How AI Impacts the Quest for Knowledge: Evidence from AlphaFold2

Myra Mohnen and Joshua S. Gans*

15 June 2026

Abstract

We study how artificial intelligence changes the direction of scientific research. Our setting is AlphaFold2 (AF2), which predicts protein structures. We develop a model in which AI can both replace some experiments and make other experiments easier. Using 530,495 proteins linked to the Protein Data Bank and differences in the quality of AF2 predictions after its release, we find that AF2 increased experimental activity but redirected it towards proteins closer to existing knowledge. For proteins fully covered by usable AF2 predictions, total experimental deposits rose by 41% and follow-up deposits on previously studied proteins rose by 64%. By contrast, first depositions for previously unsolved proteins did not rise, and the entry that occurred concentrated near already-solved proteins. AF2 also changed the kinds of data produced: structures showing new conformations and interactions with other molecules became more common, while structures extending coverage to new protein families and novel folds became less common. Thus, AI can deepen knowledge while slowing its expansion into less-explored areas. This matters because today's experiments provide training data for tomorrow's AI systems. More experimental data need not mean broader experimental data.

Keywords: artificial intelligence; scientific experimentation; AlphaFold2; knowledge frontier; training data.
JEL codes: D62, O31, O32, O33.

*University of Ottawa (Mohnen) and University of Toronto and NBER (Gans). We thank David Autor, Pierre Azoulay, Jason Garred, Scott Stern, and seminar participants at the universities of Luxembourg, Carlos III de Madrid, Toronto, Utrecht, and the NBER productivity lunch for helpful comments, the SSHRC for funding, and ChatGPT 5.5 Pro, Claude Opus 4.8, and Claude Fable 5 for research assistance. We are especially grateful to Dustin Britton and Lindsey Guan (Keating Lab, MIT), Andrew Savinov (Gene-Wei Li Lab, MIT), Nasim Abdollahi (SDL3 & SDL5, Acceleration Consortium, University of Toronto), Aled Edwards (Structural Genomics Consortium, University of Toronto), Juan Mateos-Garcias (Nesta and Google DeepMind), and Yossi Matias (Google) for their insights into experimental and computational practices. Responsibility for all errors remains our own.

1 Introduction

A new artificial intelligence can take science in either of two directions. It can propose what no human would have considered, opening unexplored ground; or, because it learns from where research has already concentrated, it can predict best in well-charted regions and draw effort inward, towards what is already close to settled. Which tendency dominates is an empirical question that determines whether AI broadens or narrows the frontier of science.

The direction of the response matters because the findings from today's research become the data that trains the AI systems of tomorrow. In purely machine-learning settings, models recursively trained on earlier models' outputs can exhibit distributional collapse: tails truncate, rare events disappear, and variance shrinks across generations (Shumailov et al., 2024; Alemohammad et al., 2024). In science, the same loop runs through human choices rather than synthetic data: existing knowledge trains an AI tool, the tool changes which problems scientists judge worth pursuing, and the new findings they generate enter the record from which later systems learn. A tool that accelerates discovery today can therefore enrich the data available tomorrow while still narrowing the regions in which new knowledge is created.[1]

Our setting is the public release of AlphaFold2 (AF2), a deep-learning system that predicts protein structures with near-experimental accuracy (Jumper et al., 2021). In structural biology, each protein poses an open problem: determining the three-dimensional structure into which its amino-acid chain folds, which governs the protein's function and underpins drug discovery and disease biology. AF2 learned this mapping from experimentally determined structures, evolutionary sequence data, and its own high-confidence predictions on sequences with no experimental structure. Experimental structures remain scarce relative to the universe of known proteins (Tunyasuvunakool et al., 2021): fewer than 5% of the proteins we study had any experimental structure before AF2. By 2022, the AlphaFold Protein Structure Database held over 200 million predicted structures, supplying a first structural model even for proteins that had never been experimentally solved.[2] The result is a broad but heterogeneous information endowment — confidence varies sharply across proteins and across regions within a protein — and that variation lets us study how an AI tool changes the direction of research.

[1] Writing about successor systems, Abramson et al. (2024) anticipate that experimental advances will "provide a wealth of new training data," and that parallel advances in experimental and computational methods will "propel us further into an era" of structurally informed biology.

[2] Announcing the 2022 expansion, Demis Hassabis stated that the database could "open up completely new avenues of scientific discovery" (EMBL-EBI, 2022).

We construct a dataset linking approximately 530,000 proteins to their experimental structures in the Protein Data Bank (PDB), the public archive where laboratories deposit solved structures. We call each deposited structure a deposition, the basic measure of structural-biology output, and classify each by its scientific content and by margin. For each protein, we measure AF coverage: the share of its amino-acid residues that AF2 predicts at high confidence, a continuous measure of how much usable predicted structure the tool supplies for the protein.

Two margins determine whether this activity pushes the scientific frontier or deepens the interior. The first is entry: a first deposition records the first experimental structure of a protein new to the PDB, whereas a follow-up deposition adds to a protein already in the archive. The second is location in the structural knowledge landscape, the protein's distance — along a tree of sequence similarity — to the nearest protein with a pre-2017 experimental structure: a first deposition far from any solved relative pushes the structural map outward (frontier expansion), while one in the well-mapped interior fills it in (deepening), as follow-up depositions do by construction. A first deposition is thus necessary but not sufficient for frontier expansion.

We develop a model to organise these margins. Proteins lie on a sequence-similarity tree, where a new solved structure creates local informational spillovers. Laboratories face protein-specific setup costs, and AF2 arrives as a public signal about the dominant structural scaffold. The model isolates two opposing forces. *Information substitution* reduces the value of experiments that recover scaffold information AF2 already supplies. *Experimental complementarity* lowers the cost of experiments that use the predicted scaffold as an input but produce knowledge AF2 cannot provide, such as alternative conformations (other functional shapes the same protein can adopt) and ligand-bound structures (the protein bound to a ligand — a drug, substrate, ion, or partner molecule). The balance between these forces depends on the experiment and on its location in the knowledge landscape. The induced choices then change the composition of the experimental corpus available to later prediction systems. We show that this creates a directional externality: laboratories need not capture the value that a remote experimental anchor contributes to the ground-truth corpus that successor systems inherit.

Identification rests on a continuous-intensity difference-in-differences: we interact AF coverage with a post-release indicator and absorb protein and year fixed effects. The concern is that coverage is not randomly assigned — it reflects evolutionary depth, template availability, and accumulated knowledge that may shape research trajectories independently of AF2 — so we let these predetermined pre-AF characteristics trend differentially after the release, identifying the effect from within-protein changes net of those correlates

rather than from fixed differences across proteins.

Our first finding is that AF2 redirects experiments towards deepening rather than frontier expansion. For a fully covered protein, AF2 raises the total deposition rate by 41% and the follow-up deposition rate on already-mapped proteins by 64%. First deposition on previously unsolved proteins, by contrast, does not increase, and the remaining entry moves inwards: it is more likely near pre-2017-solved homologues and less likely at greater frontier distances.

Our second finding identifies the composition of this deepening. Mechanistic depositions, including conformational variants and ligand complexes, expand sharply, while scaffold follow-up depositions, which extend or refine the static structural scaffold, rise more modestly and less precisely. Cryo-electron microscopy expands significantly, whereas the X-ray response is statistically indistinguishable from zero. Follow-up activity increases both within laboratories' existing pipelines and for proteins new to the depositing laboratory. These patterns are consistent with AF2 lowering the cost of richer experiments, not only the cost of obtaining an initial scaffold.

Our third finding is that the experimental data feed into later AI is selective. New conformations and biologically relevant ligand structures expand. At the same time, first-in-cluster structures and structurally novel folds contract. AF2, therefore, enriches the experimental corpus along mechanistic dimensions while reducing observations that broaden sequence-family and fold coverage. More experimental data need not mean broader experimental data.

This compositional shift has a directional welfare reading. Mechanistic deepening is valuable in its own right — successor systems need precisely the conformational and ligand data that AF2 induces laboratories to produce. But the externality is directional rather than scalar: by underweighting the remote first-ever structures that map new territory, AI can erode private incentives to broaden the ground truth later systems inherit.

This paper contributes to four strands of the literature. First, we build on theories of search and discovery. Carnehl and Schneider (2025) study costly search on a Brownian knowledge path and show why patient institutions may value moonshots with low short-run payoff but substantial future spillovers. Gans (2026) distinguishes cost reductions within established domains from increases in the downstream use value of discoveries within an AI tool's range. Related work studies general-equilibrium growth and distortions in research direction (Gans, 2025; Gans and Goldfarb, 2026; Bryan and Lemus, 2017; Bryan et al., 2022; Hopenhayn and Squintani, 2021). We adapt these ideas to structural biology by separating information substitution, experimental complementarity, and the composition of the experimental data supplied to later systems. The model identifies a directional

data-supply externality: the experiments made privately attractive by the current tool need not include the remote experimental anchors with the greatest range-expansion value for later systems.

Second, we contribute to a rapidly growing empirical literature on AF2. Several contemporaneous papers use related protein-structure and publication data to study how AF2 changes scientific activity (Hill and Stein, 2026; Qian, 2026; Sun et al., 2026).[3] Rather than treating experimental output as a single margin, we show that AF2 redirects its composition. First deposition does not rise and shifts towards proteins near existing structural knowledge, while mechanistic follow-up work expands. The resulting experimental corpus is richer in conformational and ligand information but thinner in first-in-cluster structures and novel folds. This decomposition connects the immediate impact of AF2 on experimental choice to its longer-run implications for the development of predictive systems.

Third, we engage the literature on data-driven exploration and extrapolation technologies. Kim (2025) study the release of Phaser, a crystallographic method that requires a prior solved homologue, and partition sequence space into regions where molecular replacement is and is not usable. Our frontier-distance measures are a continuous-tree analogue applied to a tool that can generate a prediction for essentially every protein. We find that broad applicability does not eliminate inherited data geography: AF2 continues to redirect effort towards regions connected to existing structural knowledge. This result relates to work on scientific search, streetlight effects, and technologies that either reinforce or weaken path dependence (Azoulay et al., 2026; Hoelzemann et al., 2025; Tranchero, 2025; Aldighieri and Malpassi, 2025).

Fourth, we add to the literature on growth through recombination, in which the pace and direction of discovery depend on how densely the knowledge space is populated

[3]These papers are closely related and complementary. Hill and Stein (2026) emphasise publication responses; Qian (2026) studies changes in human-produced structural knowledge, including heterogeneity across experimental techniques; and Sun et al. (2026) studies the expansion of scientific frontiers. Relative to this work, our analysis adds three elements. First, we combine a protein-year panel with deposition-level measures of scientific content. This higher level of disaggregation allows us to distinguish first deposition from follow-up work, locate new entries relative to the inherited knowledge frontier, and characterise the induced deposits by mechanistic content, experimental method, and laboratory pipeline. Second, we interpret these margins through a formal model of experimental choice on a sequence-similarity tree. The model identifies the trade-off between *information substitution*, which reduces the value of experiments whose scaffold information AF2 already supplies, and *experimental complementarity*, which raises the return to experiments that use a predicted scaffold as an input. Third, we study a dynamic implication that is distinct from the immediate effect of AI on human research: the induced reallocation of experiments changes the composition of the experimental corpus available to train and improve successor AI systems. Thus, our object is not only whether an AI tool changes scientific activity, but also how that response endogenously shapes the data environment faced by future AI tools.

(Romer, 1990; Weitzman, 1998; Agrawal et al., 2019). Structural biology provides a concrete mechanism: a solved protein creates local experimental spillovers for its sequence neighbours.

The remainder of the paper is organised as follows. Section 2 describes protein structural biology and AF2. Section 3 develops the model. Section 4 describes the data. Section 5 reports descriptive statistics. Section 6 presents the empirical strategy. Section 7 reports the results. Section 8 concludes.

2 Protein Structural Biology and AlphaFold2

Proteins are the molecular machinery of the cell: they catalyse reactions, transmit signals, copy genetic information, and form the structural components of every tissue. A protein is a linear chain of amino acids—each called a *residue*—and a typical protein contains several hundred. Because a protein's biochemical function follows from its three-dimensional structure, knowing the structure is central to drug discovery, disease biology, synthetic biology, and basic cellular biology.

Structural biology answers three broad questions about a protein. The first is its static three-dimensional fold—the protein folding problem of predicting the structure from the amino-acid sequence. It resisted computational solutions for decades.[4] The second is its conformational dynamics: how the protein moves and changes shape as it functions, beyond what a single snapshot captures. The third is the architecture of the complexes it forms with partner molecules—other proteins, DNA, RNA, or small-molecule drugs—which is studied computationally using ligand docking. Unlike the folding problem, the last two have no clean computational statement and have historically been approached by experiment. Some applications also require near-atomic geometric accuracy, particularly in drug design and the interpretation of disease-causing mutations. The methods of structural biology, both experimental and computational, are organised around providing different combinations of these.

[4]The difficulty has two parts. The first is a *search* obstacle: a 100-residue chain can adopt more than 10^{100} three-dimensional shapes, so exhaustive enumeration is infeasible. The second is a *scoring* obstacle: even with candidate shapes in hand, no available scoring function—a rule that evaluates how physically plausible a candidate shape is—is accurate enough to pick out the correct shape from the many incorrect ones that score almost as well.

