.org/abs/2212.12263.
Griffin, Conor, Don Wallace, Juan Mateos-Garcia, Hanna Schieve, and Pushmeet Kohli. 2024. 'A New Golden Age of Discovery'. Google DeepMind, November 20. https://deepmind.google/public-policy/ai-for-science/.
Hammond, Mark. 2023. 'The Path to Applied Science Co-Pilots.' Substack newsletter. *Accelerating Applied Science by Deep Science Ventures*, December 15. https://acceleratingappliedscience.substack.com/p/the-path-to-applied-science-co-pilots.
Hazra, Satakshi, and Sanjukta Patra. 2021. 'Importance of Protein Structure and Function in Pathogenesis: Highlights on the Multifaceted Organism *Mycobacterium Tuberculosis*'. In *Microbial and Natural Macromolecules*, edited by Surajit Das and Hirak Ranjan Dash. Academic Press. https://doi.org/10.1016/B978-0-12-820084-1.00030-2.
Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. *Predicting Results of Social Science Experiments Using Large Language Models*. August 8.
Higgins, Matthew K. 2021. 'Can We AlphaFold Our Way Out of the Next Pandemic?' *Journal of Molecular Biology*, From Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology, vol. 433 (20): 167093. https://doi.org/10.1016/j.jmb.2021.167093.
Hill, Ryan, and Carolyn Stein. 2020. *Race to the Bottom: Competition and Quality in Science*. November 5.
Huang, Kexin, Payal Chandak, Qianwen Wang, et al. 2024. 'A Foundation Model for Clinician-Centered Drug Repurposing'. *Nature Medicine*, September 25, 1–13. https://doi.org/10.1038/s41591-024-03233-x.
Idalski Carcone, April, Mehedi Hasan, Gwen L Alexander, et al. 2019. 'Developing Machine Learning Models for Behavioral Coding'. *Journal of Pediatric Psychology* 44 (3): 289–99. https://doi.org/10.1093/jpepsy/jsy113.
Jakubik, Johannes, Sujit Roy, C. E. Phillips, et al. 2023. 'Foundation Models for Generalist Geospatial Artificial Intelligence'. arXiv:2310.18660. Preprint, arXiv, November 8. https://doi.org/10.48550/arXiv.2310.18660.
Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. 'Highly Accurate Protein Structure Prediction with AlphaFold'. *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.
Kempen, Michel van, Stephanie S. Kim, Charlotte Tumescheit, et al. 2024. 'Fast and Accurate Protein Structure Search with Foldseek'. *Nature Biotechnology* 42 (2): 243–46. https://doi.org/10.1038/s41587-023-01773-0.
Krenn, Mario, Robert Pollice, Si Yue Guo, et al. 2022. 'On Scientific Understanding with Artificial Intelligence'. *Nature Reviews Physics* 4 (12): 761–69. https://doi.org/10.1038/s42254-022-00518-3.
McNamee, Laura M., and Fred D. Ledley. 2017. 'Modeling Timelines for Translational Science in Cancer; the Impact of Technological Maturation'. *PLOS ONE* 12 (3): e0174538. https://doi.org/10.1371/journal.pone.0174538.
Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. 'Scaling Deep Learning for Materials Discovery'. *Nature* 624 (7990): 80–85. https://doi.org/10.1038/s41586-023-06735-9.
Messeri, Lisa, and M. J. Crockett. 2024. 'Artificial Intelligence and Illusions of Understanding

in Scientific Research'. *Nature* 627 (8002): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. 'ColabFold: Making Protein Folding Accessible to All'. *Nature Methods* 19 (6): 679–82. https://doi.org/10.1038/s41592-022-01488-1.

Noy, Shakked, and Whitney Zhang. 2023. 'Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence'. *Science* 381 (6654): 187–92. https://doi.org/10.1126/science.adh2586.

Paszke, Adam, Sam Gross, Francisco Massa, et al. 2019. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. 'Scikit-Learn: Machine Learning in Python'. *Journal of Machine Learning Research* 12 (Oct): 2825–30.

Perrakis, Anastassis, and Titia K Sixma. 2021. 'AI Revolutions in Biology'. *EMBO Reports* 22 (11): e54046. https://doi.org/10.15252/embr.202154046.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. 'OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts'. arXiv:2205.01833. Preprint, arXiv, June 17. https://doi.org/10.48550/arXiv.2205.01833.

Saplakoglu, Yasemin. 2024. 'How AI Revolutionized Protein Science, but Didn't End It'. Quanta Magazine, June 26. https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/.

Savage, Tom, Nausheen Basha, Jonathan McDonough, James Krassowski, Omar Matar, and Ehecatl Antonio del Rio Chanona. 2024. 'Machine Learning-Assisted Discovery of Flow Reactor Designs'. *Nature Chemical Engineering* 1 (8): 522–31. https://doi.org/10.1038/s44286-024-00099-1.

Schweidtmann, Artur M., Adam D. Clayton, Nicholas Holmes, Eric Bradford, Richard A. Bourne, and Alexei A. Lapkin. 2018. 'Machine Learning Meets Continuous Flow Chemistry: Automated Optimization towards the Pareto Front of Multiple Objectives'. *Chemical Engineering Journal* 352 (November): 277–82. https://doi.org/10.1016/j.cej.2018.07.031.

Skarlinski, Michael D., Sam Cox, Jon M. Laurent, et al. 2024. 'Language Agents Achieve Superhuman Synthesis of Scientific Knowledge'. arXiv:2409.13740. Preprint, arXiv, September 26. https://doi.org/10.48550/arXiv.2409.13740.

Sønderby, Casper Kaae, Lasse Espeholt, Jonathan Heek, et al. 2020. 'MetNet: A Neural Weather Model for Precipitation Forecasting'. arXiv:2003.12140. Preprint, arXiv, March 30. http://arxiv.org/abs/2003.12140.

Sumner, James. 2024. 'The AI Revolution Is Always Just out of Reach'. *Nature Reviews Physics* 6 (11): 1–2. https://doi.org/10.1038/s42254-024-00777-2.

The UniProt Consortium. 2023. 'UniProt: The Universal Protein Knowledgebase in 2023'. *Nucleic Acids Research* 51 (D1): D523–31. https://doi.org/10.1093/nar/gkac1052.

Toner-Rodgers, Aidan. 2024. *Artificial Intelligence, Scientific Discovery, and Product Innovation*.

Tranchero, Matteo. 2024. 'Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation'. *Academy of Management Proceedings* 2024 (1): 13751. https://doi.org/10.5465/AMPROC.2024.241bp.

Tranchero, Matteo, Gustavo Manso, Johannes Hoezlemann, and Abhishek Nagaraj. 2022. 'The Streetlight Effect in Data-Driven Exploration'. Preprint, October 26. https://www.abhishekn.com/publications-all/the-streetlight-effect-in-data-driven-exploration.

Van Noorden, Richard, and Jeffrey M. Perkel. 2023. 'AI and Science: What 1,600 Researchers Think'. *Nature* 621 (7980): 672–75. https://doi.org/10.1038/d41586-023-02980-0.

Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, 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): D439–44. https://doi.org/10.1093/nar/gkab1061.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2023. 'Attention Is All You Need'. arXiv:1706.03762. Preprint, arXiv, August 2. http://arxiv.org/abs/1706.03762.

Wang, Dashun. 2024. 'How I'm Using AI Tools to Help Universities Maximize Research Impacts'. *Nature* 630 (8018): 794–794. https://doi.org/10.1038/d41586-024-02081-6.

Wang, Hanchen, Tianfan Fu, Yuanqi Du, et al. 2023. 'Scientific Discovery in the Age of Artificial Intelligence'. *Nature* 620 (7972): 47–60. https://doi.org/10.1038/s41586-023-06221-2.

Yu, Zhengyi. 2024. 'The Impacts of Prediction AI on Scientists: Evidence from AlphaFold'. SSRN Scholarly Paper No. 4711334. Rochester, NY, February 5. https://doi.org/10.2139/ssrn.4711334.

Zunger, Alex. 2018. 'Inverse Design in Search of Materials with Target Functionalities'. *Nature Reviews Chemistry* 2 (4): 1–16. https://doi.org/10.1038/s41570-018-0121.

Appendix

A. Data Sources

We consider several sources of data to analyse the influence of AlphaFold 2 and related innovations. We focus on publications, patents, PDB submissions, and clinical articles between 2018 and 2025, spanning from AlphaFold 2's development and release until almost the present day[23].

OpenAlex is a database of academic papers and authors (Priem et al. 2022). It provides detailed information such as which papers cite each other, the main topics and concepts of research, details about authors, and measures of impact like yearly citation counts. We use OpenAlex to create datasets on papers, authors, and labs, collecting nearly 5 million unique publications. For each publication, we collect external IDs, dates, authorships, citation counts, concepts, medical concepts (Medical Subject Headings), topics, grants, year-normalised citation counts, and field-weighted citation counts. Institutional data, including location, publication counts, h-index, and i10 index, are collected as extended controls.

PubMed is a widely used search engine for medical and life science research. Using PubMed IDs obtained from OpenAlex, we supplement our dataset with PubMed data to identify the clinical impact of research related to AlphaFold 2. In particular, we match OpenAlex publications to their PubMed counterparts, and collect lists of PubMed IDs corresponding to clinical citations available as metadata, which were classified using machine learning on article metadata and MeSH terms (Hutchins et al. 2019). In extensions of this work, we further classify these clinical citations to differentiate between clinical trials, randomised trials, and observational studies.

The Lens combines academic and patent information. It links research papers to patents by matching unique DOIs. This connection provides information into how relevant research contributes to commercial science. We obtain patent citation counts for all DOI-matching publications. We additionally collect patent-to-patent citations and classifications of technology codes. An approximate 29,000 citations are identified as relevant to the OpenAlex data we collect.

The **Protein Data Bank (PDB)** is the primary repository for experimentally determined protein structures. We collect metadata on over 227,000 protein submissions, identifying nearly 79,000 publications linked to these structures in OpenAlex. To address missing PMIDs and DOIs, we supplement PDB records with identifiers from the RCSB repository, achieving 87% matching coverage. The PDB metadata includes attributes such as R-free values and resolution. Additionally, we use SIFTS to map RCSB IDs to UniProt IDs,

[23] As of June 2025, PDB submissions from the RCSB's main API accesses are only available up to the end of 2024.

enabling the collection of data for all associated sequence chains (Dana et al. 2019). We also use Foldseek to measure similarity metrics between structures (van Kempen et al. 2024).

UniProtKB is a comprehensive resource for protein sequence and functional information. Using mappings from PDB, we integrate additional metadata, including the timeline of structural definitions, functional annotations, references to associated diseases, and sequence-level details (The UniProt Consortium 2023). These data points allow the derivation of variables such as the number of associated diseases, functional complexity, and organism rarity.

Semantic Scholar adds more detail to the citation data from OpenAlex. It provides information about the intent behind citations and flags particularly influential papers. Intent labels include background, result, and methodology (Cohan et al. 2019). Their labelling algorithm relies on open-access publications, limiting the scope to publications with an open URL. An approximate 37% of citation links are found to contain intent measures.

B. Empirical Strategy

We use a difference-in-differences (DiD) framework to estimate the effects of AlphaFold 2 on research productivity and translational impact. By leveraging the staggered adoption of AlphaFold 2 and differences in exposure to counterfactual technologies, this design choice allows us to measure its effects relative to these alternatives, as well as assess the joint influence of other AI and non-AI frontier research on authors' and labs' outcomes.

The main regression model is specified as:

$$Y_{i,t} = \beta_0 + \beta_1 \mathrm{AF}_{i,t} + \beta_2 \mathrm{CT}_{i,t} + \beta_3 (\mathrm{AF}_{i,t} \times \mathrm{CT}_{i,t}) + \mathbf{X}_{i,t}\gamma + \alpha_i + \delta_t + \epsilon_{i,t}$$

where $Y_{i,t}$ represents the outcome of interest (e.g., publications, PDB submissions, or citations) for researcher or lab i at time t. The variable $\mathrm{AF}_{i,t}$ indicates AlphaFold 2 adoption, and $\mathrm{CT}_{i,t}$ denotes the use of complementary technologies, either AI-enabled or traditional. The interaction term $\mathrm{AF}_{i,t} \times \mathrm{CT}_{i,t}$ captures potential synergies or redundancies between AlphaFold 2 and these tools. Fixed effects (α_i) control for unobserved differences across researchers or labs, while time fixed effects (δ_t) absorb temporal trends. Standard errors are clustered at the researcher or lab level to account for within-entity correlation over time.