2.1 Structure Determination and Target Selection

Experimental determination of a protein structure is a long, multi-stage pipeline with substantial attrition at every step. A team must clone the target gene, express the protein in a host system, and purify it in sufficient quantity and purity. The purified protein is then taken through one of three principal experimental methods—X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (cryo-EM)—each with its own specialised data collection, model building, and refinement (Kuhlman and Bradley, 2019).[5] Each method also has a characteristic bottleneck in turning raw measurements into an atomic structure. For instance, X-ray crystallography faces the phase problem where the diffraction data yield the intensities but not the phases of the scattered waves, and both are required to reconstruct the three-dimensional structure. In cryo-EM, the experiment yields a three-dimensional density map of the protein at limited resolution; the positions of individual atoms must be inferred separately by fitting an atomic model into the map.[6] A single structure can require months to years of work and cost on the order of $100,000–$300,000. Even the optimised high-throughput efforts of the Protein Structure Initiative (PSI), a structural-genomics programme running from 2000 to 2017, carried fewer than 3% of selected targets from selection to PDB deposition, at roughly $66,000–$138,000 per deposited structure (see Figure 9 in the Appendix) (Chandonia and Brenner, 2006a; Terwilliger et al., 2009).

Alternatively, computational methods predict a protein's structure without running a new experiment on the target. These methods are most developed for the static fold; computational tools for conformational dynamics and ligand docking also exist, but have historically been considerably less mature. Within static-fold prediction, methods differ in how much prior structural information they require. At one end, comparative

[5]The three experimental methods differ in how the data are acquired, the class of protein each is best suited to, and the resolution they achieve. In X-ray crystallography, the purified protein is grown into a well-ordered crystal and exposed to an X-ray beam; the scattered X-rays produce a diffraction pattern from which the electron density of the protein, and ultimately an atomic model, is computed. It delivers the highest resolution but requires the protein to crystallise, which favours compact, rigid, soluble targets. In NMR spectroscopy, the protein is dissolved in solution and placed in a strong magnetic field; radio-frequency pulses excite specific atomic nuclei, and the resulting resonance signals encode local distances and angles between atoms, from which the structure is calculated. NMR uniquely captures protein dynamics and disorder but is practical only for relatively small proteins. In cryo-EM, many copies of the protein are flash-frozen in a thin film of vitreous ice and imaged in an electron microscope; each image is a two-dimensional projection of an individual particle at a random orientation, and combining hundreds of thousands of images yields a three-dimensional density map. Cryo-EM neither requires crystals nor imposes a tight size ceiling, and since the "resolution revolution" in detector and algorithmic technology, it has become the method of choice for large complexes and membrane proteins that the other two struggle with (Kühlbrandt, 2014).

[6]In NMR, the experiment yields a sparse and noisy set of distance constraints between atoms in the protein, from which the three-dimensional structure must be inferred.

(homology) modelling uses the solved structure of a close evolutionary relative as a template, producing a model at negligible cost when such a relative exists (Martí-Renom et al., 2000). At the other end, template-free methods rely only on physical principles and the amino-acid sequence: physics-based simulation runs molecular folding dynamics in silico, while ab initio prediction infers structure from statistical regularities learned across known structures. Until AF2, template-free methods were far less reliable than template-based modelling and were confined to small proteins (Shaw et al., 2010; Kuhlman and Bradley, 2019). Computational methods are therefore most informative where the experimental record is already dense, and they degrade sharply as the target moves away from solved neighbours. Within this template-rich region, predictions can stand on their own for questions that need only coarse structural information, such as identifying compact domains and generating initial functional hypotheses.

In practice, experimental and computational methods are often combined (Ward et al., 2013). Molecular replacement addresses the phase problem in X-ray crystallography by using a known related structure as a starting model to interpret raw diffraction data; it now accounts for the great majority of new depositions (McCoy et al., 2022). In cryo-EM, the moderate-resolution density map of a multi-protein complex can be interpreted at atomic detail by fitting in previously solved structures of its component proteins. Both uses require a close homologue: when none exists, computational prediction falls back to weak ab initio methods, and orphan or lineage-specific proteins remain inaccessible from both sides.

These methods have been applied to a small, non-random subset of proteins. For the human proteome—among the most intensively studied—only 17% of protein residues had been resolved in an experimental structure before AF2, and around a quarter of human proteins had no experimental structural data at all (Porta-Pardo et al., 2022); coverage of the broader protein universe is sparser still (Zhang et al., 2006).

Three forces drive the selection. The primary driver is biological demand: the need for a structure originates in a specific question—a disease mechanism, a drug target, a pathway puzzle—and the protein selected is the object that question points to. Laboratories focus on proteins for which a solved structure yields the greatest scientific and clinical return (Petsko, 2007).[7] The second is experimental tractability. Because success rates vary enormously across proteins, teams favour targets predicted to be soluble, single-domain,

[7]The structural-genomics consortia of the 2000s—most prominently the NIH Protein Structure Initiative—were a deliberate exception, running "structure first" by selecting targets from large protein families that lacked any structural representative in order to maximise coverage of fold space per structure solved, an approach that explicitly decoupled target choice from immediate biological demand (Brenner, 2000; Marsden and Orengo, 2008; Liu et al., 2007a).

and crystallisable, and avoid large, membrane-embedded, or intrinsically disordered ones (Slabinski et al., 2007). The third is proximity to existing structural knowledge. Proteins are not isolated problems: because proteins sharing more than 30% sequence identity almost always share the same three-dimensional fold (Rost, 1999), resolving one protein generates a positive externality for its sequence neighbours. A close homologue makes the next experiment cheaper—it serves, for instance, as a molecular-replacement template (Terwilliger et al., 2014)—and itself yields a homology model that researchers often use in place of an experimental structure when one is not available (Martí-Renom et al., 2000). The cumulative result, after decades, is an experimental record heavily skewed towards proteins that are biomedically prominent, evolutionarily well-connected, and experimentally convenient (Peng et al., 2004; Chandonia and Brenner, 2006b).

2.2 AlphaFold2

AlphaFold2 (AF2) abruptly closed a large part of this coverage gap. DeepMind announced the system's near-experimental accuracy in blind structure prediction at the 14th Critical Assessment of Protein Structure Prediction (CASP14) in November 2020. The economically relevant availability shock came in July 2021, when DeepMind and EMBL-EBI publicly released predicted structures for approximately 350,000 proteins, including nearly the entire human proteome, and made the AF2 code available for research use. The database expanded to over 200 million proteins in July 2022. In recognition of the underlying advance, Demis Hassabis and John Jumper, the lead developers of AF2, were awarded one half of the 2024 Nobel Prize in Chemistry for protein structure prediction; the other half went to David Baker for computational protein design (The Royal Swedish Academy of Sciences, 2024).

AF2's success stems from a deep learning architecture that integrates biological and physical knowledge (Jumper et al., 2021). The system takes three inputs: the amino acid sequence, a multiple sequence alignment (MSA) of evolutionarily related proteins, and an optional structural template from the PDB. The MSA provides the primary predictive signal—correlated mutations across species reveal which residue pairs are spatially proximate in the folded protein. The number of homologous sequences in the alignment—the *MSA depth*—determines how strong this signal is: a deep MSA covers many independent evolutionary observations and yields confident contact predictions, while a shallow MSA leaves AF2 with little to learn from. Attention-based neural networks jointly reason over these inputs, iteratively refining inter-residue distance and orientation predictions to assemble a full three-dimensional structure. The network was trained on the ∼170,000 ex-

perimental structures in the PDB released before April 30, 2018. Beyond these experimental structures, AF2 also trained on its own high-confidence predictions for sequences with no solved structure—a self-distillation step in which the model learns from pseudo-labels it generated itself. High-quality predictions can be generated from sequence and MSA alone—AF2 internalises evolutionary signals and physical regularities rather than relying on solved templates or explicit folding simulations. All three inputs were already available to structural biologists before 2021: sequences from the universal protein-sequence database, MSAs from standard alignment tools, and templates from the PDB. What AF2 adds is the deep network itself — a learned non-linear function that extrapolates from these inputs with an accuracy that template-based homology modelling and physics-based simulation could not approach.

Crucially, AF2 does not deliver uniform certainty across all proteins or all regions of a protein. Each prediction is accompanied by confidence metrics, most prominently the per-residue predicted Local Distance Difference Test (pLDDT) score (see Figure 10) and the predicted aligned error (PAE). These measures reveal substantial heterogeneity: well-structured, evolutionarily conserved regions are often predicted with very high confidence, while loops, disordered segments, and inter-domain orientations are markedly less reliable. Proteins with limited evolutionary information—such as lineage-specific proteins or those with few homologues—also tend to receive lower confidence scores (Akdel et al., 2022). As a result, AF2 provides precise structural information for some parts of the proteome while leaving other regions uncertain.

The release of AF2 fundamentally altered the workflow: high-confidence predictions now serve as default starting points, shifting structure determination from unguided discovery towards targeted validation (Campbell, 2024). Benchmarks confirm AF2 predictions are accurate enough to support molecular replacement in most cases (Terwilliger et al., 2023, 2024), including for structures that had previously resisted standard techniques (Barbarin-Bocahu and Graille, 2022). For experimentally intractable proteins, predictions can also guide downstream functional studies without any experimental structure (Varadi et al., 2023; Kovalevskiy et al., 2024).[8]

2.3 The Evolution of AlphaFold's Training Data

To understand how AF2 reshapes experimental choice, it is useful to distinguish three sources of information. First, the PDB records experimentally determined structures.

[8]There are limited contexts—such as early-stage drug discovery—where researchers rely directly on high-confidence AlphaFold predictions when experimental structures are unavailable or unnecessary (Ren et al., 2023).

Second, sequence databases contain evolutionary information that the model processes via MSAs. Third, prediction systems can use high-confidence model outputs as additional training inputs. The interaction among these sources creates a feedback loop: past experiments help train current prediction systems, current predictions change experimental choices, and new experiments become part of the corpus available to train later systems.

From experimental anchors to AF2. The first AlphaFold system (AF1) won CASP13 in 2018, establishing deep learning's viability for structure prediction (Senior et al., 2020); its code was released, but with no public structure database. AF2 should not be understood as a mechanical template-lookup system. It learns from experimental anchors and evolutionary sequence information (Jumper et al., 2021).

AF3 and the value of new experiments. Successor systems extend the prediction target beyond the static single-chain scaffold. AF3, for example, released in 2024 through a restricted, non-commercial server, models complexes containing proteins, nucleic acids, small molecules, ions, and modified residues (Abramson et al., 2024). These developments reinforce rather than eliminate the value of experimental data: experiments remain especially useful when biological interpretation requires ligands, alternative conformations, interactions, or validation. We return to this connection in Section 7.3, where the empirical outcomes identify which kinds of data AF2 induces researchers to produce.

A potentially virtuous circle. The history of AlphaFold suggests a potentially virtuous circle. Experimental structures train prediction systems; prediction systems reduce the cost of resolving additional biological questions; and the resulting experiments improve the corpus available for later models. The circle is not automatic. An AI tool may encourage experiments that deepen knowledge within its existing range while discouraging experiments on remote proteins whose structures would expand that range. The relevant empirical question is therefore not simply whether AF2 increases experimental activity. It is how AF2 changes the location and composition of experiments, and what that reallocation supplies to the next generation of AI. This data-supply consequence need not be internalised by the laboratory choosing the current experiment.

3 A Model of Experimental Choice

This section develops a stylised model that organises the empirical analysis. A usable AF2 prediction supplies a *scaffold* — the protein's static three-dimensional backbone, or

overall fold. This can reduce the return to an experiment that would reproduce the scaffold, while lowering the cost of an experiment that uses the scaffold to study a conformation, ligand, or interaction. The first force is *information substitution*; the second is *experimental complementarity*. Their balance determines which experiments laboratories conduct and, in turn, the composition of the experimental corpus inherited by later prediction systems. The model is deliberately parsimonious in the main text. Appendix A.1 gives a local foundation for the knowledge landscape, states the regularity conditions, and proves the formal results.

Figure 1 summarises the argument. Past experiments help train AF2. AF2 and the inherited structural landscape jointly change the return to current experiments through substitution and complementarity. The resulting choices then alter the experimental corpus supplied to later prediction systems. Laboratories need not internalise this last effect. Section 3.4 derives the resulting frontier-data wedge within the same bundle-based model.

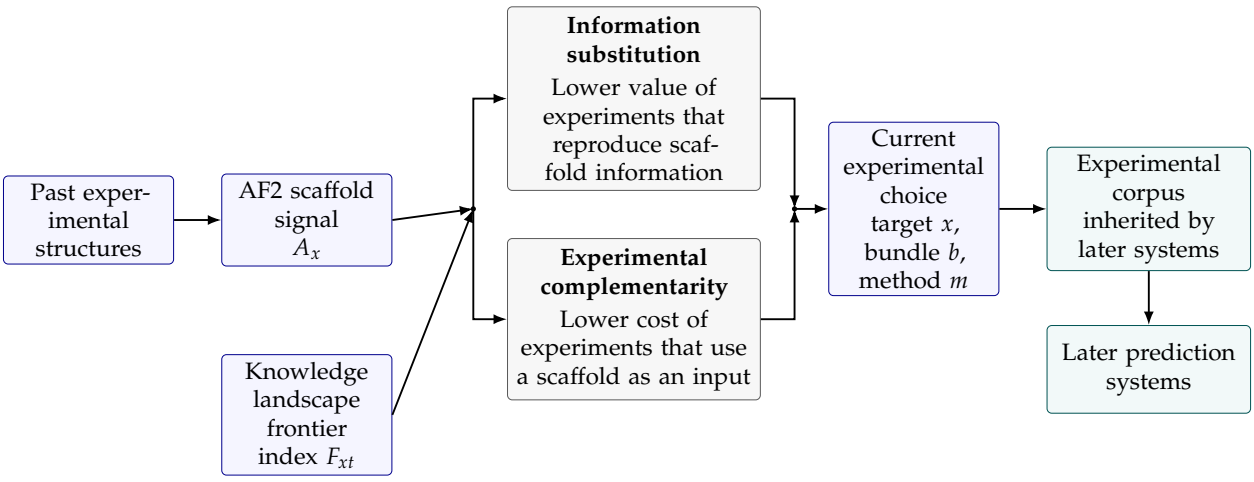


Figure 1: AF2, current experimental choice, and later experimental data supply

Notes. Solid arrows describe the baseline mechanism. Past experiments help train AF2. AF2 and the local knowledge landscape jointly change the payoff from current experiments through information substitution and experimental complementarity. Current choices alter the experimental corpus available to later systems.

3.1 What AF2 Supplies and What Experiments Produce

AF2 is informative about a protein's dominant single-chain scaffold. Its empirical usefulness is heterogeneous: a prediction may confidently cover most residues for one protein but leave important regions unresolved for another. Let $A_x \in [0,1]$ denote the usefulness of AF2's released scaffold prediction for protein x.[9] Our empirical measure, AF coverage, is a monotone but imperfect proxy for this object. It does not fully capture how a protein's structural modules are oriented relative to one another, the other molecules it binds, how multiple chains assemble into a larger complex,[10] which stretches are intrinsically disordered (floppy regions that never settle into a fixed shape), or whether AF2 happens to be confident on the part of the protein that an experiment actually cares about.

Structural knowledge is multifaceted. The formal model retains five primitive content components because a single deposition may combine several components and because first-ever feasibility requires distinguishing new scaffold coverage (R) from refinement of an existing scaffold (Q). The empirical analysis groups these primitives into three action classes:

$$\mathcal{A} = \underbrace{\{R, Q\}}_{\text{scaffold}} \cup \underbrace{\{S, L\}}_{\text{mechanistic}} \cup \underbrace{\{C\}}_{\text{corroboration}} .$$

scaffold work includes initial or extended residue-level scaffold information (R) and static refinement (Q); *mechanistic* work includes distinct conformational states (S) and biologically relevant ligand-bound or interaction structures (L); and *corroboration* (C) independently validates an existing public scaffold. A PDB deposition can contain more than one underlying component. A first-ever ligand-bound structure, for example, both establishes a scaffold and records ligand information. Thus, an experiment selects a protein x, an experimental method m that shifts costs, and a non-empty component bundle $b \subseteq \mathcal{A}$. A first-ever public structure must contain R, because it creates the first public scaffold for the protein, but it need not be only R: the same deposition may also carry conformational (S) or ligand-bound (L) mechanistic content. A Q-only refinement, by contrast, requires an existing scaffold. A follow-up experiment may instead use an existing or predicted scaffold as an input into a different question.

Table 1 summarises the distinction. AF2 may substitute strongly for scaffold work

[9] Notation is collected in Table 4.

[10] A protein's structure is described at four levels: the *primary* structure is its amino-acid sequence; the *secondary* structure is the local folding of that chain into helices and sheets; the *tertiary* structure is the full three-dimensional fold of a single chain (the scaffold); and the *quaternary* structure is the arrangement of several chains into a multi-subunit complex. AF2 predicts single-chain (tertiary) structure; predicting how chains assemble into complexes was the aim of AlphaFold-Multimer and, subsequently, AlphaFold3 (Abramson et al., 2024).

that loads directly on the static structure. It can partially substitute for some design or hypothesis-generation steps in mechanistic work, but it does not directly observe the experimentally realised conformational ensemble or biologically relevant ligand-bound state. For those actions, AF2 is primarily an input into the experiment rather than its output. Corroboration occupies an intermediate position: the informational value of another generic scaffold measurement falls, but the value of validating an important prediction may rise.

Table 1: AF2 and grouped experimental actions in the model

Input or action	Knowledge produced at x	Principal relationship to AF2	Value for later AI data supply
AF2	Predicted dominant scaffold.	Public scaffold signal with heterogeneous usefulness.	Changes experimental choices; does not add experimental ground truth.
Scaffold (R/Q)	Initial or extended residue-level scaffold and static refinement.	Relatively close substitute.	New anchors and static precision, especially when fold- or sequence-novel.
Mechanistic (S/L)	Distinct conformations, ligand-bound states, and interaction structures.	Predicted scaffold can be an input; substitution is partial.	Conformational and interaction observations.
Corroboration (C)	Independent validation of a public scaffold.	Generic information value falls, but validation demand may rise.	Reliability and error-correction observations.

3.2 Experimental Choice: Substitution versus Complementarity

For each protein x, let $J_t(x) \in \{0, 1\}$ indicate whether any experimental structure has been deposited publicly. Let $I_{\ell xt} \in \{0, 1\}$ indicate whether laboratory ℓ has previously built protein-specific capital for x. A protein may therefore be familiar to the field but new to a particular laboratory. Define

$$h^{sc}_{\ell xt}(b) = \mathbf{1}\{I_{\ell xt} = 0,\ J_t(x) = 1,\ b \cap \{S, L, C\} \neq \emptyset\}.$$

The indicator $h^{sc}_{\ell xt}(b)$ captures the scaffold-input cost faced by a laboratory that enters an already solved protein in order to conduct a scaffold-dependent follow-up experiment. Before AF2, such a laboratory may need to obtain a usable scaffold before it can study a conformation, ligand, or interaction. A laboratory already inside the protein's pipeline has built some of this capital itself. A laboratory making a first deposition must establish the

first public scaffold within the experimental episode.

Let F_{xt} denote a protein's remoteness from existing structural knowledge. The index is higher where the nearest experimental anchor is distant, and the cost of an experiment is scaled by a positive increasing function $\chi(F_{xt})$. Let $s_b(A_x, F_{xt}) \geq 0$ denote the loss in the informational value of bundle b because AF2 already supplies part of its output. This loss may also vary with remoteness, because a scaffold prediction can displace more value than an experiment would otherwise have supplied in scarce scaffold information. Let $\delta^{sc}_m(A_x)$ denote relief on the scaffold-input cost and let $\delta^{g}_{b,m}(A_x)$ denote broader bundle-specific cost relief, including relief in construct design, model building, density interpretation, or the co-production of mechanistic information with a first scaffold. Define

$$\mathcal{R}_{b,m}(A_x, I_{\ell xt}, J_t(x)) = \underbrace{\delta^{sc}_m(A_x) h^{sc}_{\ell xt}(b)}_{\text{scaffold-input relief}} + \underbrace{\delta^{g}_{b,m}(A_x)}_{\text{bundle-specific variable-cost relief}} . \tag{1}$$

The direct AF2 payoff shock is

$$\Delta_{\ell t}(x, b, m) = \underbrace{\eta_m \left[\delta^{sc}_m(A_x) h^{sc}_{\ell xt}(b) + \delta^{g}_{b,m}(A_x)\right] \chi(F_{xt})}_{\text{experimental complementarity}} - \underbrace{s_b(A_x, F_{xt})}_{\text{information substitution}} . \tag{2}$$

The two terms name the central trade-off. *Information substitution* reduces the value of experiments whose output AF2 partly supplies. *Experimental complementarity* lowers the cost of experiments that can use a predicted scaffold as an input. Where both forces act on the same margin, the net sign is contingent. Appendix Proposition 4 formalises the experiment-specific conditions, and Table 2 maps these channels onto the empirical margins.

Table 2: AF2 payoff channels across empirical margins

Margin	Information substitution	Experimental complementarity
First-ever structure	Material: the experiment must produce the first public scaffold.	Limited to variable or co-production relief.
Scaffold follow-up	Material: the experiment extends or refines a scaffold.	Possible variable-cost relief.
Mechanistic follow-up	Partial: AF2 does not observe the realised state or ligand-bound structure.	Predicted scaffold can lower design, construction, or interpretation costs.
Outside-pipeline follow-up	Partial for mechanistic bundles.	Additional scaffold-input relief when a laboratory enters a mapped protein.
Inside-pipeline follow-up	Partial for mechanistic bundles.	Variable-cost relief remains even after the laboratory has built protein-specific capital.
Corroboration	Generic informational value falls.	Validation value or validation cost relief may rise.

The first-ever and inside-pipeline cases are algebraically similar but economically distinct. A first-ever experiment receives no scaffold-input relief because it must establish the first public scaffold itself. An inside-pipeline laboratory receives no such relief because it has already built protein-specific capital. By contrast, an outside-pipeline laboratory can use AF2 to avoid part of the setup cost that previously made scaffold-dependent follow-up work expensive. The model, therefore, predicts a reallocation towards mechanistic experiments for which AF2 supplies an input without supplying the experiment's main output. It also allows follow-up activity to rise both outside and inside existing laboratory pipelines: the former identifies scaffold-input relief, while the latter points to variable-cost relief or reallocation from more substitutable experiments.

The ranking across methods is contingent rather than mechanical. X-ray crystallography can gain through molecular replacement, while cryo-EM and NMR can gain through model building and interpretation. Institutions may also differ in their rewards. Industry laboratories may place greater weight on ligand-bound and corroborating experiments when a validated structure has commercial value. Academic laboratories may place broader weight on scientific spillovers. These comparisons help identify the channels through which AF2 changes experimental costs and rewards.

In the short run, laboratories choose among protein–bundle–method combinations with finite experimental capacity. A payoff increase raises an experiment's intensity or its rank relative to alternatives whose payoffs rise less or fall. In the longer run, laboratories

may hire, redirect grants, acquire equipment, or enter the field. Aggregate experimental volume can therefore rise or fall with the average payoff shock. The model predicts the direction and composition of activity; it does not impose a fixed aggregate response.

3.3 Why New Structures May Move Inwards

Proteins are embedded in a sequence-similarity tree.[11] A solved structure is an experimental anchor: it improves knowledge at the focal protein and can provide information for nearby proteins through homology. Let F_{xt} denote the remoteness of target x from the inherited structural map. A remote experiment can be socially valuable because it opens a new region of sequence space, but it is also harder because usable templates and nearby experimental knowledge are thinner.

The empirical analysis uses nearest-solved distance as the primary landscape measure. The distance D_{xt} records how far target x is from a usable structural anchor. It is the relevant cost object because a close-solved homologue can serve as a homology model or molecular-replacement template; when no close anchor exists, both experimental and computational methods lose an important starting point. Conditional on a protein being previously unsolved, the same distance also indexes range-expansion value: a first-ever structure far from solved homologues supplies ground truth in a region later systems cannot easily infer from the inherited corpus. We use the gap G_{xt}, and the within-cluster solved share ρ^{cl}_{xt} as robustness measures. The gap records the breadth of the unsolved region surrounding the target, while the within-cluster solved share records additional close anchors inside the target's sequence cluster. The local tree foundation in Appendix A.1.1 gives their common interpretation:

$$F_D, F_G \geq 0, \qquad F_{\rho^{cl}} \leq 0.$$

Distance and gap width are measures of remoteness; the local solved share is an inverse measure of remoteness.

Remoteness sharpens the substitution-versus-complementarity trade-off. For any landscape measure $\nu \in \{D, G, \rho^{cl}\}$, the relevant structural cross-partial is

$$\frac{\partial^2 \Delta_{\ell t}(x, b, m)}{\partial A \, \partial \nu} = \left[\underbrace{\eta_m \mathcal{R}_{b,m,A}(A_x, I_{\ell xt}, J_t(x)) \chi'(F_{xt})}_{\substack{\text{frontier-amplified} \\ \text{complementarity}}} - \underbrace{s_{b,AF}(A_x, F_{xt})}_{\substack{\text{frontier-amplified} \\ \text{substitution}}} \right] F_\nu. \tag{3}$$

[11] The tree is a reduced-form geometry for sequence similarity and local homology spillovers. It is neither a literal species phylogeny nor a claim that a single scalar tree statistic fully summarises protein tractability.

For a remoteness measure such as nearest-solved distance, the interaction is positive when experimental complementarity becomes more important as the local map thins. It is negative when the informational value displaced by AF2 grows more strongly with remoteness. For the solved-share measure, the sign reverses because higher local density implies lower remoteness.

This comparative static yields the model's sharpest directional prediction. A first-ever experiment must produce an R component, so its scaffold-input indicator is zero. If AF2 substitutes most strongly for remote scaffold discovery, a more informative AF2 prediction moves the remaining first-ever experiments inwards towards proteins with nearby solved homologues. Mechanistic follow-up work can reveal the opposite gradient: where the local map is thin, an AF2 scaffold may be especially useful as input for a conformation or ligand experiment. Appendix Proposition 6 states the formal target-selection result.