The design incorporates covariates to minimise endogeneity concerns and maximise comparability between treatment and control groups. These include institutional metrics such as citedness and h-index, researcher-level attributes like prior productivity and field specialisation, and field-level fixed effects to address variation across disciplines.

Additionally, COVID-19-related publication intensity during 2020 helps adjust for unobservables that may be proxied by pandemic-induced shifts in research focus.

Selection bias remains a key challenge, as researchers building on AlphaFold 2 may differ systematically from non-users. To balance pre-treatment characteristics between treated and control groups, we use CEM to match on citation counts, publication records, institutional location, and COVID-19 research activity. We also attempt to minimise concerns about omitted variable bias by considering subgroup analyses based on high PDB activity and field-specific outputs, which provide additional controls for differences in exposure and research focus.

C. Sample Construction

The goal is to construct datasets that allow for a detailed examination of AlphaFold 2's influence on structural biology and related fields. This involves identifying relevant publications, constructing citation chains, and addressing potential confounders such as citation intent and prior expertise. We create datasets at the paper, author, and laboratory level to account for multiple diffusion pathways. The analysis is restricted to the period from the first quarter of 2018 to the first quarter of 2025.

Baselines The core data for this study focuses on three pivotal AlphaFold 2 papers, which form the starting point for the analysis:

1. **Highly accurate protein structure prediction with AlphaFold 2** (Jumper et al, 2021), which introduced the breakthrough innovation in protein structure prediction.
2. **Protein complex prediction with AlphaFold 2-Multimer** (Evans et al, 2022), an upgraded model that allows for the prediction of multi-chain protein structures.
3. **AlphaFold 2 Protein Structure Database** (Varadi et al, 2022), which made AlphaFold 2 predictions available to structural biologists.

These three publications form the initial reference point for our sample construction. In order to define a baseline, we turn to OpenAlex to collect relevant metadata from all structural biology papers published over the past decade. OpenAlex tags publications with multiple concepts, enabling us to filter papers relevant to the domain. Google DeepMind provided a curated list of structural biology concepts, including terms such as "protein structure prediction," "protein folding" and "protein sequencing," which were used to identify relevant publications. This approach results in a dataset comprising 102,457 structural biology papers published between 2018 and early 2025.

Citation chains To measure the direct and indirect influence of AlphaFold 2, we construct citation chains from our initial set of 102,457 structural biology papers published since 2018. This process begins by identifying all publications that cite the core AlphaFold 2 papers; these form the first level of foundational papers, yielding 32,810 unique entries

tied to AlphaFold 2. Any of the 102,457 baseline structural biology papers not citing AlphaFold 2 constitute the pool for our baseline and counterfactual groups.

To ensure a clean comparison, we perform several reassignments. First, any paper from this baseline pool that is found to cite AlphaFold 2 is reassigned to the AlphaFold 2 chain. Second, papers identified as belonging to the counterfactual frontier groups (as described in the "Counterfactual construction" subsection below) are also separated out. To manage the computational demands of processing citation intents for the large number of remaining non-AF and non-CT foundational papers (a significant bottleneck in our approach) this final set was randomly sampled by one-half. This procedure resulted in 29,912 foundational papers for the "other structural biology" group (a more recent update of the data brings this number up to 41,039).

Finally, we expand the network for all groups by tracking two additional levels of downstream citations to capture broader research applications. These publications capture broader research, including studies that apply or extend methodologies developed in the foundational papers or build upon their findings in adjacent fields. This approach allows us to trace the diffusion of influence from structural biology to more applied domains, particularly within medical sciences. This yields 514,348 applied papers for the AlphaFold 2 chain and 579,797 for the "other structural biology" group (a recent update of the data brings the number of applied papers to 640,018). To avoid duplication, articles appearing at multiple levels of separation are assigned exclusively to their closest link to a foundational paper.

Citation intent One challenge in using citations is that references can serve multiple purposes: they may acknowledge prior work, report results, or, crucially for our design, describe methodological uses. The broad diffusion of AlphaFold 2 in structural biology means that citations may reflect different types of usage. To account for this variation, we classify citations according to their intent. Specifically, we distinguish between citations that serve as:

- Background references (i.e., acknowledging the state of the field or providing context)
- Research results (i.e., citing AlphaFold 2 or a baseline paper for its findings)
- Methodological uses (i.e., citing AlphaFold 2 or a baseline paper as a tool or technique employed in new research).
- Influential citations (i.e., the cited publication has a significant impact on the citing publication)

To identify relevant uses, we use Semantic Scholar's intent and influence metrics to classify whether a publication credits AlphaFold 2 or a baseline structural biology paper for specific contributions to their research methodology. This approach allows us to focus on the subset of citations that reflect a direct contribution to the advancement of the field

through a method, rather than through a citation merely acknowledging a paper's existence.

We collect citation context data for 36% of AlphaFold 2-connected publications. Among these, 68% of foundational papers exhibit methodological use of the technology, while 40% of applied publications demonstrate at least partial methodological use of downstream AlphaFold 2-related research. For foundational papers not connected to AlphaFold 2, citation intent data is available for 32% of publications. Within this group, 33% of applied papers show at least partial methodological use of their foundational references.

Counterfactual construction To evaluate whether the impacts of AlphaFold 2 are distinct or reflective of broader trends in scientific innovation, we construct a set of counterfactual publications. These are high-impact structural biology papers that align with AlphaFold 2 in terms of citation dynamics and research focus but do not involve AlphaFold 2. This approach provides a robust basis for comparison, isolating the unique contributions of AlphaFold 2 from general developments in the field.

We define two sets of counterfactuals:

- **AI-intensive** developments, which leverage machine learning or AI-based techniques for protein structure prediction.
- **Non-AI protein prediction developments,** which rely on more mature, non-AI innovations specifically for predicting protein structures (e.g., homology modeling).
- **Other frontier structural biology methods**, which represent significant non-AI innovations in areas outside of direct structure prediction (e.g., advancements in cryo-EM or X-ray crystallography).

These counterfactual papers help us compare AlphaFold 2's effects with the impacts of other innovations in the field of structural biology. Citation chains are similarly constructed for the publications indirectly influenced by these counterfactual papers. This procedure allows us to identify the dynamics surrounding the three innovation pathways and compare them to AlphaFold 2's influence on the field.

To construct the counterfactual groups, we filter foundational non-AlphaFold 2 papers based on OpenAlex concept matches and a minimum threshold of publications. We calculate the Euclidean distance between the proportions of methodological and influential citations for each paper, relative to AlphaFold 2 citation patterns. Papers in the bottom quintile of these distances are identified as having citation intent distributions most similar to AlphaFold 2. From this subset, we manually curate a list of 90 counterfactuals, evenly divided between AI-intensive and non-AI-driven innovations.

Citation chains are then constructed for these counterfactual papers, yielding 18,272 foundational papers and 582,964 applied papers. By design, the citation intent and influence distribution of these counterfactuals closely resembles that of AlphaFold 2, with

74% of foundational citations and 35% of applied publications indicating at least partial methodological use of the upstream innovation.

D. Human capital and knowledge diffusion

Beyond the use of direct citations, AI innovations like AlphaFold 2 may influence subsequent research trajectories, collaboration patterns, and productivity. To investigate these effects, we focus on the roles of individual authors and the labs in which foundational and applied research is conducted.

Authors We collect first-author publication records for all researchers involved in any citation chain linked to AlphaFold 2, counterfactual innovations, or baseline structural biology papers. These records cover the period from 2015 Q1 onwards, capturing longitudinal trends in research output. In total, 3,826,195 unique authors are identified. To examine differences across career stages, we categorise researchers based on their publication timelines. Early career researchers (ECRs) are defined as those who began publishing in 2020 or later, with no prior records in the dataset.

These classifications provide a straightforward segmentation to study AlphaFold 2's differential impact on researchers at varying stages of their careers. In particular, it provides a basis to investigate whether ECRs, less bound by established research practices, adopt and benefit from AI innovations differently compared to ERs, who may rely on domain expertise to integrate new methodologies.

Laboratories Research labs serve as hubs for foundational and applied research. They are key units of scientific productivity, often led by principal investigators (PIs) who shape research direction and collaboration dynamics. Our approach combines author metadata from OpenAlex to identify these PIs and their associated labs.

Building on the existing literature (Galasso et al, 2023), we exploit authorship patterns across publications to determine leadership roles. Specifically, PIs are identified as authors who frequently occupy the last-author position in papers, a standard indicator of laboratory leadership in many scientific fields.

To formalise this, we calculate an Author Position Factor (APF) for each author, which weights their contribution since 2014 across first, middle, and last-author roles. High APF scores over a rolling three-year window are used to flag researchers as candidate PIs.

$$\mathrm{APF}_{i,t} = \left[p_i^{\text{author last}} + \frac{1}{2} p_i^{\text{author first}} \right] \cdot \left(1 - p_i^{\text{author middle}}\right)$$

Lab identification is refined through additional criteria, including publication and citation metrics. Candidates are considered if they appear in at least three consecutive years. The score of a lab is defined as:

$$LIS^{b}_{i,t} = 0.4 \cdot APF_{i,t} + 0.1 \cdot \widehat{APF}_{i,t} + 0.2 \cdot \widehat{APF^{3y}}_{i,t} + 0.2 \cdot PC_{i,t} + 0.1 \cdot \log(CC_{i,t}+1)$$

Candidates are considered if their PIs exhibit consistent lab identification scores, defined by at least three consecutive years of top quartile scores for foundational candidates, and three consecutive years of top decile scores for applied candidates.

This procedure identifies 9,722 highly-likely foundational PIs and 12,981 highly-likely applied PIs, associated with a total of 1,043,702 unique foundational papers and 1,659,277 unique applied papers. Among foundational PIs, 8,251 cite AlphaFold 2 at least once, with 56% of these labs showing at least one methodological citation. For applied labs, 9,612 cite downstream applications of AlphaFold 2 at least once, 63% of which have at least a partially methodological citation.[24]

E. Sub-groups of interest

Fields To better understand how AlphaFold 2 and other AI innovations influence research dynamics, we examine variation across specific sub-groups defined by their primary research focus. Using OpenAlex's field taxonomy, developed by Van Eck and Waltman (2024), we assign each publication a primary topic. This allows us to capture field-specific diffusion patterns and assess whether AlphaFold 2 has distinct impacts across foundational and applied domains.

Our analysis focuses on two primary topic categories:

- **Biochemistry, Molecular Biology, and Genetics**: This field encompasses much of the foundational research within the citation chains, including structural biology papers that contribute to core methodological advances.
- **Medicine**: This category represents applied research, where structural biology findings are leveraged for clinical applications, including drug discovery and diagnosis.

This classification helps disentangle AlphaFold 2's effects on fundamental scientific discovery from its impacts on translational research. Figures in the Descriptive section illustrate the distribution of topics across foundational and applied work in the citation chains, highlighting the concentration of AlphaFold 2-related studies within these domains. Secondary topics, while not the primary focus, offer additional context for specific research directions. To facilitate quarterly analysis, we calculate both quarter-specific field shares and the mode of publications' primary field.

[24] An earlier version of this paper in June 2024 had only 3,044 foundational PIs citing AlphaFold 2 at least once, with 42% of these labs showing at least one methodological citation. For applied labs, 9,612 cited downstream applications of AlphaFold 2 at least once, 55% of which have at least a partially methodological citation

Experimental exposure We identify a subset of experimental structural biologists particularly engaged with protein prediction methodologies by leveraging submission data from the Protein Data Bank (PDB). The PDB, as a repository for experimentally validated protein structures, provides a proxy for experimental structural biology work, which is particularly exposed to predictive tools like AlphaFold 2 and its counterfactuals.

To define exposure levels, we compute the proportion of PDB-linked publications relative to the total number of publications for both authors and labs. In particular, we compute shares of PDB submissions in the years 2018 and 2019 for authors and labs, and classify as high-exposure those in the top quartile.

F. Coarsened Exact Matching

To enable robust comparisons between researchers and labs using AlphaFold 2, counterfactual innovations, and baseline structural biology approaches, we implement Coarsened Exact Matching (CEM) as described by Iacus et al. (2012). This method helps identify comparable control authors and labs by aligning pre-treatment characteristics, reducing potential bias in the analysis.

The matching procedure uses several pre-treatment variables to align groups. Citation and publication counts before 2020 capture baseline research impact and productivity. PDB submission shares reflect engagement in experimental structural biology, while institutional affiliation and country of origin account for differences in research environments. COVID-19-related output is also included to control for its influence on research priorities during the pandemic.

CEM involves grouping continuous variables, such as citations and PDB activity, into discrete bins to simplify matching. Researchers and labs are then matched exactly across these bins to ensure similarity in their pre-treatment characteristics. Once matched, researchers and labs are assigned to AlphaFold 2 user, counterfactual user, or baseline control groups. This procedure produces a well-balanced dataset, with 86% of AlphaFold 2 users successfully matched to counterparts in control groups.[25]

G. Bootstrapping Citation Intent

Citation intent data is only available from Semantic Scholar for 31% of publications in our dataset that contains structural biology publications, including those stemming from AlphaFold 2 and counterfactuals. For descriptive measures of reach, we use a stratified bootstrapping method.

To estimate numbers of publications, we first take the publications in citation chains for which complete citation intent data is available. This is then stratified according to

[25] This proportion drops to 13% for principal investigators due to the now-imbalanced sample.

properties of each publication: the primary field associated according to OpenAlex and the month and year of publication, and whether the publication is directly or indirectly citing a core publication. We then randomly resample this set many times with replacement to generate a distribution of the proportion of strong and weak citation chains. These distributions are then applied to the size of the entire dataset to create a distribution of the number of papers in citation chains that are strong versus weak.

We take a similar approach to estimating the number of researchers impacted by AlphaFold and the counterfactual papers. The difference is that within each stratum and iteration, each paper in the unlabelled data is randomly labelled as being a strong or weak citation chain, according to the resampled proportions. We then count the number of unique authorships and generate a distribution.

H. Descriptive Statistics and Group Matching

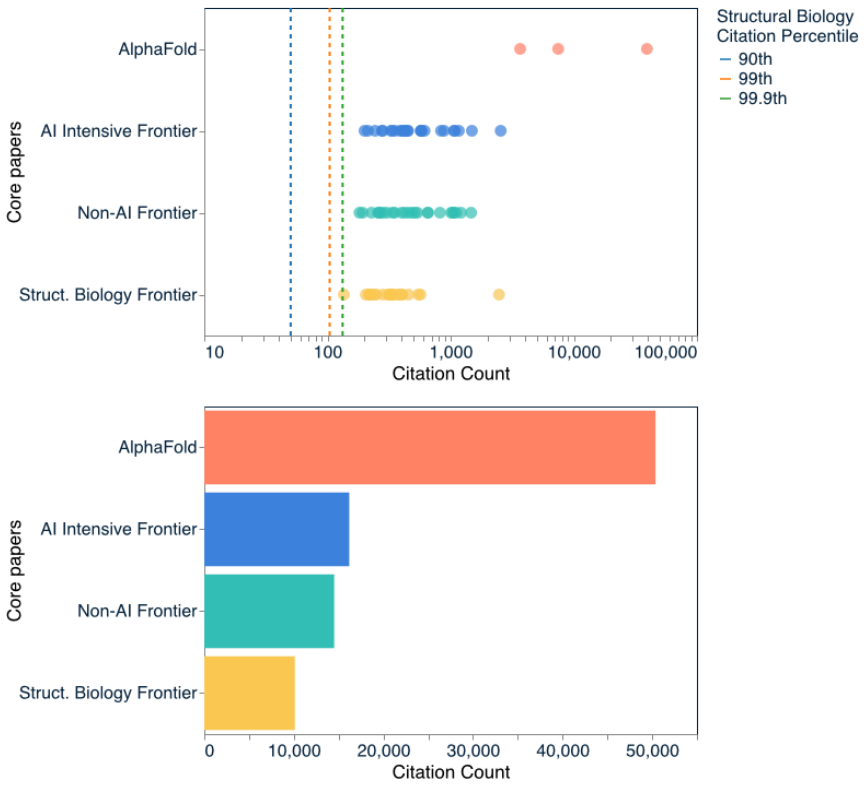

Figure A.0. Top: Distribution of citation counts for AlphaFold 2 and counterfactual core papers compared to 90th, 99th and 99.9th percentile of citations in typical structural biology papers. Bottom: Total direct citation counts for the core AlphaFold 2 papers and counterfactual sets, according to citation counts in OpenAlex. Data is not updated for 2024-2025.

Descriptive analysis

Citation Chain Papers

Source		AlphaFold		AI Frontiers		PP Frontiers		Other SB Frontiers		Other Research	
	2021Q2	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.
Publications	Pre	-	-	3,286	69,368	3,887	64,953	3,303	68,188	20,896	407,638
Publications	Post	32,726	513,838	3,264	536,890	4,223	514,040	3,247	451,042	4,441	725,157
PDB Submissions	Pre	-	-	137	603	148	1,289	103	578	1,168	5,144
PDB Submissions	Post	2,884	4,153	97	2,446	122	2,603	49	786	64	1,397
Clinical Citations	Pre	-	-	16	1,208	36	2,303	45	724	240	2,296
Clinical Citations	Post	287	12,765	19	7,061	40	9,786	94	7,352	15	11,744
Patent Citations	Pre	-	-	777	8,779	842	12,881	441	9,016	2,463	49,469
Patent Citations	Post	788	3,750	154	7,298	111	7,377	83	5,974	25	10,057

Early Career Researchers

Source		AlphaFold		AI Frontiers		PP Frontiers		Other SB Frontiers		Other Research	
	2021Q2	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.
Publications	Pre	940	9,754	303	9,628	235	5,793	114	3,902	595	32,966
Publications	Post	30,691	346,238	3,469	269,700	2,362	167,368	1,329	111,553	4,769	486,810
PDB Submissions	Pre	23	5	4	8	2	11	1	0	8	29
PDB Submissions	Post	776	657	46	193	23	175	12	39	19	110
Clinical Citations	Pre	121	657	19	1,642	4	548	3	578	26	2,428
Clinical Citations	Post	535	7,395	67	5,028	34	4,438	28	4,191	69	13,406
Patent Citations	Pre	140	551	5	538	27	217	15	412	52	1,150
Patent Citations	Post	585	2,804	117	2,318	48	1,672	21	1,079	58	4,021

Established Researchers

Source		AlphaFold		AI Frontiers		PP Frontiers		Other SB Frontiers		Other Research	
	2021Q2	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.
Publications	Pre	395,738	1,675,188	41,532	601,720	23,489	309,107	10,996	199,638	65,802	474,413
Publications	Post	465,941	2,133,312	30,480	661,473	14,799	341,592	6,573	202,403	13,778	383,308
PDB Submissions	Pre	5,587	1,627	887	636	625	481	348	240	3,661	3,457
PDB Submissions	Post	7,471	2,580	640	698	383	514	121	141	657	921
Clinical Citations	Pre	65,361	386,920	5,091	111,998	2,487	72,714	1,308	62,858	2,888	84,216
Clinical Citations	Post	19,672	127,066	1,277	32,042	674	21,323	313	16,818	332	21,231
Patent Citations	Pre	46,990	113,454	4,995	37,765	3,672	16,950	1,524	10,589	6,302	29,556
Patent Citations	Post	10,023	29,873	1,042	10,077	705	4,898	208	3,153	349	6,304

Principal Investigators' Laboratories

Source		AlphaFold		AI Frontiers		PP Frontiers		Other SB Frontiers		Other Research	
	2021Q2	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.	Adjacent	Downstr.
Publications	Pre	970,644	1,363,378	38,241	155,552	17,684	52,592	6,386	18,716	22,494	28,886
Publications	Post	826,516	1,243,989	20,200	108,738	9,418	35,910	2,873	13,115	8,700	17,398
PDB Submissions	Pre	20,352	3,136	358	121	223	39	67	16	377	26
PDB Submissions	Post	12,715	2,482	103	42	37	18	7	4	46	3
Clinical Citations	Pre	295,146	891,692	13,211	91,719	9,102	47,976	4,050	20,778	6,338	15,175
Clinical Citations	Post	49,693	176,806	1,810	17,360	988	7,164	254	2,861	673	2,116
Patent Citations	Pre	218,573	195,135	5,575	16,015	2,964	5,529	858	1,798	2,828	2,599
Patent Citations	Post	17,694	18,463	289	1,199	108	412	41	103	103	131

Appendix Table 1: Summary of total publications, PDB submissions, patents and clinical trial citations linked to each group of core publications and other structural biology papers, broken down by citation chains, ECRs, established researchers and PI labs.

Comparative overview of the sample

After applying Coarsened Exact Matching to align baseline characteristics, Table 2 compares key metrics of the four groups in our study: researchers building on AlphaFold 2 (AF), frontier AI (AI), frontier non-AI (no AI), and other structural biology researchers and labs (SB)[26]. The baseline years are 2015 through 2020 for established researchers and labs, and 2020 for early career researchers.

Early Career Researchers

	AlphaFold2		AI Frontiers		PP Frontiers		SB Frontiers		Baseline			Effect Analysis	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Δ	Cohen's d	p-value
Number of Publications	**4.95**	64.63	**2.77**	1.84	**2.81**	2.31	**2.55**	1.42	**2.74**	2.46	2.23	**0.056**	**<0.001**
Citations per Publication	**15.32**	27.92	**13.99**	35.36	**13.13**	25.22	**13.07**	30.66	**12.88**	30.42	2.05	**0.083**	**<0.001**
Log Field-Weighted Citation Index	**0.64**	0.64	**0.63**	0.64	**0.54**	0.55	**0.50**	0.60	**0.56**	0.62	0.08	**0.129**	**<0.001**
Number of Clinical Citations	**0.13**	0.77	**0.10**	0.69	**0.15**	0.90	**0.29**	1.56	**0.16**	0.99	-0.04	**-0.036**	**<0.001**
Number of Patents	**0.19**	3.16	**0.07**	0.45	**0.07**	0.48	**0.06**	0.40	**0.09**	1.00	0.12	**0.049**	**<0.001**
Number of PDB Submissions	**0.00**	0.07	**0.00**	0.00	**0.00**	0.00	**0.00**	0.00	**0.00**	0.00	0.00	**0.058**	**<0.001**
Log Max TM-Score	**0.69**	-	-	-	-	-	-	-	-	-	-	-	**<0.001**
Institution h-index	**2.60**	1.10	**2.46**	1.14	**2.48**	1.12	**2.31**	1.20	**2.51**	1.15	0.16	**0.084**	**<0.001**
Institution 2-Year Mean Citedness	**2.47**	1.10	**2.26**	1.11	**2.37**	1.11	**2.29**	1.09	**2.36**	1.15	0.15	**0.096**	**<0.001**

Established Researchers

	AlphaFold2		AI Frontiers		PP Frontiers		SB Frontiers		Baseline			Effect Analysis	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Δ	Cohen's d	p-value
Number of Publications	**2.26**	7.06	**2.18**	6.06	**2.09**	6.11	**2.10**	7.56	**2.07**	6.21	0.15	**0.028**	**<0.001**
Citations per Publication	**25.66**	85.58	**23.58**	88.64	**21.52**	69.70	**20.39**	65.34	**19.51**	56.77	4.41	**0.078**	**<0.001**
Log Field-Weighted Citation Index	**0.70**	0.70	**0.68**	0.69	**0.63**	0.66	**0.61**	0.66	**0.63**	0.67	0.06	**0.101**	**<0.001**
Number of Clinical Citations	**0.49**	7.12	**0.40**	3.75	**0.48**	3.78	**0.64**	5.28	**0.35**	2.92	0.03	**0.023**	**<0.001**
Number of Patents	**0.17**	2.56	**0.14**	2.32	**0.13**	1.93	**0.13**	1.79	**0.13**	2.77	0.04	**0.015**	**<0.001**
Number of PDB Submissions	**0.01**	0.11	**0.01**	0.10	**0.01**	0.11	**0.01**	0.10	**0.02**	0.18	0.00	**-0.114**	**<0.001**
Log Max TM-Score	**0.53**	0.20	**0.55**	0.19	**0.56**	0.19	**0.56**	0.19	**0.53**	0.20	-0.02	**-0.025**	**<0.001**
Institution h-index	**2.58**	1.10	**2.53**	1.11	**2.53**	1.12	**2.59**	1.14	**2.59**	1.13	0.02	**-0.008**	**<0.001**
Institution 2-Year Mean Citedness	**2.59**	1.11	**2.44**	1.12	**2.40**	1.12	**2.41**	1.12	**2.38**	1.12	0.19	**0.193**	**<0.001**

Principal Investigator's Laboratories

	AlphaFold2		AI Frontiers		PP Frontiers		SB Frontiers		Baseline			Effect Analysis	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Δ	Cohen's d	p-value
Number of Publications	**3.48**	9.87	**3.78**	4.91	**4.49**	13.67	**5.28**	6.47	**6.92**	9.13	-1.63	**-0.353**	**Small**
Citations per Publication	**40.72**	88.14	**38.63**	78.84	**37.30**	77.09	**38.93**	100.29	**37.22**	74.93	2.70	**0.041**	**<0.001**
Log Field-Weighted Citation Index	**1.00**	0.68	**0.98**	0.65	**0.97**	0.65	**1.01**	0.68	**1.03**	0.67	0.01	**-0.030**	**<0.001**
Number of Clinical Citations	**1.21**	7.35	**1.59**	8.34	**3.22**	15.38	**6.14**	22.86	**7.35**	29.15	-3.36	**-0.439**	**Small**
Number of Patents	**0.74**	4.66	**0.45**	2.19	**0.53**	2.74	**0.70**	3.74	**0.75**	3.77	0.13	**-0.003**	**<0.001**
Number of PDB Submissions	**0.07**	0.35	**0.01**	0.10	**0.01**	0.09	**0.01**	0.09	**0.01**	0.13	0.07	**0.194**	**<0.001**
Log Max TM-Score	**0.59**	0.17	**0.60**	0.16	**0.56**	0.18	**0.63**	0.14	**0.58**	0.17	-0.01	**0.031**	**<0.001**
Institution h-index	**2.59**	1.12	**2.39**	1.11	**2.47**	1.10	**2.59**	1.12	**2.61**	1.11	0.07	**-0.014**	**<0.001**
Institution 2-Year Mean Citedness	**2.47**	1.11	**2.44**	1.13	**2.42**	1.12	**2.43**	1.11	**2.54**	1.13	0.01	**-0.066**	**<0.001**

Appendix Table 2. Summary statistics and Group Balance for researchers building on AlphaFold 2 (AF), frontier AI (AI), frontier non-AI protein prediction (PP), other structural biology frontier (SB), and baseline researchers and labs (SB) pre-release of AlphaFold (2015-2020).