3.4 Experiments as Inputs into Future AI and the Frontier Externality

Current experimental choices affect the corpus inherited by later prediction systems. The model represents this consequence as *directional experimental data supply*, not as a prediction of realised successor-model performance. Let

$$\mathbf{T}^{exp}_{g+1}(z) = \left(T^{\text{scaf}}_{g+1}(z), T^{\text{mech}}_{g+1}(z), T^{\text{corr}}_{g+1}(z)\right) \tag{4}$$

denote the experimental input stocks available to a later prediction-system generation at target z. The components record scaffold-range (R/Q), mechanistic (S/L) and corroboration (C) observations from experiments.

Because these contributions are distinct (Table 1), a scalar count of depositions cannot distinguish them; the data supply is therefore represented as a vector rather than a single count.

Appendix A.1.4 formalises this mapping. Let M^b_g denote deposits with bundle b generated while prediction-system generation g is available, let $\mathbf{\Gamma}_b$ record which data-supply components bundle b improves, and let $\mathcal{K}_b(z, x)$ capture local propagation from an experiment at x to a later prediction problem at z. Then

$$\mathbf{T}^{exp}_{g+1}(z) = \sum_{b \subseteq \mathcal{A},\, b \neq \emptyset} \int_{\mathcal{T}} \mathbf{\Gamma}_b \mathcal{K}_b(z, x)\, dM^b_g(x). \tag{5}$$

This expression allows a first-ever ligand-bound structure, for example, to broaden scaffold range and add mechanistic data at the same time.

The same structure yields a welfare implication. Let $\boldsymbol{\omega}(z)$ be a vector of marginal social weights on the experimental input stocks available at target z. The marginal downstream data-supply value of conducting bundle b at target x is then

$$\tau_{b,g}(x) \equiv \int_{\mathcal{T}} \boldsymbol{\omega}(z)' \boldsymbol{\Gamma}_b \mathcal{K}_b(z,x)\, d\mu(z), \tag{6}$$

the contribution of that experiment to the weighted downstream value of the inherited corpus, $\int_{\mathcal{T}} \boldsymbol{\omega}(z)' \mathbf{T}^{exp}_{g+1}(z)\, d\mu(z)$. Appendix A.1.4 gives the derivation. The object $\tau_{b,g}(x)$ is directional. It can be large because an experiment supplies mechanistic or corroboration data within a mapped region. For a first-ever structure in a remote region, it can also be large because the experiment creates an experimental anchor where fold or sequence-family coverage is thin.

A laboratory need not capture this full downstream value when it chooses its current experiment. Let $\pi^{cur}_{\ell t}(x,b,m)$ denote the laboratory's current payoff after the AF2 shock in (2), and let $\theta_\ell \in [0,1]$ denote the share of successor-system data-supply value internalised by the laboratory or its funder. The private ranking payoff is

$$\pi^{priv}_{\ell t}(x,b,m) = \pi^{cur}_{\ell t}(x,b,m) + \theta_\ell \tau_{b,g}(x), \tag{7}$$

whereas the corresponding social ranking payoff is

$$\pi^{soc}_{\ell t}(x,b,m) = \pi^{cur}_{\ell t}(x,b,m) + \tau_{b,g}(x). \tag{8}$$

The missing reward is therefore

$$\pi^{soc}_{\ell t}(x,b,m) - \pi^{priv}_{\ell t}(x,b,m) = (1-\theta_\ell)\tau_{b,g}(x). \tag{9}$$

This wedge does not require mechanistic deepening to be socially unimportant. It says that the private ranking of experiments can underweight the distinct range-expansion value of experimental ground truth.

Proposition 1 (AI-induced frontier-data underweighting)**.** *Let $H \subseteq \mathcal{T}$ be a high-frontier region. Suppose AF2 raises the selected activity for a set of follow-up bundles but lowers the selected activity for a set of first-ever bundles containing R in H. Suppose the increase in follow-up activity is large enough that total private experimental activity rises, while the induced weighted change in scaffold-range data supply in H is negative. Suppose further that the induced weighted change is positive for the mechanistic component. If $\theta_\ell < 1$ and $\tau_{b,g}(x) > 0$ for a frontier-expanding bundle, private choice underweights that experiment relative to the social ranking by $(1-\theta_\ell)\tau_{b,g}(x)$. If the*

total downstream data-supply value $\tau_{b,g}(x)$ increases with remoteness for these bundles, the total missing reward is largest at the frontier. If only the range-expansion component of $\tau_{b,g}(x)$ increases with remoteness, then the range-expansion part of the missing reward is largest at the frontier.

The proposition identifies a directional externality rather than an unambiguous welfare loss. AF2 can raise the production of valuable mechanistic and corroboration observations while reducing the remote experimental anchors that broaden the scaffold range. The empirical analysis, therefore, measures first-in-cluster structures and structurally novel folds alongside new conformations, biologically relevant ligands, and binding-affinity information. The sign of the net effect on successor-model performance remains outside the scope of the model because it also depends on architecture, pseudo-data, curation choices, proprietary data, and the tasks used to evaluate performance. Appendix A.1.5 proves Proposition 1; Appendix A.1.6 extends the same wedge to rewards for future human research and moonshot experiments.

3.5 Theory-Guided Empirical Predictions

The model delivers contingent predictions rather than unconditional signs. This is useful empirically: the estimates reveal which force dominates on each margin. The central payoff equation separates substitution from complementarity; the remoteness interaction identifies how their balance shifts across the landscape of inherited knowledge; and the data-supply vector explains why the composition of experiments matters for later AI. The welfare wedge adds a final implication: aggregate expansion does not resolve the allocation problem when private rankings underweight range-expanding ground truth. Table 3 collects these predictions by margin.

Table 3: Theory-guided empirical predictions

Margin	Conditional prediction
Total depositions	Rise when average cost relief exceeds information substitution.
First-ever structures and frontier location	Contract and move inwards when substitution for the required R component is amplified by remoteness more strongly than cost relief.
Mechanistic versus scaffold follow-up	mechanistic S/L bundles rise relative to scaffold R/Q bundles when scaffold-input or variable-cost relief exceeds partial substitution.
Novel folds and first-in-cluster deposits	Contract when frontier-expanding scaffold contributions fall.
Conformations and ligand structures	Rise if S/L complementarity dominates partial substitution.
Inside versus outside pipeline	Outside-pipeline mechanistic or corroborating work receives scaffold-input relief; inside-pipeline gains require variable-cost relief or reallocation from alternatives with weaker payoff shocks.
Methods, institutions, and corroboration	Responses are larger where AF2 relieves costly construction or interpretation and where validation or downstream value is high.
Total activity versus frontier-data supply	Total deposits can rise while range-expanding deposits fall when complementarity dominates in aggregate, but substitution dominates for remote first-ever bundles.

4 Data

To study whether and how AF2 redirected scientific attention across sequence space, we construct a balanced panel observed annually from 2017 through 2025. We gather a rich set of protein-level characteristics, including each protein's position within sequence space relative to the pre-AF2 structural knowledge frontier. For every protein, we measure the quality and coverage of its AF2 prediction and track scientific attention.

4.1 Proteins as Scientific Problems

The starting point is the Universal Protein Resource Knowledgebase (UniProtKB), the central repository for protein sequence and functional information(The UniProt Consortium, 2025).[12] Each UniProtKB record is indexed by a stable accession number and provides rich

[12]UniProt constructs protein sequences by translating gene sequences from major genome repositories, providing a standardised and stable mapping from DNA to protein identifiers. More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources. UniProtKB excludes protein sequences

information on the protein, including its sequence length, family, organism, membrane status, associated diseases, and scientific publications—allowing us to characterise proteins in terms of their structural complexity, functional scope, and clinical and scientific relevance. Protein-related publications are obtained using the PubMed identifiers (PMIDs) associated with each protein in UniProt. We obtain bibliographic information for these publications from PubMed.

UniProtKB currently contains over 250 million protein entries, of which 573,661 have been manually curated through literature review and curator-verified computational analysis (SwissProt). We restrict the analysis to proteins in SwissProt deposited before 2017, ensuring that all proteins in the sample were known before the treatment period began and that exposure to AF2 does not select on entry into the protein universe. This yields 530,495 proteins that constitute the sample for this study.

We construct a sequence-similarity tree for the SwissProt proteins using neighbour joining. The construction proceeds in three steps. First, we group proteins into *clusters* of closely related sequences using MMseqs2 (Steinegger and Söding, 2017), a standard tool for sequence-similarity search and clustering at scale: proteins sharing at least 50% sequence identity are assigned to the same cluster, with one representative sequence per cluster. This reduces the ∼573K proteins to ∼127K clusters. Second, we compute pairwise sequence identity between all cluster representatives, retaining pairs that share at least 30% identity—the threshold above which proteins reliably share the same three-dimensional fold (Rost, 1999). Clusters that share less than 30% identity with every other cluster have no edges in this network; proteins in these clusters are classified as *isolated*. Third, we build a neighbour-joining tree from the connected clusters by iteratively joining the pair of nodes that minimises the total branch length, yielding a tractable tree metric for sequence-similar clusters (see Appendix A.2.5 for construction details). Cluster size also serves as our protein-level proxy for MSA depth: the number of SwissProt entries in a protein's MMseqs2 cluster counts the close sequence homologues documented for that protein in the curated database, which is what AF2's MSA pipeline ultimately draws on to infer inter-residue contacts.

from most non-germline immunoglobulins and T-cell receptors, synthetic sequences, patent application sequences, small fragments (< 8 amino acids), and pseudogenes.

4.2 Experimental Structures

The Protein Data Bank (PDB) is the worldwide archive of experimentally determined structures of biological macromolecules (Berman et al., 2000).[13] Each PDB entry is a revealed experimental choice: a research team decided to determine the structure of protein p at time t, given the existing landscape of structural knowledge. The accumulation of experiments across proteins and time therefore records where scientific effort has been directed.

As of January 2026, there are 372,413 PDB depositions associated with 38,561 proteins. Each PDB entry records the year it was made, the depositing team, the method used (X-ray crystallography, cryo-EM, or NMR), the resolution of the structure, and the associated publication.[14] We exclude two classes of PDB depositions that do not reflect ordinary experimental choices: batch deposits arising from structural genomics consortia and high-throughput pipelines, which mechanically inflate counts, and SARS-CoV-2 protein structures, whose post-2020 deposition surge reflects pandemic-era emergency research that overlaps with our treatment window.[15]

We identify the laboratory of each PDB deposition by linking the deposition's primary citation to OpenAlex. We take the disambiguated OpenAlex ID of the last listed author as the laboratory identifier. We group lab types into academic (education, facility, healthcare, government, nonprofit), industry (company), and other, based on the last author's institution. Each PDB deposition is characterised along four orthogonal dimensions.

Frontier expansion and deepening. We distinguish between depositions that provide the *first-ever* experimentally resolved structure for a protein and *follow-up* depositions on proteins that had already been structurally characterised. We also determine the location of the targeted protein relative to the structural map. A first deposition most clearly expands the structural map when the protein is remote from existing structural knowledge. A

[13] Mandatory deposition was driven by the structural biology community beginning in 1989 and solidified when the Research Collaboratory for Structural Bioinformatics (RCSB) took over PDB management in 1998. Since the early 2000s, virtually all journals have required deposition in the PDB as a condition for publication, making the archive close to exhaustive for published experimental structures.

[14] Resolution (in Ångströms) captures the level of structural detail. Values of 1.0–2.0 Å represent very high detail with confidently positioned side chains; 2.5–3.5 Å is typical quality with clear backbone but fuzzier side chains; above 4.0 Å only coarse domain placement is possible. Depositions include three dates: deposition, initial public release, and subsequent revisions. We use the initial release date as the relevant timing for scientific availability.

[15] Batch deposits are identified as PDB depositions of multiple structures by the same team on the same date, typically generated by structural genomics consortia, large-scale pipeline projects, and industrial high-throughput operations. SARS-CoV-2 structures are identified by NCBI Taxonomy ID 2697049. We report results on this cleaned sample throughout; estimates without these restrictions appear in Appendix Table 39 and yield qualitatively identical conclusions.

first deposition near a solved homologue instead fills in the mapped interior. Follow-up depositions deepen knowledge by construction.

Content components. We classify depositions according to the type of structural knowledge they contribute. Multiple depositions may exist for a given protein, reflecting different sequence regions, conformational states, ligand-bound forms, or experimental refinements, and a single deposition often resolves only part of the full protein sequence. We therefore construct several non-mutually-exclusive categories based on observable features of the deposition record.

We link depositions to UniProtKB proteins using the Structure Integration with Function, Taxonomy and Sequences (SIFTS) resource, which provides residue-level alignments between PDB depositions and UniProtKB accessions (Velankar et al., 2013). Residues—the individual amino acids in a protein chain—are indexed according to the UniProtKB reference sequence, and SIFTS identifies which residues are experimentally observed in each deposition. We define a deposition as contributing to *static* structural knowledge if it extends residue coverage beyond what had previously been resolved for the protein.