[26] If a researcher or lab cites both AlphaFold 2 and other frontier innovations, we classify them as AlphaFold 2 'adopters'. In our regressions, interactions between uses of AF and any group of frontier research developments do not produce significant results.

Established researchers building on AlphaFold 2 show slightly higher mean institutional citedness and publication counts compared to the counterfactual groups. However, other key metrics like the institutional h-index are very similar across all groups, suggesting a strong baseline comparability on observable characteristics.

A similar pattern is observed for Early Career Researchers. While some minor differences may be statistically significant due to the large sample size (e.g., for institutional h-index), the corresponding effect sizes (Cohen's d) are negligible. This indicates that the matching process has effectively balanced the pre-treatment characteristics across the ECR groups.

At the laboratory level, labs that adopt AlphaFold 2 report moderately higher patenting rates than the counterfactual groups. In contrast, other key metrics, including the number of PDB submissions, are now highly comparable across all lab types. Any minor differences, primarily in patent activity for labs, are accounted for in our model, providing a solid foundation to study how AlphaFold 2's adoption compares with frontier research.

I. Trend, Regression, and Event Study Results

Pre-treatment trends and the common trends assumption

A core assumption of the difference-in-differences (DiD) framework is that treatment and control groups would have followed parallel trends in the outcome variables had the treatment not occurred. While our Coarsened Exact Matching (CEM) procedure ensures the groups are balanced on observable characteristics at baseline , the validity of our DiD estimates also rests on this "common trends" assumption.

To empirically assess this, we plot the annual trends for key outcome variables across our analytical groups for both the pre-treatment period (2015-2020) and the post-treatment period (2021-2024). Appendix Figures A.1 and A.2 below illustrate these trends for established researchers and laboratories, respectively. Visual inspection confirms that for most outcomest, the trend lines for AlphaFold 2 adopters, the three counterfactual frontier groups, and the wider structural biology baseline are largely comparable prior to 2021. While minor year-to-year fluctuations are present, particularly in publication volume, no group exhibits a systematically different pre-treatment trajectory for our core outcomes.

Furthermore, we acknowledge that adoption of AlphaFold 2 is not randomly assigned, and researchers who use it may systematically differ from non-adopters in unobservable ways. Our empirical strategy uses a multi-pronged approach to mitigate this risk. In addition to balancing on observables with CEM and validating pre-treatment trends, our regression models include strict researcher or laboratory fixed effects. This controls for any time-invariant unobserved heterogeneity (ie. innate ability or persistent funding levels) that might otherwise confound the results. A placebo test, which displaces the treatment date three years earlier within the pre-treatment period (ending 2021 Q2), confirms the absence of significant pre-trends, save for moderate effects on publication volume.

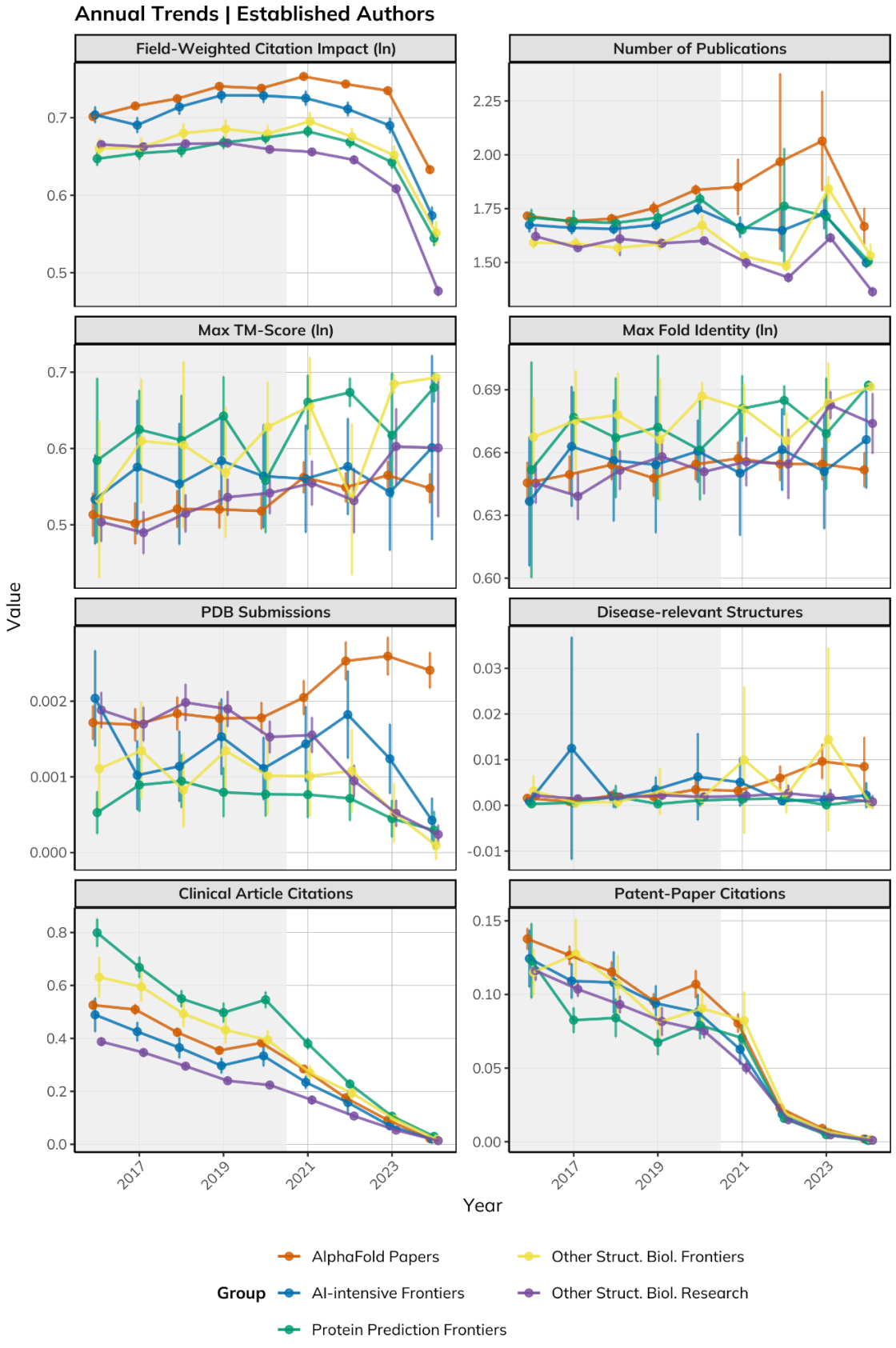

Figure A.1: Annual trends in key outcome variables (2015-2024) across all analytical groups in our researcher data, illustrating parallel pre-treatment trends and post-treatment divergence.

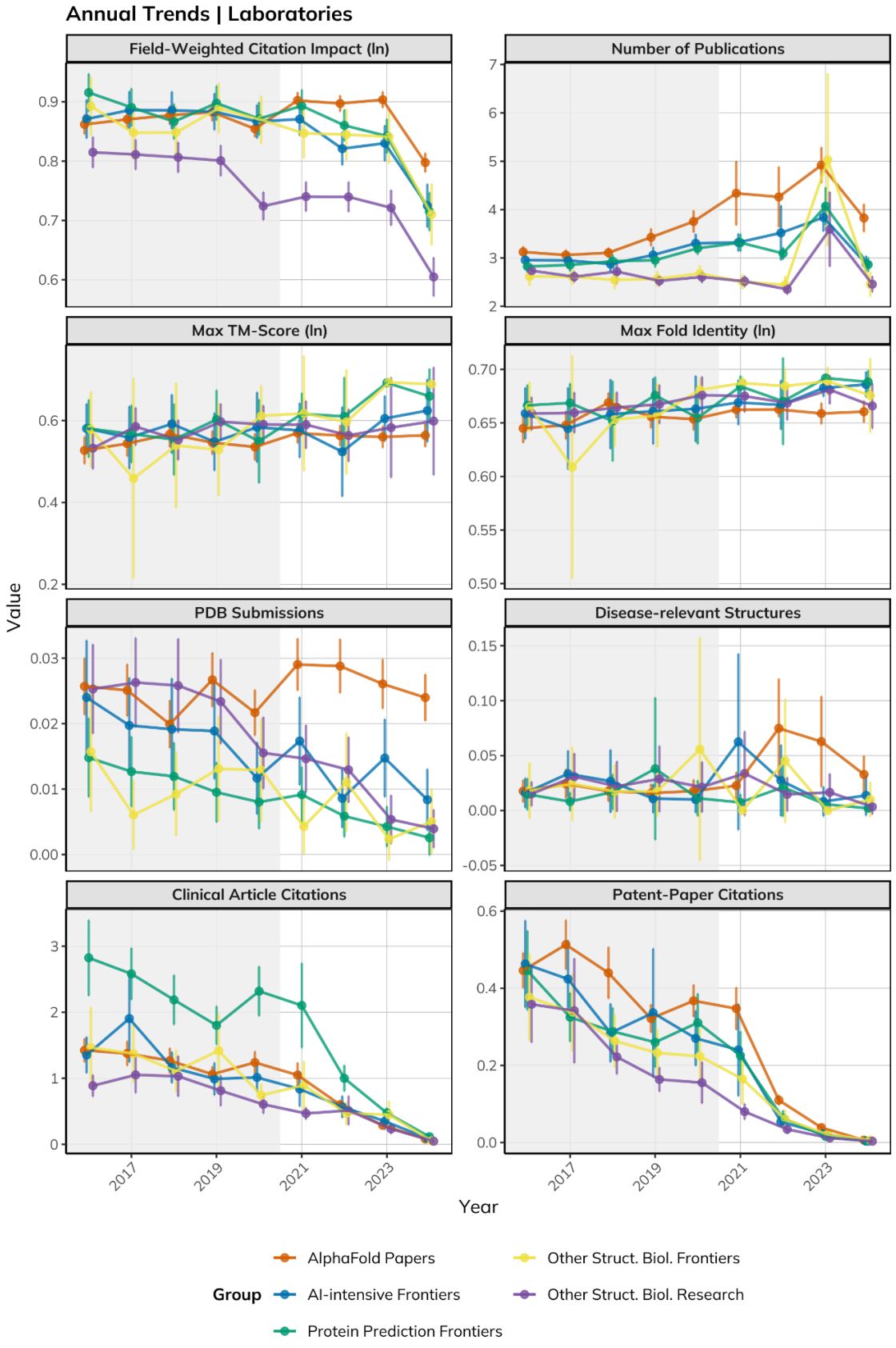

Figure A.2: Annual trends in key outcome variables (2015-2024) across all analytical groups in our laboratories data, illustrating parallel pre-treatment trends and post-treatment divergence.