A deposition contributes to the *mechanistic* knowledge of a protein if it reveals a previously unobserved conformational state or introduces biologically relevant ligand information. To identify conformational novelty, we compare each deposition to earlier structures of the same UniProt protein using Foldseek structural clustering.[16] We classify a deposition as ligand-relevant if it lists at least one non-cryoprotectant chemical component, reports a quantitative binding affinity (e.g., K_d, K_i, or IC_{50}), or if its primary citation carries a binding-related MeSH descriptor.[17]

We define a *refinement* deposition if it improves on the best previously available within-method quality metric (e.g., a higher-resolution X-ray structure, a lower X-ray R_{free} or clashscore, or a higher-resolution cryo-EM map). Finally, a *corroborating* deposition is a follow-up structure that re-measures an already-resolved scaffold without contributing additional residue coverage, conformational novelty, ligand information, or quality improvements. Appendix A.2.3 provides the detailed field-level definitions and overlap

[16] Foldseek is a structural alignment and search tool (van Kempen et al., 2024). It encodes each residue's local three-dimensional environment into a 20-letter "3Di" alphabet, allowing efficient structure-versus-structure comparisons using a sequence-search engine. We apply Foldseek to PDB structures associated with the same UniProt protein to identify follow-up depositions representing distinct conformational states.

[17] Examples of binding-related MeSH descriptors recognised in our pipeline include *Protein Binding*, *Binding Sites*, *Ligands*, *Substrate Specificity*, *Allosteric Site*, *Allosteric Regulation*, *Protein-Protein Interaction Maps*, *Molecular Docking Simulation*, *DNA-Binding Proteins*, *RNA-Binding Proteins*, and *Drug Design*. These are NLM Medical Subject Headings that flag the primary scientific content of a paper as concerning the binding or interaction between a protein and a small molecule, nucleic acid, or partner protein.

statistics across categories.

Training value. The content-component flags above classify what each deposition reveals about its own protein; a complementary view asks what each deposition contributes to the training corpora of next-generation structure predictors (Appendix A.2.4). The novelty axis isolates structural information absent from the public corpus at three nested scopes. *New conformation* marks a state of the same protein not previously observed — the information a model needs to learn how a fold moves rather than how it sits. *First in cluster* marks the first deposition anywhere in the protein's sequence family, bringing a template into a region of sequence space where a sequence-to-structure model previously had none. *Novel structure* marks a fold absent from the entire prior archive — a region of structural space a predictor cannot reach without direct observation. The biological richness axis isolates information that AF2 cannot derive from sequence alone, but successor models explicitly aim to predict. *Binding affinity* restricts to depositions reporting a quantitative binding strength, the central training input for protein–ligand affinity prediction. *Biological ligand* restricts to depositions containing a biologically meaningful substrate, cofactor, inhibitor, or drug-like molecule — a stricter curation than the broader ligand content-component flag, which still admits crystallographic additives — supplying the substrate-bound complexes from which next-generation models of protein–ligand interaction must learn.

Pipeline. We classify each deposition according to its relationship to the depositing laboratory's prior research pipeline. A deposition is *outside the lab's pipeline* if the laboratory had not previously deposited on protein p, and *inside the lab's pipeline* otherwise. We further distinguish whether outside-pipeline depositions correspond to proteins that were entirely new to the field or had already been structurally characterised by other laboratories. These categories are constructed by comparing each deposition's release date to the earliest prior depositions on the same protein, both globally and within the laboratory. Same-day multi-depositions by the same laboratory are treated as a single lab event.

4.3 Structural Knowledge Frontier

To study how AF2 redirects experimental choices, we therefore need a measure of each protein's proximity to existing structural knowledge at each point in time. The sequence-similarity tree is fixed—determined by protein sequences, not by experiments. What changes over time is which proteins have been *solved*: structural knowledge is cumulative, so the set of solved proteins expands monotonically as new depositions accumulate. In

2000, only roughly 2,900 clusters had any PDB coverage; by 2025, this had grown to roughly 24,800 (19.5% of all clusters). As more proteins are solved, the knowledge frontier contracts and distances shrink for the remaining unsolved proteins.

For each protein, we characterise its position relative to the pre-AF2 structural knowledge frontier through a tree-path metric on the 2016 neighbour-joining sequence-similarity tree (Appendix A.2.5). Let $\mathcal{S}_{2016}$ denote the subset of SwissProt proteins with at least one PDB structure deposited on or before 31 December 2016. The *distance to the nearest usable solved protein* is the empirical snapshot

$$D_p^{2016} \equiv D_{p,2016}, \tag{10}$$

the protein-level tree-path distance to a member of $\mathcal{S}_{2016}$. Following the construction in Appendix A.2.5, Step 6, we use the within-cluster path whenever p's cluster already contains a solved member and otherwise use the shortest path to an external solved cluster. Inter-cluster edges carry weight $1 - \text{pident}/100$ where pairwise identity is detectable ($\geq$ 30%) and weight 1 otherwise, plus an intra-cluster spoke from p to its cluster representative. Lower D_p^{2016} corresponds to a closer chain of pre-2017-solved homologues and therefore a more usable homology-modelling template. We complement this measure with a tree-anchored *gap*,

$$G_p^{2016} \equiv G_{p,2016},$$

the sum of tree-path distances to the nearest pre-2017-solved external anchors reached from opposing sides of p's cluster. The gap captures the width of the unsolved region surrounding p along the sequence-similarity tree: a protein bracketed by two distant flanking anchors sits in a sparse stretch of sequence space, regardless of whether one anchor happens to be close; Figure 2 illustrates the distance and gap on a stylised tree. The *share of solved cluster members* accounts for the density of structural knowledge within p's own cluster.

These measures are the empirical counterparts to $(D_{xt}, G_{xt}, \rho_{xt}^{cl})$ in Section 3.3. Nearest distance measures the closest usable structural template and is the primary measure used in the main tables. Gap width distinguishes a locally close template from a broadly mapped neighbourhood, while the solved-cluster share records extra precision supplied by nearby anchors. We use these two alternatives as robustness measures. Lemma 2 shows why larger distance and gap measures, but smaller solved shares, shift a protein towards the structural frontier.[18]

[18]For proteins in clusters with no detectable sequence similarity ($\geq$ 30% identity) to any other cluster, no tree path to a solved neighbour exists. We max-impute nearest distance, flag the observation with the

4.4 AlphaFold2

The AlphaFold Protein Structure Database (AFDB) provides predicted three-dimensional structures for nearly all known proteins, indexed by UniProt accession numbers. Each AFDB entry includes local confidence measures. The predicted Local Distance Difference Test (pLDDT) is a per-residue score from 0 to 100 capturing expected local accuracy, with established confidence bands: above 90 (very high—reliable backbone and side-chains), 70–90 (confident—reliable backbone), 50–70 (low—possible disorder), and below 50 (very low—typically disordered) (Jumper et al., 2021).[19]

To capture the usefulness of AF2's prediction at the protein level, we define the fraction of residues i in protein p whose predicted pLDDT exceeds threshold q:

$$\text{AF cov}_p^q = \frac{1}{L_p} \sum_{i=1}^{L_p} \mathbf{1}\{\text{pLDDT}_{pi} \geq q\}, \tag{11}$$

where L_p is the protein's sequence length and $q \in \{0, 50, 70, 90\}$ represents increasingly stringent confidence thresholds. In our baseline specification, we use $q = 70$, the threshold at which backbone coordinates are reliably predicted, and the structure is suitable for molecular replacement, construct design, and cryo-EM model fitting—the primary channels through which experimentalists use AF2 predictions (Akdel et al., 2022; Terwilliger et al., 2024).[20] This measure provides a continuous notion of exposure to the AF2 shock: the fraction of the scientific problem that AI solved at a usable level of confidence.

5 Descriptive Statistics

Our sample consists of 530,495 proteins observed in an annual panel from 2017 to 2025, comprising 93,522 PDB depositions. Experimental activity is sparse and dominated by follow-up (Panel A of Table 6). A typical protein receives almost no experimental attention in a typical year. The average protein–year contains 0.019 PDB depositions and 0.178 publications. The distribution is severely right-skewed: the maximum deposition count in a single protein–year is 132, and a small number of well-studied proteins accumulate dozens of structures across the window, while most proteins receive zero. Within the

isolated indicator, and treat undefined gap values separately in specifications that use them.

[19] pLDDT reflects a protein's intrinsic predictability by deep learning models. Later AFDB updates may adjust confidence scores, but these do not represent new information becoming available to scientists; rather, they refine estimates of a property already revealed with the initial release.

[20] For 2,508 proteins, AF2 did not provide predictions. These are mostly viruses (mimivirus, bacteriophage T4, vaccinia, monkeypox, cytomegalovirus).

rare event of a deposition, follow-up work dominates. The same picture appears at the deposition level (Panel B): 88.5% of depositions are follow-ups.

Aggregate deposition counts mask substantial heterogeneity in the contributions of each deposition (Panels B and C). By content component within follow-up depositions, 15% extend static residue coverage, 79% are mechanistic, and 15% are corroborating. By experimental method, cryo-EM accounts for 30% of depositions over our window and X-ray crystallography for 67%; Figure 3 shows the underlying time trends. Academic laboratories account for 62% of depositions, while only 5% of depositions originate from industry firms.

Activity is concentrated on a small number of productive laboratories. Of the 9,119 distinct depositing labs over 2017–2025 (36.7% incumbents with a pre-2017 deposit), the top 10% account for 70.6% of depositions (Panel D of Table 6; Appendix Table 13). Only 33.7% make any first deposition, so two-thirds operate exclusively at the intensive margin, and pipeline reuse is frequent: 43.8% of depositions come from labs already inside the focal protein's pipeline; Appendix Figure 14 plots annual depositions by lab affiliation, entrant status, and pipeline.

The structural knowledge frontier is sparse but contracting over time: Appendix Figure 13 shows the cross-protein average distance to the nearest solved cluster and the within-cluster density, both moving steadily towards the centre since the 1990s. Figure 6 plots the cross-section distribution of D_p^{2016}, the 2016 tree-path distance to the nearest pre-2017-solved protein, with the 14.1% of proteins in clusters sharing no measurable sequence similarity piling up at the imputed maximum. The mean tree-path distance is 0.94, the mean gap is 2.33, and only 4.2% of a typical cluster's members have any pre-2017 PDB.

Table 14 reports the joint distribution of depositions along the extensive–intensive margin and the dense/sparse/isolated neighbourhood density (tertiles of the 2016 tree-path distance, with isolated proteins pooled into the top bin; see the table notes), separately for the pre-AF (2017–2020) and post-AF (2021–2025) periods. Two compositional shifts are visible without any controls: the share of follow-up depositions in the isolated neighbourhood rises from 10.0% to 13.2%, and the share of first depositions in the dense neighbourhood falls from 4.9% to 3.6%.

Most proteins have no experimental structure. Only 4.3% of proteins in our sample had PDB entry deposited before 2017 (Panel B of Table 5); the unconditional mean of PDB depositions per protein is 0.248 but is concentrated on a small well-studied subset, with a maximum of 684 depositions on a single protein. Against this backdrop, AF2 delivers a structural prediction for essentially every protein, but the predicted accuracy varies sharply across the proteome (Panel D of Table 5). At the standard usable-backbone threshold

(pLDDT $\geq$ 70), the cross-protein mean coverage is 0.868 (see the distribution in Figure 5). This means AF2 transforms the field from sparse and highly uneven experimental coverage to near-uniform predicted coverage.

The cross-protein variation in AF cov_p is not random. AF2 was trained on the heavily selected pre-2018 PDB and on multiple sequence alignments concentrated in well-sampled evolutionary families; Appendix Figure 11 documents this selection by organism, family, and drug-target and membrane status.[21] The transmission to AF coverage, however, runs mainly through evolutionary information rather than prior structural attention: AF coverage rises strongly with MSA depth but is approximately orthogonal to the 2016 distance to the nearest solved protein (Figure 4).

Figure 7 plots annual deposition means by AF coverage quartile (Q1 lowest, Q4 highest). Follow-up depositions on already-solved proteins (panel c) rise most in the highest-coverage quartile after 2021, while the first-deposition rate on unsolved proteins (panel b) is persistently lower for high-coverage proteins. Total depositions on the full proteome (panel a), by contrast, rise most in the *lowest*-coverage quartiles. This last pattern is unconditional and reflects composition rather than an AF coverage effect. The post-2021 cryo-EM expansion, for instance, disproportionately resolved large and membrane proteins, which have systematically lower AF coverage, lifting the low-coverage quartiles for reasons unrelated to AF2; and total activity blends the first-ever and follow-up margins, which need not move together. The protein fixed effects and the pre-AF features interacted with Post_t are designed to absorb exactly this composition, and the year fixed effects absorb the common 2020 uptick (the COVID-era backlog and the November 2020 CASP14 preview).

6 Empirical Strategy

Our goal is to recover the causal effect of AF2 on structural-biology activity. The counterfactual is the post-2021 trajectory of scientific output that would have prevailed had AF2 not been released—full proteome as a shock — and estimating how scientific activity responds to cross-protein variation in the informativeness of AF2 used to extrapolate from them (template-based homology modelling, biophysical simulation, expert priors). What AF2 adds, and what we want to isolate, is the highly non-linear extrapolation from those same inputs that the deep network performs. Our specification approximates this counterfactual through fixed effects and interactive controls that absorb the deposition

[21] Appendix Figure 12b adds the downstream evidence that AF coverage is systematically lower on intrinsically harder proteins, again consistent with the MSA-depth channel.

trajectory predicted by the observable pre-AF information set; the surviving variation in $\mathrm{AF\,cov}_p$ is intended to load on AF2's algorithmic content.

We exploit the AF2 release of the full-proteome as a shock and estimate how scientific activity responds to cross-protein variation in AF2 informativeness. For protein p in year t, we estimate the continuous difference-in-differences

$$\mathrm{link}\big(\mathbb{E}[Y_{p,t} \mid \mathbf{X}_p, \alpha_p, \lambda_t]\big) = \alpha_p + \lambda_t + \beta\left(\mathrm{Post}_t \times \mathrm{AF\,cov}_p\right) + \mathrm{Post}_t \times \mathbf{X}_p'\boldsymbol{\psi} \tag{12}$$

where $\mathrm{link}(\cdot)$ is the identity link when $Y_{p,t}$ is binary (estimated by OLS), and the log link when $Y_{p,t}$ is a non-negative count (estimated by Poisson pseudo-maximum likelihood, PPML). The treatment variable $\mathrm{AF\,cov}_p$ is the share of p's residues predicted with pLDDT ≥ 70 and Post_t is an indicator equal to one for $t > 2021$. Protein fixed effects α_p absorb all time-invariant protein characteristics, including sequence length, taxonomic family, intrinsic disorder, baseline experimental tractability, biomedical importance, pre-AF publication count, and the availability of pre-existing structural templates; the level term in $\mathrm{AF\,cov}_p$ is therefore absorbed by α_p. Year fixed effects λ_t absorb the aggregate time trend and any year-specific shock common to all proteins, including the average post-2021 level shift. Standard errors are clustered at the protein level throughout.

AF coverage enters as a continuous dose rather than a binary indicator. AF2 issued a prediction for essentially every protein, so there is no untreated control in a present-versus-absent sense; what varies is the *intensity* of usable scaffold information, from near-complete confident coverage to almost none (Section 4.4). A binary split would discard most of that variation and impose an arbitrary cutoff, and the continuous dose is the empirical counterpart of the model's latent usefulness $A_x \in [0, 1]$. This intensity is fixed at the release and time-invariant thereafter, and it is not randomly assigned: $\mathrm{AF\,cov}_p$ is largely determined by AF2's own inputs — MSA depth, biophysical foldability, and the density of PDB templates in p's sequence neighbourhood, all predetermined relative to the release — and is therefore correlated with pre-existing deposition trajectories, since proteins with deeper MSAs and denser template coverage already attracted attention through homology spillovers, methodological waves, and cumulative interest. With only α_p and λ_t, β would conflate AF2's algorithmic contribution — the network's non-linear extrapolation from these inputs — with the continuation of those pre-existing trends.

We therefore control for seven predetermined protein characteristics, $\mathbf{X}_p$, each interacted with Post_t. They absorb pre-AF heterogeneity along two margins: the cost of conducting a depositable experiment (supply) and its scientific or clinical value (demand). Family size (the number of SwissProt proteins sharing p's primary Pfam family) proxies

for the availability of homology-derived starting points and MSA depth; large families also attract recurring methodological waves — a kinase-optimised cryo-EM workflow, an ion-channel NMR protocol — that propagate to all members regardless of AF2. Sequence length captures the cost of expression, purification, and refinement, and is the single best protein-level predictor of crystallographic difficulty. The number of PDB depositions pre-2017 captures both experimental tooling and accumulated demand that survives into the post-2021 period and would have driven follow-up activity regardless of AF2. The number of cryo-EM depositions pre-2017 enters separately because the cryo-EM expansion — driven by detector and software advances independent of AF2 — preferentially benefits proteins with established cryo-EM workflows. The membrane-protein indicator marks a sharp cost discontinuity on which the post-2021 cryo-EM expansion has acted asymmetrically. Membrane proteins are also disproportionately drug targets (GPCRs, ion channels, transporters). The number of publications pre-2017 captures accumulated research attention, and the disease-association indicator captures clinical and therapeutic interest. After conditioning on these $\text{Post}_t \times \mathbf{X}_p$ terms, β measures the post-AF differential that survives those pre-AF correlates: AF2's algorithmic contribution rather than the inherited PDB bias the algorithm carries forward.

Because no protein is treated before 2021 and the dose $\text{AF}\,\text{cov}_p$ is realised only at release, our setting is the canonical continuous difference-in-differences of Callaway et al. (2024).[22] The level dose-response — the effect of a given coverage relative to zero coverage — is identified under the parallel-trends condition that proteins with different AF coverage would have followed common deposition trends absent AF2. The reported β is not this level effect but a linear summary of the *average causal response* to coverage: the slope of the dose-response curve. Reading that slope causally requires a stronger condition: parallel trends in potential outcomes at *every* dose, not only relative to zero coverage (Callaway et al., 2024; Baker et al., 2025). This condition concerns how a protein would have trended under coverage levels it never actually received — counterfactuals that are never in the data, pre or post, so no test computed from realised outcomes can verify it.

We estimate three main protein-year outcomes, one per margin. Two are deposition counts estimated by PPML, for which β is a log-rate shift, so $\exp(\hat{\beta} \cdot x) - 1$ gives the implied multiplicative change in the expected count for a protein with $\text{AF}\,\text{cov}_p = x$: *total depositions*, which capture overall activity, and *follow-up depositions* on the 22,578 proteins with at least one pre-2017 PDB structure, which isolate *deepening* — the choice to run

[22]The release is a common shock, not a staggered one: it reaches the whole proteome through a single process, so every protein shares the same treatment date, and the forbidden-comparison and negative-weight problems of staggered two-way fixed-effects designs do not arise. The remaining negative-weight concern is intrinsic to the continuous dose, which the dose-response estimators below address.

additional experiments on proteins already studied. We additionally estimate the number of depositions by categories based on content component, training value, and laboratory pipeline. We also examine the annual publication counts linked to a protein.

To explore frontier expansion, we look at an indicator for the first deposition on a protein, estimated on the at-risk sample of 507,917 proteins without a pre-2017 PDB deposition — a discrete-time hazard estimated by OLS as a linear probability model, in which proteins remain in the panel until and including the year of their first event. Here β is the post-AF change in the annual conditional probability that an at-risk protein is first deposited that year, per unit of $\text{AF}\,\text{cov}_p$ and in percentage points; with a pre-AF baseline hazard $\bar{h}_0 = 0.002$, the implied change for a fully covered protein is $\hat{\beta}/\bar{h}_0$. It isolates protein-level entry: the decision to experiment on a protein whose structural problem is unsolved. Such entry is necessary but not sufficient for frontier expansion; the distance interactions distinguish nearby fill-in from remote entry. The at-risk specification replaces α_p with a primary-Pfam family fixed effect $\gamma_{f(p)}$[23], adds $\text{AF}\,\text{cov}_p$ in level form (since the protein FE that absorbed it is gone), and includes the pre-2017 protein features $\mathbf{X}_p$ in level form alongside their $\times$ Post interactions, to control for the cross-sectional variation that the protein fixed effect would otherwise absorb.[24]

To distinguish frontier expansion from inward fill-in, we augment (12) with the triple interaction $\text{AF}\,\text{cov}_p \times \text{Post}_t \times D_p^{2016}$ and all corresponding lower-order terms, where D_p^{2016} is the tree-path distance to the nearest pre-2017 solved protein. This interaction is the reduced-form analogue of the model's cross-partial (3) — whether the AF coverage slope strengthens or weakens with remoteness — and appendix specifications repeat it with the gap G and the within-cluster solved share ρ^{cl}. Because D_p^{2016} is max-imputed for isolated proteins, which have no measurable tree neighbour, the triple is meaningful only on non-isolated proteins, where distance reflects a genuine separation from solved structures. We therefore include an $\text{isolated}_p \times \text{Post}_t$ term as a control, so that the distance triple is identified off the non-isolated variation rather than being driven by the imputed maximum distance assigned to isolated proteins.

Magnitude rows translate β into an interpretable effect — $\left(\exp(\hat{\beta}\,x) - 1\right) \times 100$ in percent for the PPML columns and $100\,\hat{\beta}\,x$ in percentage points for the LPM columns — evaluated at three coverage changes x: the full range ($x = 1$, zero to full coverage), one

[23]Proteins in the same family descend from a common evolutionary ancestor. Proteins are assigned to families using a range of sources, including protein family databases, sequence analysis tools, scientific literature and sequence similarity search tools. We use the Pfam family with the most UniProt members globally as the primary family (or the most "evolutionarily dominant" family) it belongs to. Around 3.5% of proteins have no family and are dropped from the sample.

[24]The two features that are mechanically zero on the at-risk sample by construction — log pre-2017 PDB count and log pre-2017 cryo-EM PDB count — are dropped from $\mathbf{X}_p$ in this specification.

cross-protein standard deviation ($x = \sigma = 0.190$), and the Q4–Q1 gap ($x = 0.391$). The latter two are computed once on the full cross-section of proteins and held fixed across all columns and tables, so magnitudes are comparable across specifications; distance-triple columns are evaluated at $D_p^{2016} = 0$. A separate *implied depositions per year* row prices the full-range effect against the pre-AF (2017–2021) annual baseline.

7 Results

7.1 Deepening Rather Than Frontier Expansion

We first measure reallocation at the protein level, distinguishing between extensive and intensive margins. A first deposition adds a new protein to the PDB (column 3), whereas a follow-up deposition adds another structure to a protein already there (columns 5–6).

AF2 reallocates structural-biology activity within the existing structural map rather than broadening it (Table 7). Moving up the AF coverage gradient, from no usable coverage to full coverage, the expected total deposition rate is 40.8% higher after 2021 (column 1) and the follow-up rate on proteins with at least one pre-2017 PDB is 63.8% higher (column 5), while the annual first-deposition hazard on previously unsolved proteins is 31.2% lower than its pre-AF baseline (column 3) — an estimate we read cautiously, since the falsification tests below indicate that part of this decline was already under way before AF2. Because coverage is concentrated near one, this full-range move reaches into a thin zero-coverage region; the same pattern holds over the interquartile range where coverage actually varies, with a top-quartile protein showing a 14.3% higher total and 21.3% higher follow-up deposition rate than a bottom-quartile one.

The pattern is consistent with the current-choice channels in Section 3. AF2's predicted structure can serve as a scaffold for state-specific, ligand-bound, and corroborating experiments, reducing the cost of follow-up work. At the same time, AF2 partially substitutes for the static-scaffold information that a first-ever experiment would itself yield, weakening the incentive to bring a previously unstudied protein into the PDB. The results indicate that substitution dominates on average for first deposition, while complementarity dominates for follow-up work. The heterogeneity results below show that scaffold-input relief is part, but not all, of the follow-up response. This is the aggregate pattern in Proposition 1: private experimental volume can expand even as activity shifts away from the first-ever structures that supply range-expanding anchors.

Figure 8 plots the dynamic version of the headline coefficients. The total and follow-up coefficients rise monotonically after 2021 — the follow-up coefficient reaches +0.85

log-points by 2025 and turns statistically positive by 2023, with the total coefficient on a flatter version of the same path (+0.61 by 2025). The first-deposition coefficient is negative throughout the post-AF window and statistically distinguishable from zero by 2023. The AF effect accumulates over time rather than reverting around the 2021 release, consistent with a persistent reorganisation of structural-biology effort rather than a one-off compositional shift.

Identification rests on parallel trends: conditional on the controls, proteins would have trended in parallel at *every* coverage level, not merely relative to zero coverage. This strong condition is not directly testable, because the falsification tests we can run use realised pre-period outcomes and so speak only to the weaker, level form. That level of evidence is nonetheless supportive. The cross-sectional gradient in raw outcomes is approximately flat in the pre-AF window (Figure 7), the year-by-AF coverage interaction is statistically zero on first-ever and follow-up depositions and only marginally positive on total depositions (Table 26), and observable characteristics show no differential pre-trends across coverage quartiles (Table 27). The Rambachan–Roth (2023) bounds widen the confidence intervals once deviations are allowed (Table 30, on the log-rate scale for the count outcomes), a randomisation-inference placebo places the true results far outside the chance distribution (Table 28). A 2016-event placebo on the cryo-EM–excluded sample confirms that the total and follow-up effects are not spurious (Table 29). For the first deposition, the same placebo returns a statistically significant negative coefficient, indicating that part of the inward movement in frontier entry was already underway before AF2 — a pre-existing drift as the field's most tractable targets were progressively solved. This is precisely the inherited data geography our analysis foregrounds, and AF2 reinforces it rather than originating it: the flat pre-AF coverage gradient (Table 26), the inward shift of the entry that remains, and the strong deepening response on the total and follow-up margins all point the same way and survive every falsification test.

We also rely on forward-engineered dose-response estimators that recover the strong condition without the negative weights a two-way fixed-effects slope can carry. The long-difference dose-response (Callaway et al., 2024) — which differences out the protein effect and regresses each outcome's pre/post change on AF coverage by unweighted OLS, a transparent linear summary rather than the density-weighted average causal response — confirms the direction (Table 25): the total and follow-up slopes are positive and significant, and the first-deposition slope reproduces the headline negative sign, though imprecise because the long difference of a rare extensive event carries little signal. The quartile specification (Table 36) shows the same monotone pattern. Because mean coverage is high and a genuine zero-coverage comparison group is thin, β is best read as a causal response

among treated proteins, with the trimmed sample (Table 40) probing that limited support.