Summary of regression results across observation groups

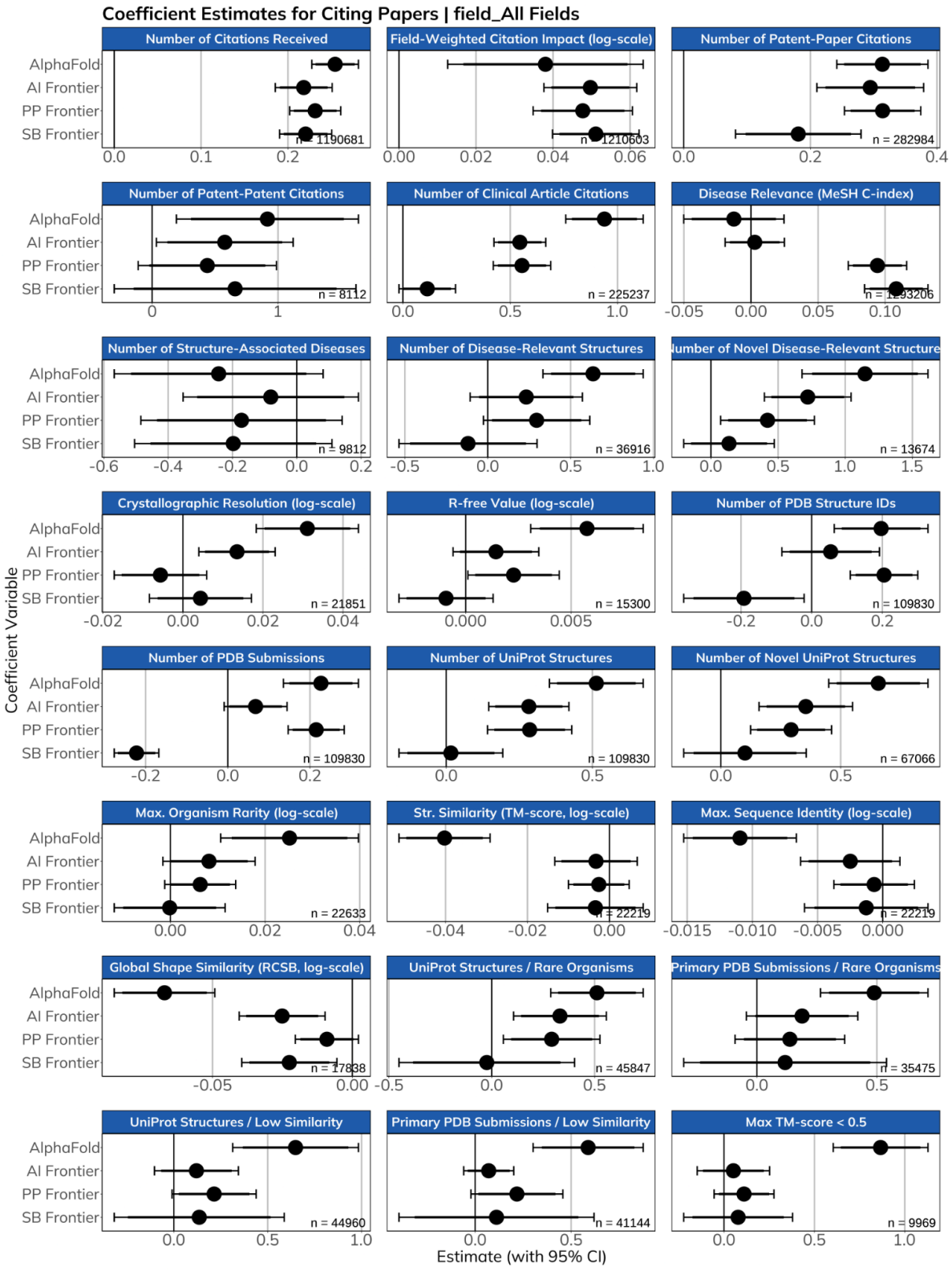

Figure A.3: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

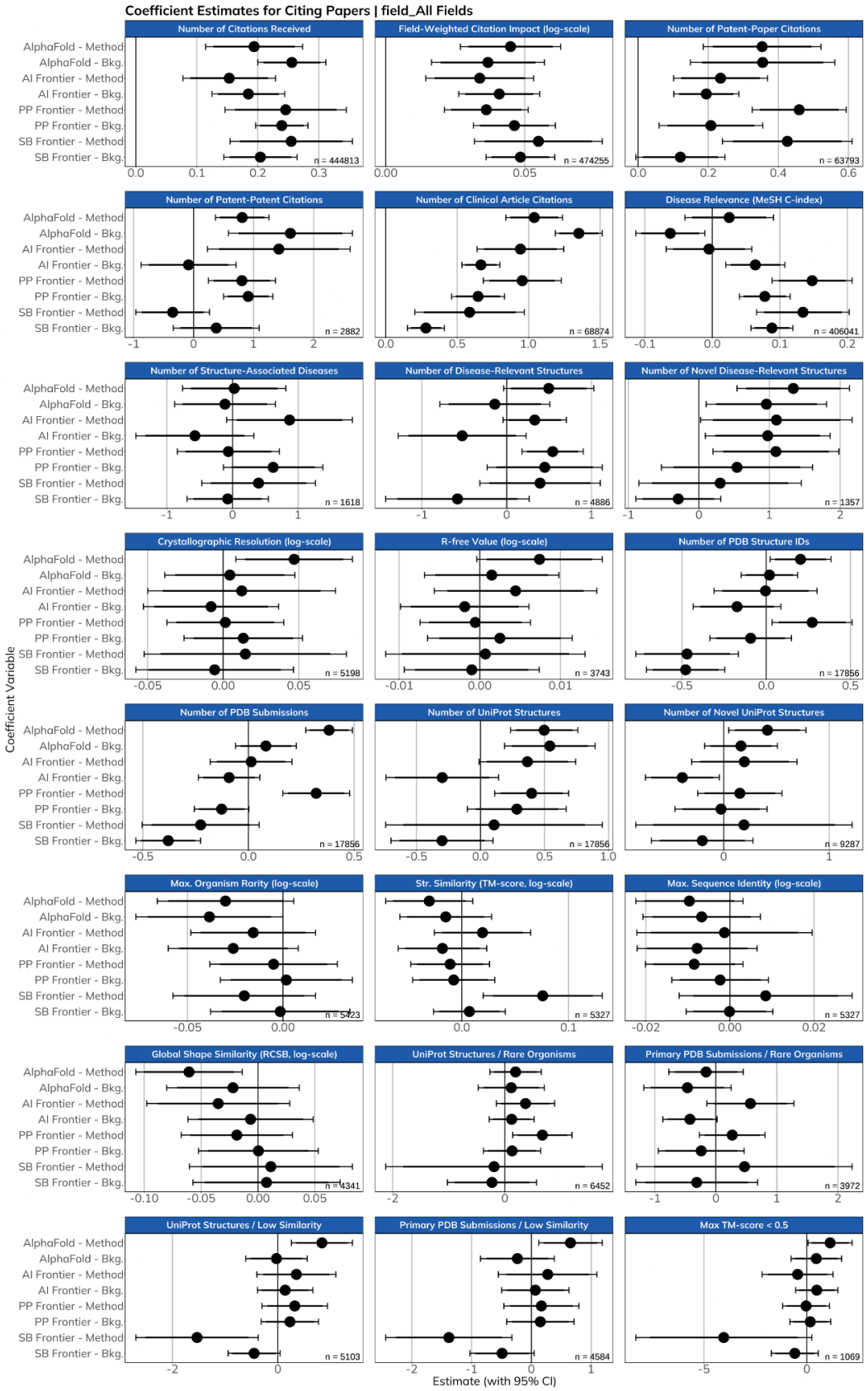

Figure A.4: Coefficient estimates from Poisson (if discrete counts) and linear (otherwise) regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals.