7.2 Where Deepening Concentrates

In Table 7, we also explore the distance to the knowledge frontier, captured by the tree-path distance D_p^{2016} from p's cluster to the nearest cluster with a pre-2017 PDB structure to sharpen the distinction between inward fill-in and frontier expansion. A first deposition at large D_p^{2016} pushes the structural map outward: this is frontier expansion. A first deposition at small D_p^{2016} instead fills in the mapped interior, while follow-up depositions deepen knowledge by construction. The headline coefficients average over location; the distance interactions reveal whether the remaining entry moves inwards or outwards.[25]

On first deposition (column 4), the post-AF effect of AF coverage on the annual first-deposition hazard is $+0.4 - 0.4\, D_p^{2016}$ percentage points on a 0.2 p.p. pre-AF baseline. It is positive for a protein adjacent to a pre-2017-solved homologue, crosses zero near $D_p^{2016} \approx 1.0$, and becomes increasingly negative beyond. Because the median at-risk protein lies past the zero-crossing, the average first-deposition coefficient is negative. AF2 reallocates the entry that remains towards nearby homologues and away from the frontier.

The follow-up pattern is different. On total depositions (column 2), the post-AF effect on the log deposition rate is $+0.45 - 0.24\, D_p^{2016}$ and remains positive across the observed support. On follow-up depositions (column 6), the post-AF effect is $+0.34 + 0.37\, D_p^{2016}$ log points. The positive distance interaction is marginally significant, so the precise degree of amplification should be interpreted cautiously. The robust result is that deepening operates throughout the observed landscape and does not disappear when the surrounding sequence neighbourhood is sparsely solved.

The two-direction frontier gradient identifies different parameter regions of the current-choice model. For first deposition, the negative AF coverage-by-distance interaction is consistent with the region in which, for the relevant first-ever bundle b containing R,

$$s_{b,AF}(A_x, F_{xt}) > \eta_m \delta^g_{b,m,A}(A_x)\chi'(F_{xt}) :$$

[25] Because D_p^{2016} is built relative to the pre-2017-solved set, it means different things in the two samples. For an unsolved (at-risk) protein, it is genuine frontier remoteness — for proteins in as-yet-unsolved clusters, the inter-cluster distance to the nearest solved cluster. An already-solved protein, by contrast, lies in a cluster that already contains a solved member, so the within-cluster path applies (Appendix Step 6) and D_p^{2016} measures how sparsely the protein's immediate neighbourhood is solved rather than distance to the frontier. The follow-up×distance interaction (column 6) should therefore be read in that local sense, not as frontier remoteness in the same sense as the first-deposition interaction (column 4); the tree-anchored gap G and the within-cluster solved share are the natural frontier objects for already-solved proteins and deliver the same qualitative gradient (Appendix Table 21).

the substitution of AF2 for scaffold discovery is more strongly amplified by remoteness than residue-level or co-production cost relief. Under the monotone-selection condition in Proposition 6, an informative AF2 prediction shifts the remaining first-ever structures towards nearby homologues rather than towards remote anchors. For follow-up work, the positive interaction is consistent with the opposite parameter region. In particular, for mechanistic S/L components, it is the sign predicted when AF2's scaffold relief becomes more valuable as the local structural map thins and dominates any partial substitution. The same gradient holds when the distance triple is replaced by the tree-anchored gap G_p^{2016} or the within-cluster solved share, both individually and jointly (Appendix Table 21).

Table 22 verifies the pattern non-parametrically with three neighbourhood bins (dense, sparse, isolated) based on tertiles of distance to the nearest solved protein among non-isolated proteins, with proteins lacking a measurable similar neighbour pooled into the isolated bin. The bins reveal the heterogeneity hidden by the average. For the first deposition, the negative net effect masks a monotonic gradient: the AF effect is negative on isolated and sparse proteins and slightly positive on dense ones. For total depositions, the positive response is concentrated in non-isolated bins. For follow-up, the AF effect is positive in all three bins and statistically indistinguishable across them. The continuous and binned specifications, therefore, agree on the central conclusion: AF2 reduces frontier expansion while increasing deepening across the structural landscape. In the welfare interpretation of Section 3.4, this is where the directional wedge can become especially consequential: the private return to remote first-ever anchors falls precisely where their uninternalised range-expansion value can be largest.

7.3 The Content of Deepening and the Data Supplied to Later AI

The training-value outcomes also have a distinct interpretation in the model. First-in-cluster and structurally novel folds proxy for the range-expansion component of $\mathbf{T}_{g+1}^{exp}$, while conformations, ligand structures, and binding-affinity records proxy for mechanistic components. These regressions do not estimate the downstream weights $\boldsymbol{\omega}(z)$ or the missing reward in (9). They identify the direction in which AF2 reallocates the experimental ground-truth corpus, which is the empirical object required by Proposition 1.

Deepening takes three economically distinct forms in the model. A scaffold deposition extends or refines the existing scaffold (R or Q); a mechanistic deposition captures a new conformational state or biologically relevant ligand (S or L); a corroborating deposition re-measures the same residues without producing new structural information (C). The component-conditional setup wedge predicts that AF2's cost-side relief should often land

on mechanistic work, because pre-AF scaffold-input costs can gate state- and ligand-bound experiments on outside-pipeline proteins and AF2 supplies relief by providing a usable scaffold input. Scaffold work, which partly supplies the object AF2, faces cost relief and benefit-side substitution simultaneously, so the net response is ambiguous. Corroboration is also ambiguous: AF2 reduces the marginal informational value of an additional scaffold signal but can simultaneously reduce its cost (Proposition 5).

Table 8 reports the composition of follow-up deepening. Scaffold depositions rise by 23.2% (column 1), but the estimate is only marginally significant and smaller than the mechanistic and corroboration responses. Mechanistic depositions rise by 97.8% (column 2). This ordering is consistent with the cost-complementarity channel: mechanistic follow-up deposits often need a baseline scaffold but do not mainly produce one, so AF2 can relax a constraint while only partially competing with the experiment's own output. Appendix Table 16 reports the corresponding full-sample specification.

Corroborating depositions rise by 175.7% (column 3), the largest action-type response. This finding is not explained by a scaffold-input story alone. AF2 lowers the informational value of another generic scaffold measurement, but it can also lower the cost of validation and increase the value of verifying a consequential prediction for publication, regulation, or commercial use. The large corroboration coefficient is therefore consistent with the validation mechanism in Proposition 5. A competing, mechanical reading is that AF2-assisted molecular replacement makes it cheap to re-deposit structures that merely re-cover an existing scaffold, registering as corroboration without adding scientific content. We cannot fully separate this from genuine validation, though molecular replacement is an X-ray technique and the X-ray margin is statistically flat (Table 12), which makes cheap re-covering through that channel unlikely to be the main driver. Either way, the model treats corroboration as ambiguous (Proposition 5), and our "richer deepening" conclusion does not rest on it: it rests on the mechanistic response — new conformations and ligand-bound structures — whose scientific content is unambiguous and which the training-value results confirm.

The mechanistic action is also where AF2's induced depositions most directly change the experimental data supply available to next-generation predictors. AlphaFold3, AF-Multimer, RoseTTAFold2NA, and successor models extend the prediction target beyond static single-chain folds towards conformational ensembles, ligand binding, protein–nucleic-acid complexes, and proteins in sequence regions where co-evolutionary signal is shallow — dimensions on which AlphaFold2 is less directly informative. This is the empirical place where the virtuous-circle discussion in Section 2.3 closes: the kinds of conformational and ligand-related work AF2 induces are precisely the kinds of obser-

vations later systems seek to model. Table 9 reports the AF effect on five training-value outcomes capturing structural novelty (new conformation, first in cluster, novel structure) and biological richness (binding affinity, biological ligand).

On the novelty measures, AF2 raises the expected rate of new-conformation deposits by 41.3% (column 1) — the same S flag that underlies the mechanistic-action result above. By contrast, the first-in-cluster deposition rate falls by 48.8% (column 2), and the structurally novel-fold rate falls by 62.0% (column 3). The deposition stream deepens within already-solved sequence and fold neighbourhoods and becomes less informative about regions the model has yet to see. AF2 also raises biologically rich deposits. The biological-ligand deposition rate rises by 21.9% (column 5), while the binding-affinity deposition rate rises by 17.5% but is not statistically distinguishable from zero (column 4).

The implication for the experimental corpus available to next-generation structure prediction is two-sided. AF2 enriches the training corpus on conformational and ligand dimensions that next-generation models seek to model — exactly the mechanistic actions AF2's predicted scaffold complements rather than substitutes. It reduces the observed experimental corpus on dimensions next-generation models may also benefit from, and that are not supplied by AF2's own predictions: structurally novel folds and first-in-cluster entries both fall. The novelty AF2 amplifies is the discovery of new states of already-mapped proteins; the novelty it reduces is the discovery of new sequence families and new folds. The deposition stream available to successor systems, therefore, supplies relatively more evidence inside AlphaFold2's existing reach and relatively less evidence that broadens the experimental map. These estimates do not rank the social value of the resulting components or establish the net effect on later model performance. They identify the compositional change required by Proposition 1: valuable mechanistic deepening expands while the supply of range-expanding experimental anchors contracts.

7.4 Who Deepens

The same cost-relief logic suggests where the AF2 response should be strongest. The benefit-side substitution channel acts symmetrically across labs at the protein level, whereas the cost-side setup-relief channel is heterogeneous along pipeline incumbency and method. We treat the following comparisons as mechanism probes: pipeline status is tied directly to scaffold-input relief, while institutional type and experimental method explore where costs, rewards, and validation motives are largest.

7.4.1 Institutional heterogeneity

Table 11 reports AF2 coefficients by lab type (academic vs. industry) and within-protein margin (total, first-ever, follow-up). The model's cost-relief channel predicts that both lab types should benefit from AF2's scaffold supply on follow-up depositions; the proportional magnitudes depend on each type's pipeline breadth, with narrower pipelines concentrating relief on fewer proteins.

Both lab types respond positively to the intensive (follow-up) margin and negligibly to the first deposition. The expected academic follow-up rate rises by 72.4% (column 3) and the industry follow-up rate by 248.3% (column 6); the first-ever response is statistically zero or marginally negative for both types. The difference in follow-up responses is marginally significant ($p = 0.095$), with industry the larger of the two in proportional terms — consistent with industry's narrower commercial-target pipeline concentrating cost relief on a smaller set of proteins.

The absolute-volume increase is dominated by academic labs. The academic pre-period mean on follow-up ($\bar{Y} = 0.445$) is more than ten times the industry mean (0.031), so AF2's effect on aggregate structural-biology output is primarily an academic-volume effect, with a proportionally larger but smaller-in-volume industry contribution on top.

7.4.2 Method heterogeneity

Method heterogeneity follows from the same scaffold-relief logic: AF2 should matter more where the pre-AF scaffold prerequisite was expensive to satisfy de novo. For X-ray crystallography, non-AF molecular-replacement templates already supply scaffolds, so AF2's marginal relief is smaller. For cryo-EM, where de novo scaffold determination is comparatively costly, AF2's templates do more economic work. Table 12 estimates separate AF2 coefficients on deposition counts by method.

The expected cryo-EM deposition rate rises by 68% (column 2), while the X-ray crystallography rate rises by 34.4% but is not statistically distinguishable from zero (column 1). NMR and Other are imprecisely estimated. The method ranking is consistent with the component-conditional setup-wedge channel: AF2 expands the method for which scaffold construction and interpretation costs were high, while the response is imprecise for the method whose scaffold cost was already lower. This pattern complements Qian (2026)'s finding that very-high-confidence X-ray work falls post-AF. Her bin-level binary very-high design captures a region in which substitution dominates the net response, while our within-protein continuous-treatment specification captures a region in which complementarity is more important. Both findings are consistent with different balances across

techniques and margins.

7.4.3 Inside vs. outside the pipeline

The model distinguishes between outside and inside pipeline follow-up. Laboratories that had not previously worked on the focal protein paid a scaffold-solving setup cost pre-AF, and AF2 can now relieve it. Inside-pipeline laboratories, having previously solved the protein themselves, receive no such scaffold-prerequisite relief. Their follow-up work can nevertheless expand through variable-cost relief or reallocation away from static-scaffold actions (Proposition 4).

Table 10 reports the decomposition of deposition rates by pipeline status. First deposition shows the same negative response as in the headline (column 1); the decomposition splits only follow-up work, which a laboratory can do either inside or outside its existing pipeline. Both inside- and outside-pipeline follow-up rates rise for high AF coverage proteins after AF2's release. The inside-pipeline response is larger (+137.5% versus +58.1%). AF2's complementarity, therefore, extends beyond the relief of the outside-pipeline scaffold-input cost. The result is consistent with variable-cost relief and reallocation towards richer experiments on proteins laboratories already know, alongside the outside-pipeline entry channel. Table 24 confirms a related pattern: the total number of distinct laboratories depositing on a protein rises with AF coverage post-AF.

As a supplementary outcome, Appendix A.4.3 reports the AF2 response in annual UniProt-curated publication counts (Table 23 and Figure 15). The publication results are positive, but smaller than the deposition responses and closer to the existing publication-focused literature. We therefore treat them as supporting evidence rather than as a central margin in the main text.