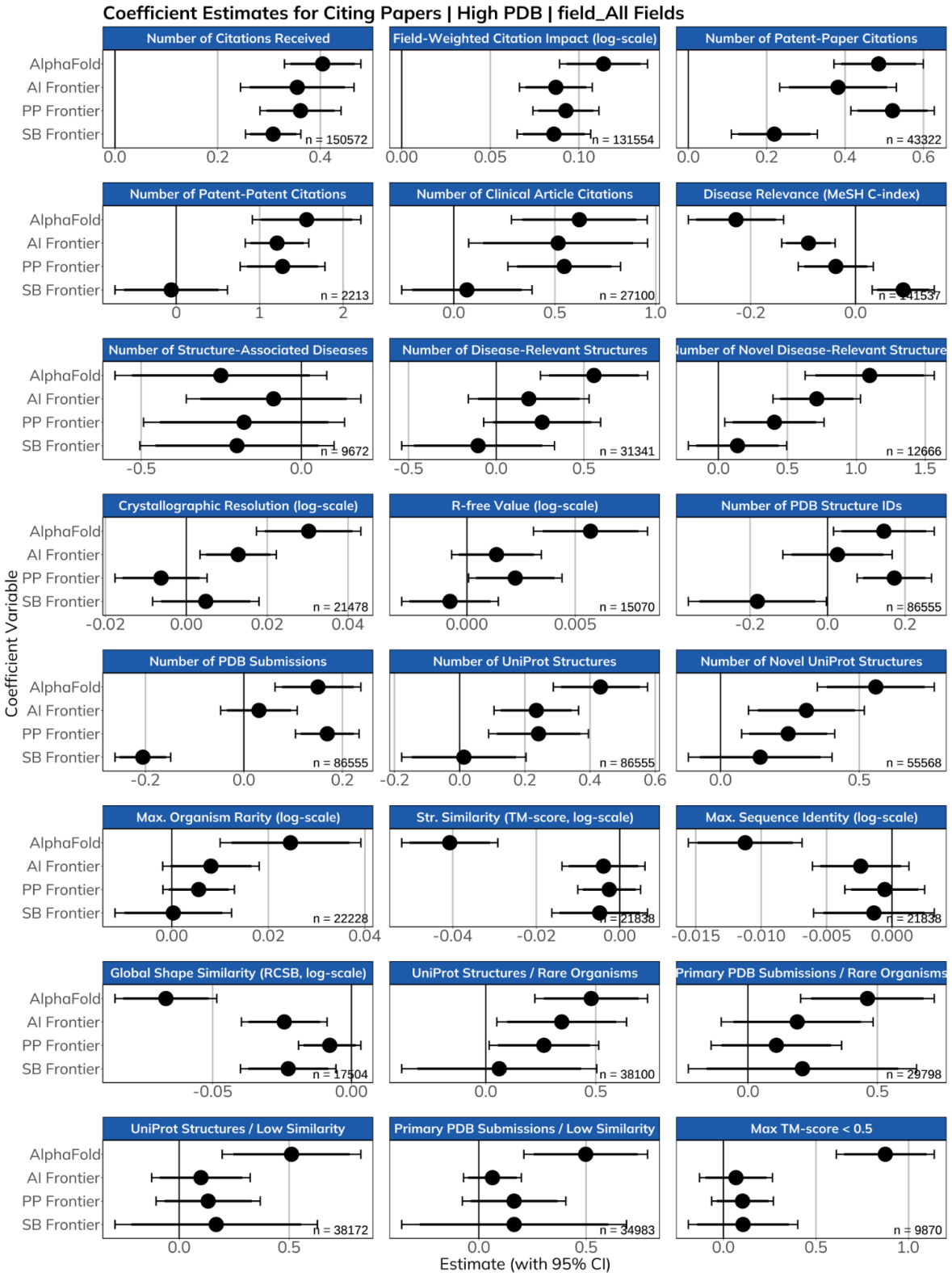

Figure A.5: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

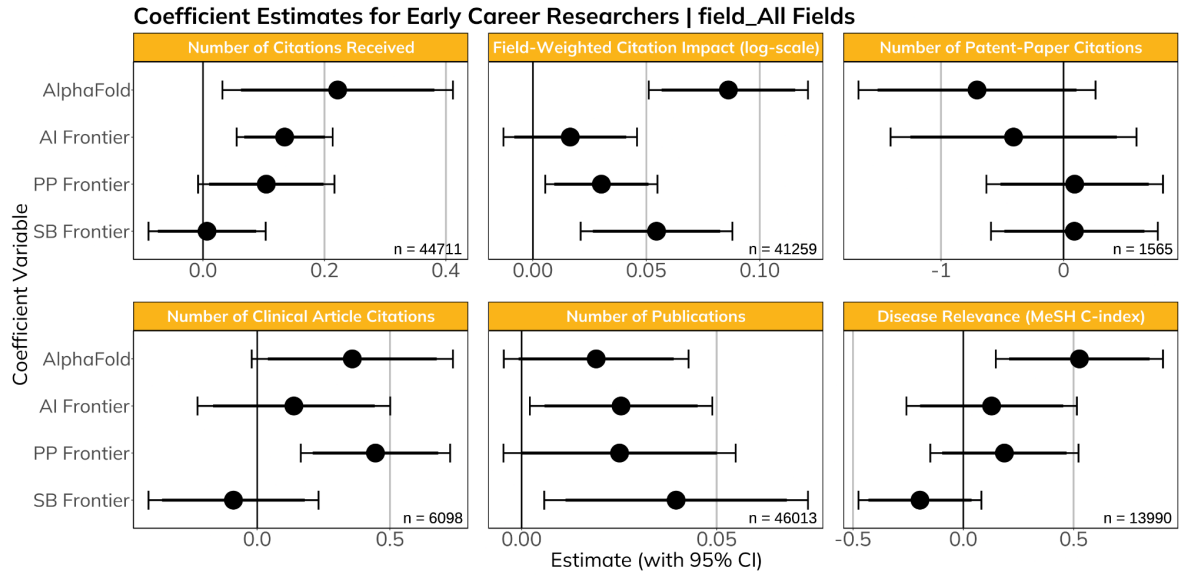

Figure A.6: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

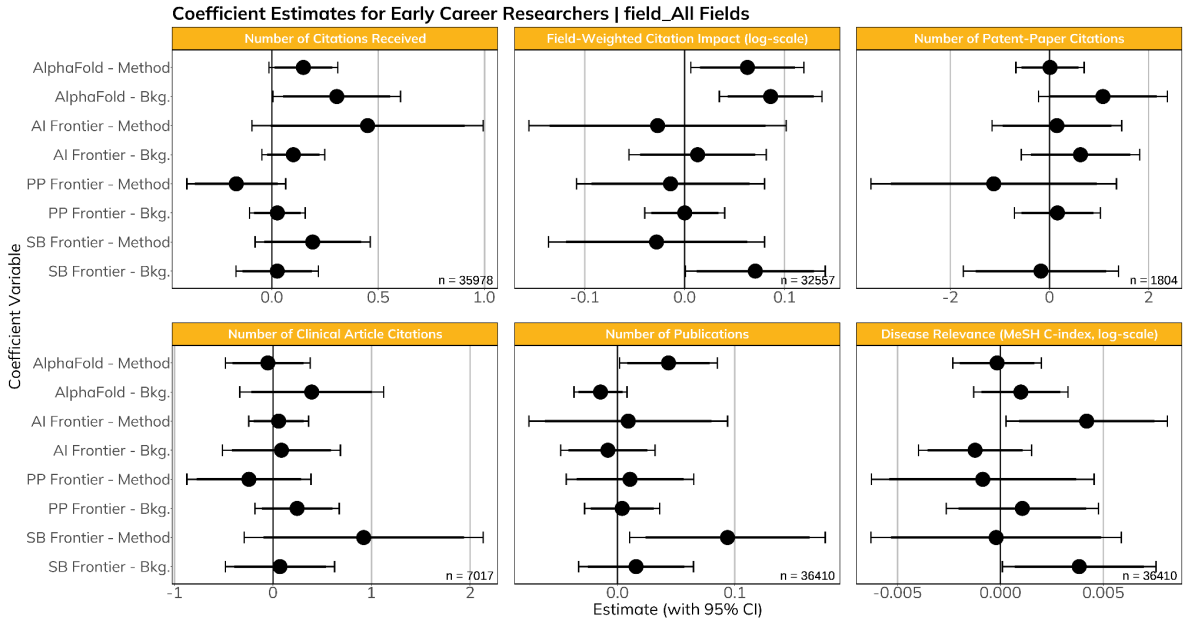

Figure A.7: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

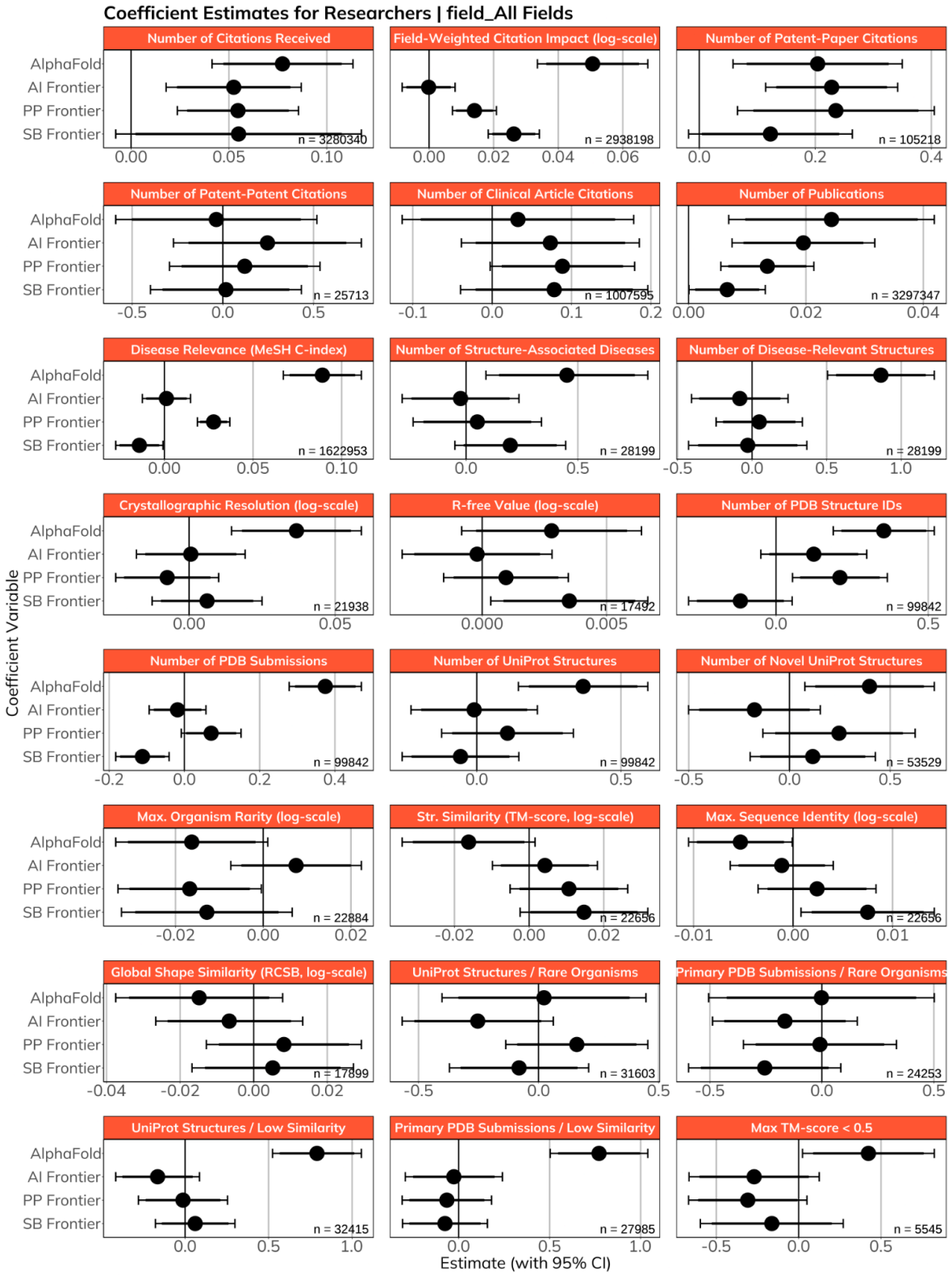

Figure A.8: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

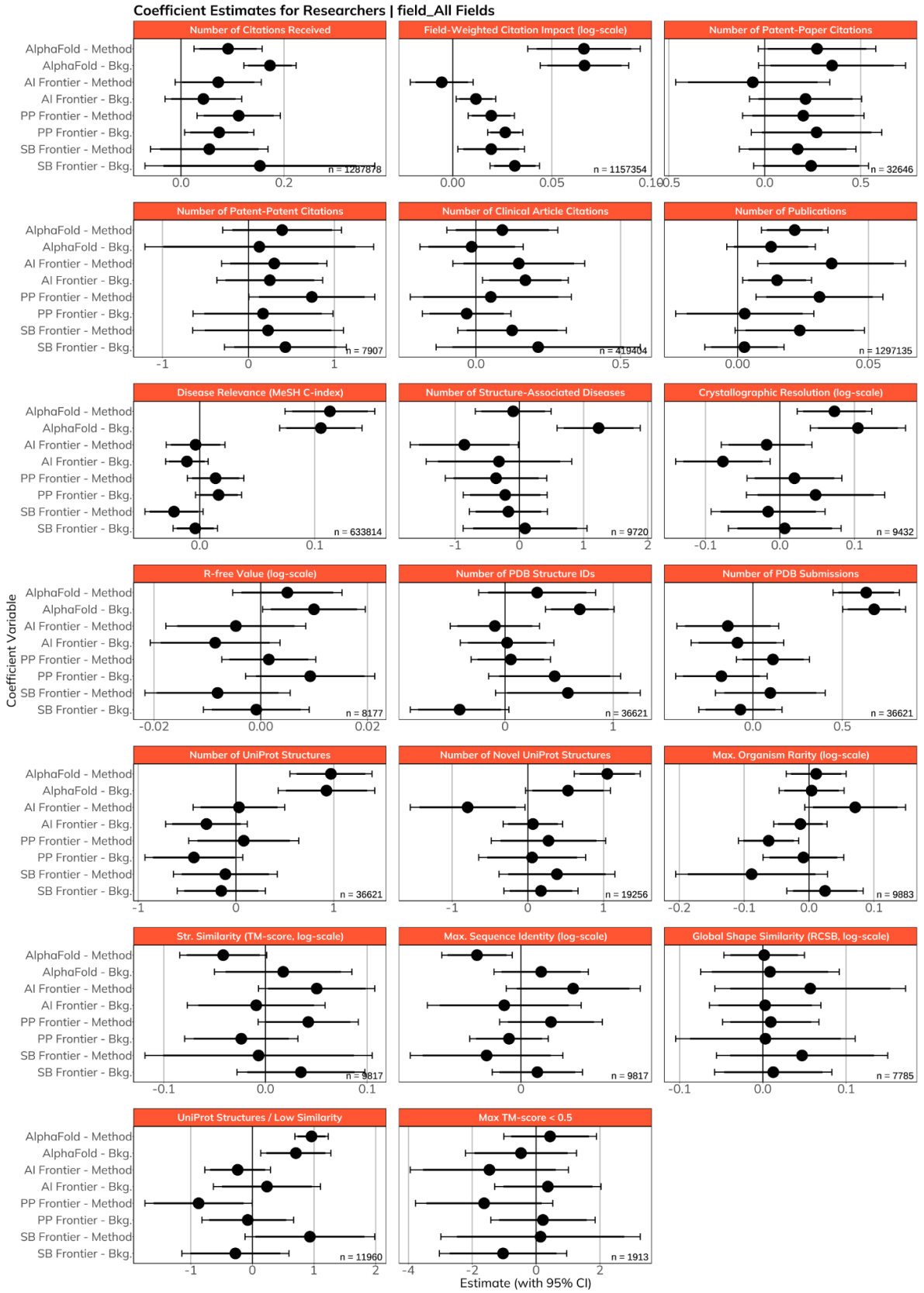

Figure A.9: Coefficient estimates from Poisson (if discrete counts) and linear (otherwise) regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals.

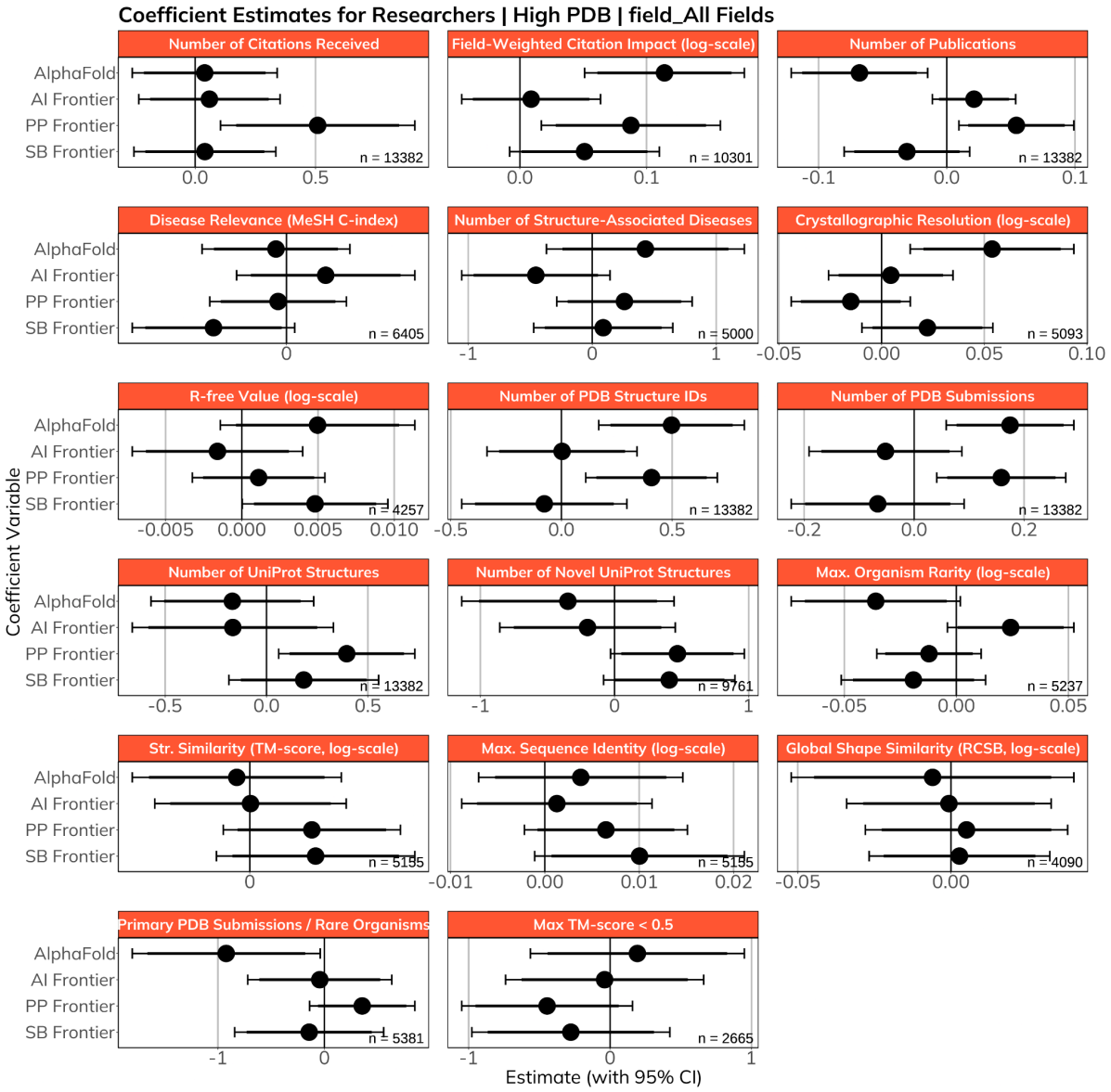

Figure A.10: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

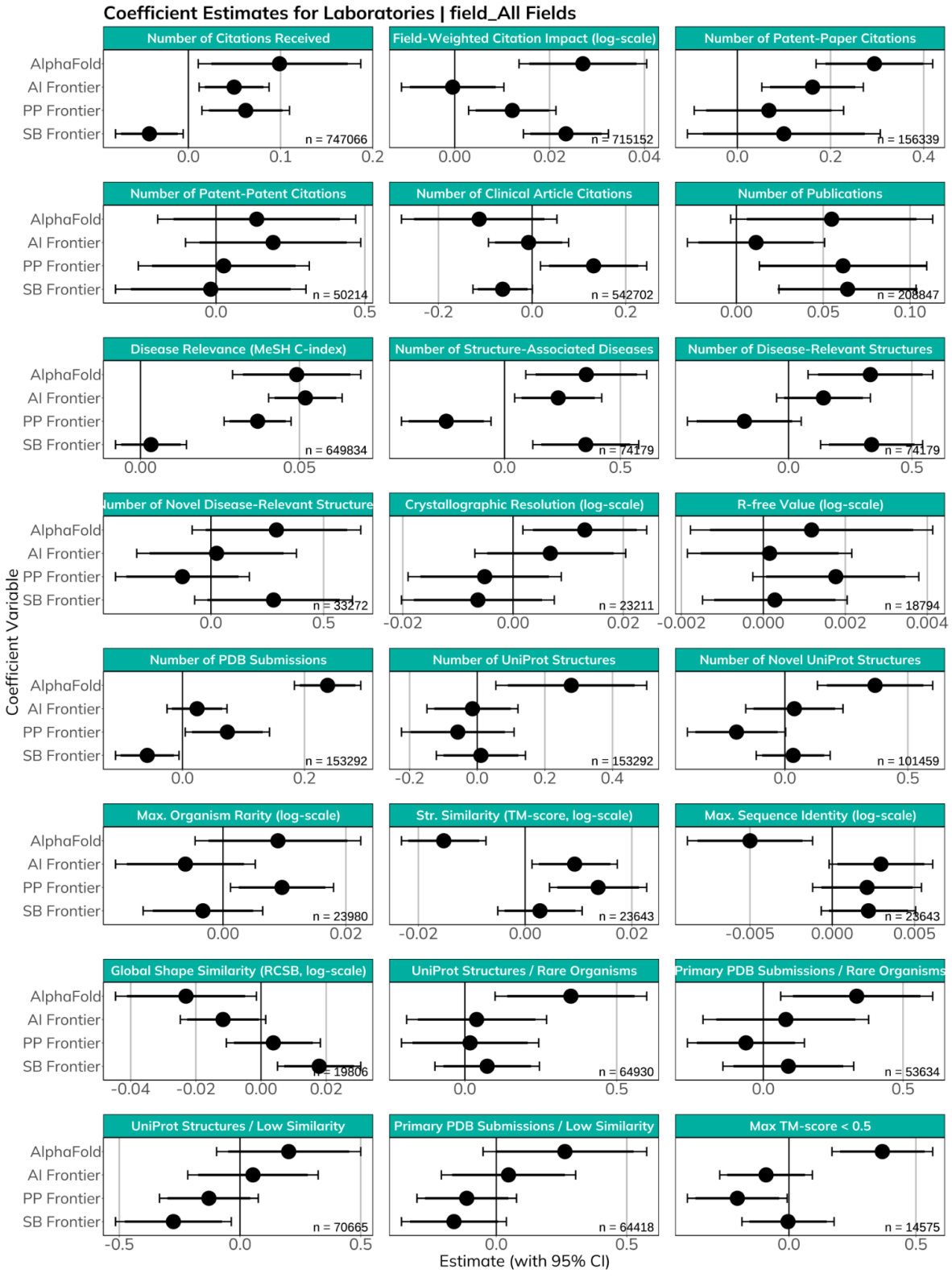

Figure A.11: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

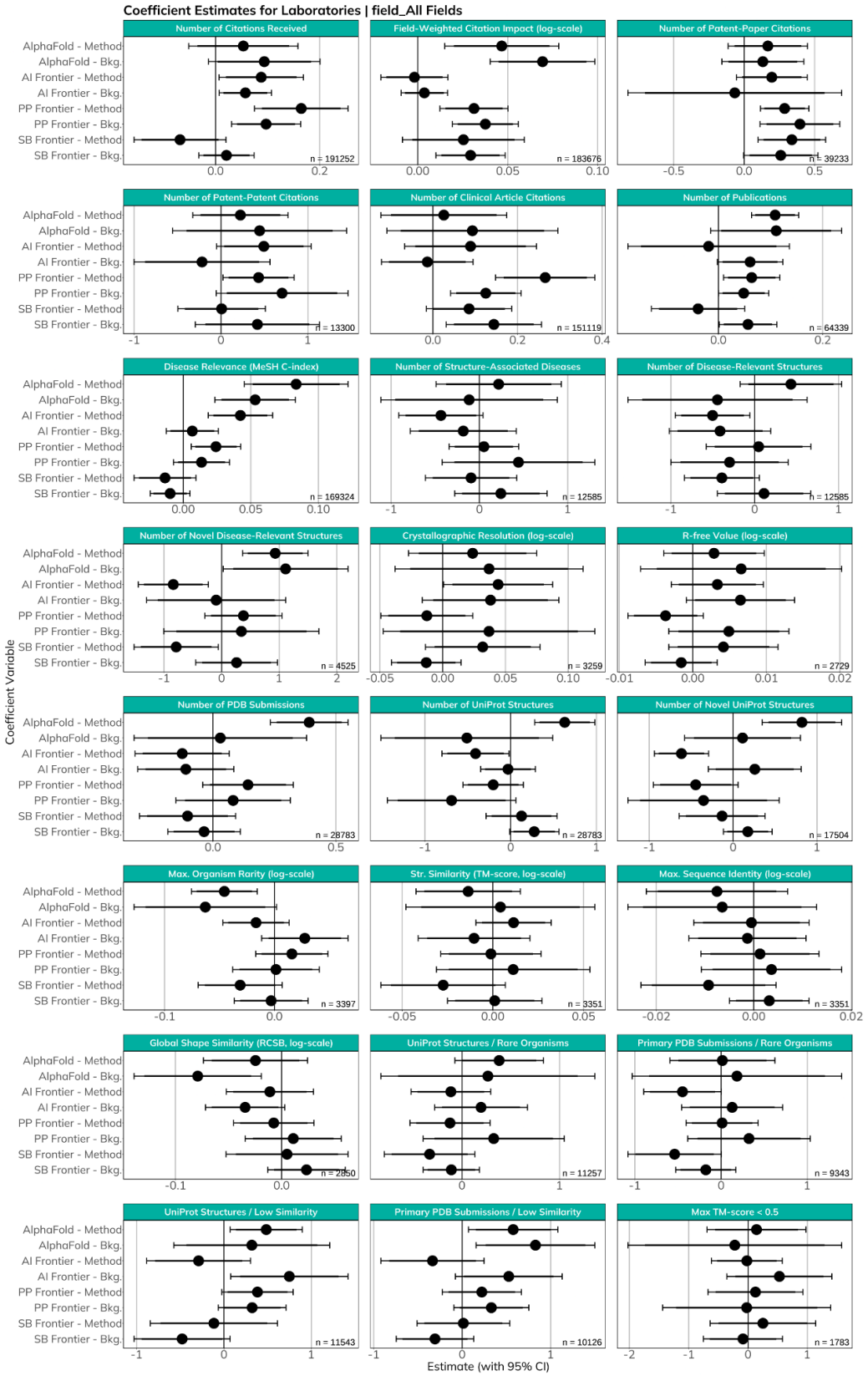

Figure A.12: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

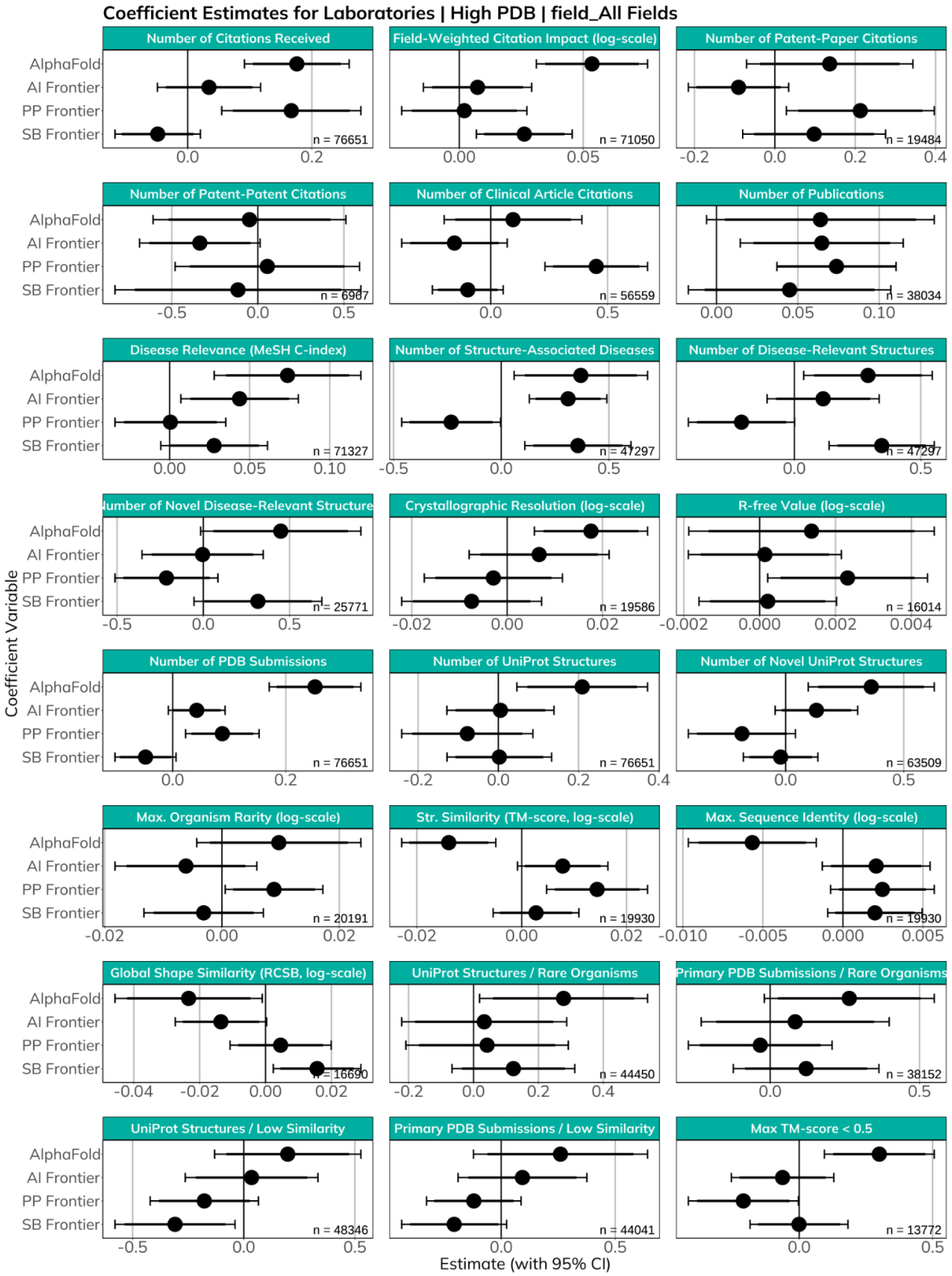

Figure A.13: Coefficient estimates from Poisson and linear regressions showing associations between frontier developments and various research outcomes. Points represent main effects with 95% confidence intervals. Each row shows a different dependent variable, with frontier effects estimated separately for each outcome.

Event study specification

To understand the dynamic effects of AlphaFold 2 adoption, we consider event study design approaches for both individual researchers and laboratories. It is important to note a key limitation upfront: citation-based metrics, including clinical and patent citations, inherently have a downward bias over shorter time horizons due to the natural lag in the publication and citation process. While the Callaway and Sant'Anna (2021) estimator is designed to be robust, results beyond one year should be interpreted with this in mind.

Our analysis at the individual researcher level benefits from a large sample size and varied control groups, providing greater statistical power. Main results are available in Figure A.14. We observe a positive, though initially modest, impact on the field-weighted citation index within the first nine months of an author adopting AlphaFold 2. The effect remains positive and gains significance over time, which is noteworthy given that the Callaway and Sant'Anna methodology produces larger standard errors (Gardner et al 2024).

In terms of research focus, we observe a nearly 10% decrease in the submission of proteins with high sequence identity to existing structures during the first year of adoption, with a similar, though smaller, effect observed for proteins within an author's established field identity. This is followed by a clear increase in total PDB submissions beginning approximately two years after adoption.

Regarding translational impact, clinical article citations increase by around 5% during the first year of adoption, with the effect becoming negligible after 18 months. Conversely, patent citations appear to decrease starting one year post-adoption, although both of these findings may be influenced by the aforementioned citation lag dynamics.

The analysis at the laboratory level, available in Figure A.15, is more challenging due to the smaller sample of matched labs, particularly as many labs in the dataset eventually adopt AlphaFold 2. This results in larger confidence intervals and a general scarcity of statistically significant effects.

Nonetheless, the direction of most effects (positive or negative) aligns with the findings at the researcher level. We find statistically significant effects for structural novelty (TM-scores), PDB submissions, and the submission of disease-related protein structures. Downward dynamics are observed for citation variables that are not field-weighted, though we caution against overinterpreting these results due to the sample limitations and potential confounding factors at the lab level.

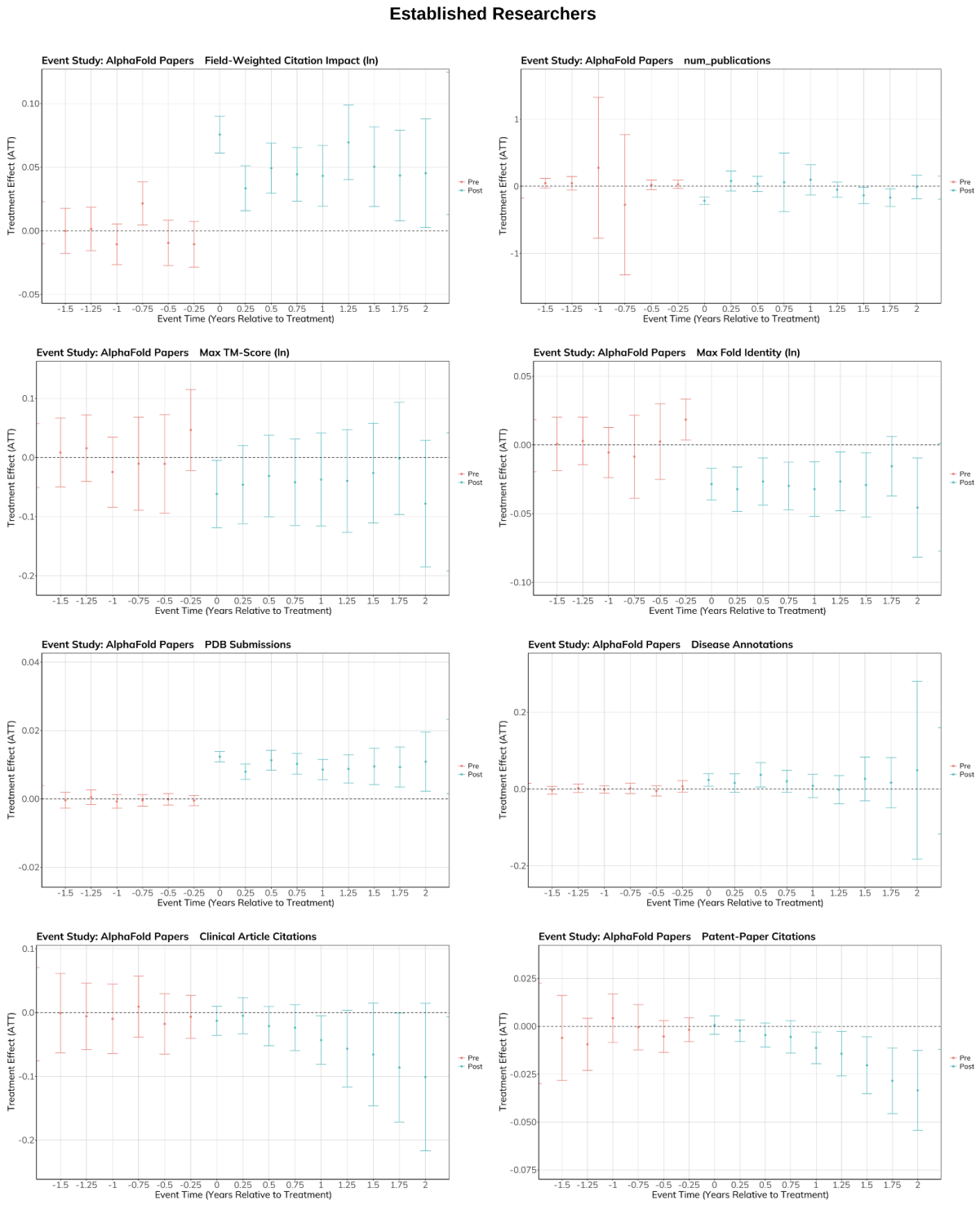

Figure A.14: Event study plots showing the dynamic effects of AlphaFold 2 adoption on primary outcome variables for established researchers. The points represent the estimated average treatment effect on the treated (ATT) for each quarter relative to the adoption event (t=0), from 8 quarters prior to 12 quarters after adoption. The reference period is implicitly embedded in the Callaway and Sant'Anna (2021) robust estimation procedure through double-differencing. Vertical lines indicate 95% confidence intervals. The pre-treatment coefficients (t<0) serve as a test of the parallel trends assumption and should be statistically insignificant if the assumption holds.

Laboratories

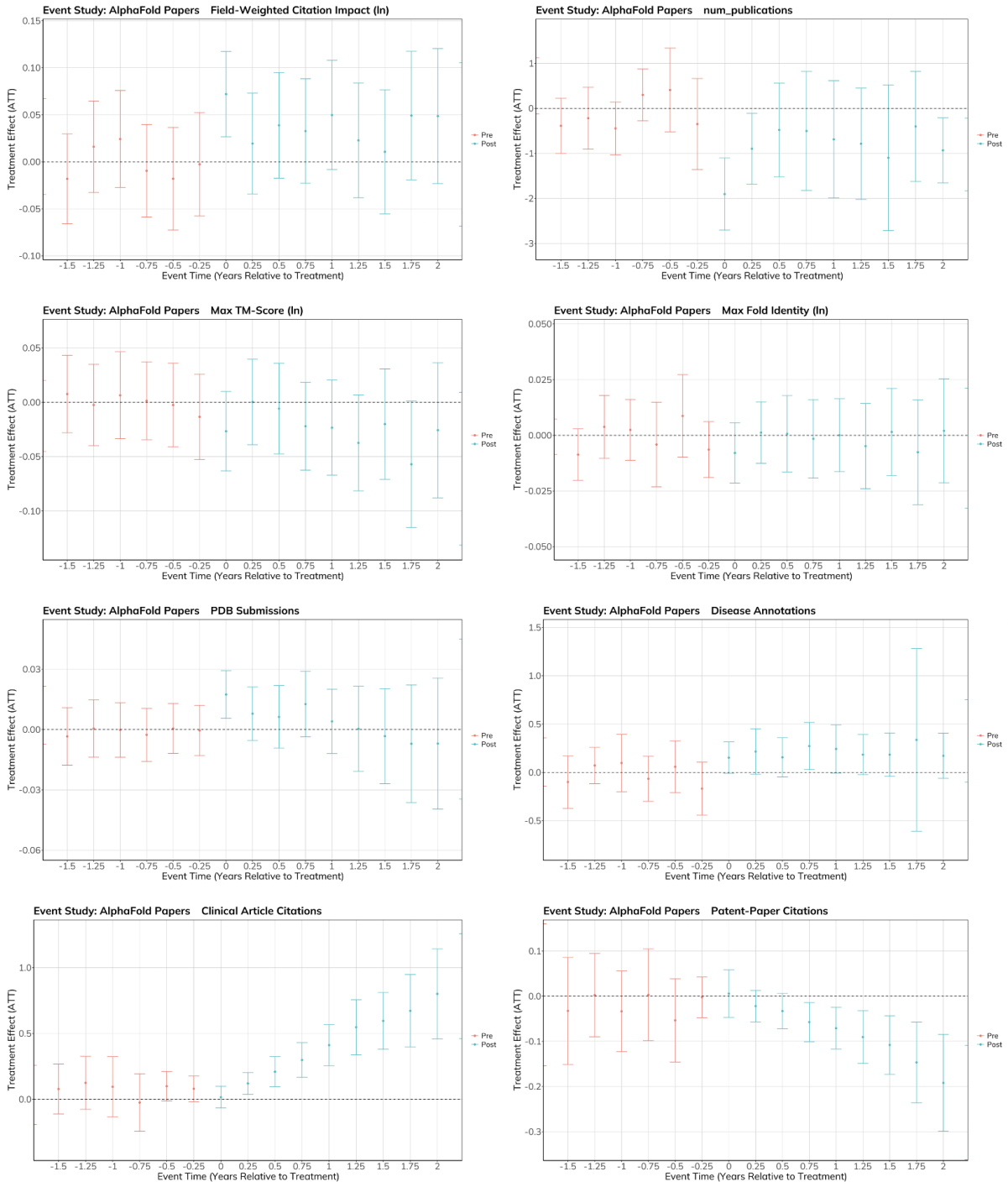

Figure A.15: Event study plots showing the dynamic effects of AlphaFold 2 adoption on primary outcome variables for principal investigators' laboratories. The points represent the estimated average treatment effect on the treated (ATT) for each quarter relative to the adoption event (t=0), from 8 quarters prior to 12 quarters after adoption. The reference period is implicitly embedded in the Callaway and Sant'Anna (2021) robust estimation procedure through double-differencing. Vertical lines indicate 95% confidence intervals. The pre-treatment coefficients (t<0) serve as a test of the parallel trends assumption and should be statistically insignificant if the assumption holds.

About the Innovation Growth Lab

The Innovation Growth Lab (IGL) is a global policy lab that helps governments develop more effective policies to increase innovation and productivity. IGL's mission is to foster productive, sustainable, and inclusive economies through novel policy ideas, experimentation, data and evidence. IGL works with policymakers, researchers, practitioners and funders to address key policy challenges in the fields of science, innovation, entrepreneurship and business policies.

IGL's Data and Technology Unit supports the organisation's mission through the development of original data driven insights, the development of tools for analysis and decision making, and supporting policymakers and practitioners to build capabilities around the use of data.

To find out more about our work, visit innovationgrowthlab.org