7.5 Robustness Checks

The results in Section 7 survive an extensive set of robustness checks, collected in Appendix A.5 and summarised here by theme.

The headline pattern is unchanged across estimators and fixed effects. It holds when we estimate the count outcomes by OLS rather than PPML (Table 31) and when we add last-author-lab or organism fixed effects (Tables 32 and 33). Aggregating to the sequence-similarity cluster — the level at which homology spillovers operate — reproduces the protein-level result (Table 34).

The result does not depend on how we measure AF exposure or date the release. Replacing AF cov_p^{70} with mean pLDDT (Table 35) or with the residue share above alternative

pLDDT thresholds (Table 37) leaves the pattern intact, and replacing the continuous dose with coverage-quartile dummies yields a monotonic response (Table 36). AF2's accuracy was also published from CASP14 in November 2020, before the July 2021 release, so a differential anticipatory response by high-coverage proteins in 2020–2021 — beyond the common 2020 uptick that year fixed effects absorb — could threaten parallel trends. Re-estimating with the post-period defined as 2022–2025 dates treatment to the database expansion and places the 2020–2021 anticipation window in the pre-period; the deepening estimates are, if anything, slightly larger (Table 38).

The findings are also stable across samples and distance measures. They hold on to the full deposition record without the batch and SARS-CoV-2 filters (Table 39) and when we restrict to non-membrane proteins (Table 41), human proteins (Table 42), or single-protein depositions (Table 43). The frontier gradient is robust to alternative distance measures — the tree-anchored gap and the within-cluster solved share (Table 44) — and adding the distance triple to the action, training-value, pipeline, lab-type, and method breakdowns leaves each set of headline magnitudes intact (Tables 15, 17, 18, 19, and 20).

A final concern is common support. AF coverage is mechanically lower for orphan and frontier proteins with shallow alignments, so it thins at the highest frontier distances, and the genuine zero-coverage group is sparse; the continuous slope is therefore identified mainly over the interior of the coverage range. Trimming proteins with AF coverage below 0.2 (Table 40) and the quartile specification (Table 36), which compares adjacent coverage groups rather than extrapolating to zero, bound this concern.

8 Conclusion

This paper studies how a major AI advance reshapes the direction of scientific effort. We model protein research as a search on a tree with local spillovers and setup frictions. AF2 enters as a broad but heterogeneous public signal. It substitutes for experiments that reproduce scaffold information and complements experiments that use a scaffold as an input. The resulting choices change the distribution of experimental data inherited by later prediction systems. Because laboratories need not capture the full value of experimental ground truth for successor systems, the model identifies a directional data-supply externality: AI can increase private experimental activity while widening the private underweighting of remote frontier anchors relative to their data-supply value.

The empirical response is deepening rather than frontier expansion. Total deposition activity rises — by 41% for a fully covered protein — but this aggregate increase is a reallocation: first deposition on previously unsolved proteins does not rise, while the

expected follow-up deposition rate on already-mapped proteins rises by 64%. The remaining entry concentrates near solved homologues and becomes negative at greater frontier distances. Follow-up work rises across the observed distance support. Its composition is mechanistically rich: new conformational states and ligand-bound structures expand, as do corroborating structures. Cryo-EM depositions rise significantly, while the X-ray response is statistically indistinguishable from zero. Academic laboratories account for most of the added volume, while industry laboratories exhibit a larger proportional follow-up response from a smaller base.

These findings inform later AI selection. AF2 induces experiments on dimensions that a static single-chain predictor leaves unresolved, and that successor systems seek to learn from: conformations, interactions, and biologically relevant ligands. But it contracts first-in-cluster structures and structurally novel folds. More experimental data is, therefore, not equivalent to broader training data. Human choices can enrich a corpus within an AI system's existing reach while reducing the observations that would broaden the experimental map inherited by later systems. The estimates do not imply that AF2 lowers welfare or that mechanistic deepening is less valuable than frontier discovery. They show that aggregate growth in scientific activity is not sufficient to eliminate a directional allocation problem.

This distinction clarifies the sense in which AF2 can be placed in a moonshot frame. AF2 is not itself a moonshot programme. It is an infrastructure shock that changes the return to candidate experiments. A virtuous circle is possible: experimental structures train a predictor, the predictor lowers the cost of richer experiments, and those experiments supply data that later predictors can use. It is not automatic. The private experiments made attractive by the current tool need not include the deliberately distant anchors that supply observations outside the current tool's reliable range. As in Carnehl and Schneider (2025), a useful moonshot is neither an incremental step nor a disconnected marsshot: it creates a new anchor that subsequent researchers can productively fill in. As in Gans (2026), rewards attached only to immediately usable downstream knowledge can instead induce scientists to work to the AI boundary.

The policy question is consequently one of direction rather than volume alone. The model's missing reward, $(1 - \theta_\ell)\tau_{b,g}(x)$, provides a way to state the problem: public consortia, grant programmes, and commercial actors developing successor models may have reasons to reward experiments that carry high data-supply value or open new regions of protein space, even when their immediate downstream return is modest. The result does not prescribe a single programme or require a literal per-structure subsidy. It identifies the relevant margin for structural-genomics consortia, targeted grants, prizes, and data-

generation contracts. AI in science reorganises which questions are worth asking. Whether that reorganisation sustains a virtuous circle depends on whether institutions preserve incentives to generate observations outside the current model's field of view.

References

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, 2024, *630* (8016), 493–500.

Agrawal, Ajay, John McHale, and Alexander Oettl, "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 149–174.

Akdel, Mehmet, Douglas E V Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati et al., "A structural biology community assessment of AlphaFold2 applications," *Nature Structural & Molecular Biology*, 2022, *29* (11), 1056–1067.

Aldighieri, Pedro and Franco Malpassi, "Technological Breakthroughs and the Progress of Science: Evidence from Early Computers," *Available at SSRN 5806445*, 2025.

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk, "Self-Consuming Generative Models Go MAD," in "International Conference on Learning Representations (ICLR)" 2024.

Azoulay, Pierre, Alessandro Bonatti, Soomi Kim, and Ran Zhuo, "The shape of scientific discovery: Evidence from structural biology," 2026. Working paper, MIT and University of Michigan.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Difference-in-Differences Designs: A Practitioner's Guide," *Journal of Economic Literature*, 2025. Forthcoming.

Barbarin-Bocahu, Irène and Marc Graille, "The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report," *Acta Crystallographica Section D*, 2022, *78* (4), 517–531.

Berman, Helen M, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne, "The Protein Data Bank," *Nucleic Acids Research*, 2000, *28* (1), 235–242.

Brenner, Steven E, "Target selection for structural genomics," *Nature Structural Biology*, 2000, *7*, 967–969.

Bryan, Kevin A and Jorge Lemus, "The direction of innovation," *Journal of Economic Theory*, 2017, *172*, 247–272.

_ , _ , and Guillermo Marshall, "R&D competition and the direction of innovation," *International Journal of Industrial Organization*, 2022, *82*, 102841.

Burra, Prasad V, Yang Zhang, Adam Godzik, and Boguslaw Stec, "Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure," *Proceedings of the National Academy of Sciences*, 2009, *106* (26), 10505–10510.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Difference-in-Differences with a Continuous Treatment," Working Paper 32117, National Bureau of Economic Research 2024.

Campbell, Iain D, "AlphaFold and beyond: the revolution in structural biology," *Nature Reviews Molecular Cell Biology*, 2024, *25*, 1–2.

Carnehl, Christoph and Johannes Schneider, "A Quest for Knowledge," *Econometrica*, 2025, *93* (2), 623–659.

Chandonia, John-Marc and Steven E Brenner, "The impact of structural genomics: expectations and outcomes," *Science*, 2006, *311* (5759), 347–351.

_ and _ , "The impact of structural genomics: expectations and outcomes," *Science*, 2006, *311* (5759), 347–351.

EMBL-EBI, "AlphaFold predicts structure of almost every catalogued protein known to science," News release July 2022. 28 July 2022.

Fraser, James S, Henry van den Bedem, Avi J Samelson, P Therese Lang, James M Holton, Nathaniel Echols, and Tom Alber, "Accessing protein conformational ensembles using room-temperature X-ray crystallography," *Proceedings of the National Academy of Sciences*, 2011, *108* (39), 16247–16252.

Gans, Joshua S, "Growth in AI Knowledge," Working Paper 33907, National Bureau of Economic Research 2025.

_ , "A Quest for AI Knowledge," June 2026. Working Paper.

_ **and Avi Goldfarb**, "O-Ring Automation," Working Paper 34639, National Bureau of Economic Research 2026.

Hill, Ryan and Carolyn Stein, "Artificial Intelligence and Scientific Discovery: The Impact of AlphaFold," 2026. Working Paper.

Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero, "The Streetlight Effect in Data-Driven Exploration," NBER Working Paper 32401, National Bureau of Economic Research 2025. Revised June 2025.

Hopenhayn, Hugo and Francesco Squintani, "On the direction of innovation," *Journal of Political Economy*, 2021, *129* (7), 1991–2022.

Jaskolski, Mariusz, Zbigniew Dauter, and Alexander Wlodawer, "The PDB at the turn of the century: a time for change," *Acta Crystallographica Section D*, 2022, *78* (11), 1303–1307.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, 2021, *596* (7873), 583–589.

Keedy, Daniel A et al., "Crystal structures of a ligand at two temperatures reveal distinct binding modes via fragment screening," *eLife*, 2023, *12*, e90606.

Kim, Soomi, "Navigating the Rugged Data Landscape: The Impact of Data-Extrapolation Technologies on Knowledge Production," 2025. Working Paper, revisions requested at *Management Science*.

Kovalevskiy, Oleg, Juan Mateos-Garcia, and Kathryn Tunyasuvunakool, "AlphaFold two years on: validation and impact," *Proceedings of the National Academy of Sciences*, 2024, *121* (34), e2315002121.

Kühlbrandt, Werner, "The resolution revolution," *Science*, 2014, *343* (6178), 1443–1444.

Kuhlman, Brian and Philip Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, 2019, *20* (11), 681–697.

Liu, Jinfeng, Gaetano T Montelione, and Deepak Bhatt, "Target selection and deselection at the structural genomics level," *Bioinformatics*, 2007, *23* (10), 1267–1274.

Liu, Tiqing, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson, "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic Acids Research*, 2007, *35* (suppl_1), D198–D201.

Marsden, Russell L and Christine A Orengo, "Target selection for structural genomics: an overview," in "Methods in Molecular Biology," Vol. 426, Humana Press, 2008, pp. 3–25.

Martí-Renom, Marc A, Ashley C Stuart, András Fiser, Roberto Sánchez, Francisco Melo, and Andrej Šali, "Comparative protein structure modeling of genes and genomes," *Annual Review of Biophysics and Biomolecular Structure*, 2000, *29* (1), 291–325.

McCoy, Airlie J, Massimo D Sammito, and Randy J Read, "Implications of AlphaFold2 for crystallographic phasing by molecular replacement," *Acta Crystallographica Section D*, 2022, *78* (1), 1–13.

Peng, Kanwei, Zoran Obradovic, and Slobodan Vucetic, "Exploring bias in the Protein Data Bank using contrast classifiers," in "Pacific Symposium on Biocomputing" 2004, pp. 435–446.

Petsko, Gregory A, "An idea whose time has gone," *Genome Biology*, 2007, *8* (6), 107.

Piehl, Dennis W, Brinda Vallat, Ivana Truong, Habiba Morsy, Rusham Bhatt, Santiago Blaumann, Pratyoy Biswas, Yana Rose, Sebastian Bittrich, Jose M Duarte, Joan Segura, Chunxiao Bi, Douglas Myers-Turnbull, Brian P Hudson, Christine Zardecki, and Stephen K Burley, "rcsb-api: Python Toolkit for Streamlining Access to RCSB Protein Data Bank APIs," *Journal of Molecular Biology*, 2025, *437*, 168970.

Porta-Pardo, Eduard, Victoria Ruiz-Serra, Samuel Valentini, and Alfonso Valencia, "The structural coverage of the human proteome before and after AlphaFold," *PLoS Computational Biology*, 2022, *18* (1), e1009818.

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

Qian, Jiarui, "AI in the lab: AlphaFold2's impacts on human-produced knowledge," 2026. Working paper, University of Virginia.

Ren, Feng, Xiao Ding, Min Zheng, Mikhail Korzinkin, Xin Cai, Wei Zhu, Alexey Mantsyzov, Alex Aliper, Vladimir Aladinskiy, Zhongying Cao, Shanshan Kong, Xi Long, Bonnie Hei Man Liu, Yingtao Liu, Vladimir Naumov, Anastasia Shneyderman, Ivan V

Ozerov, Ju Wang, Frank W Pun, Daniil A Polykovskiy, Chong Sun, Michael Levitt, Alán Aspuru-Guzik, and Alex Zhavoronkov, "AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor," *Chemical Science*, 2023, *14* (6), 1443–1452.

Romer, Paul M, "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5, Part 2), S71–S102.

Rose, Yana, Jose M Duarte, Robert Lowe, Joan Segura, Chunxiao Bi, Charmi Bhikadiya, Li Chen, Alexander S Rose, Sebastian Bittrich, Stephen K Burley, and John D Westbrook, "RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive," *Journal of Molecular Biology*, 2021, *433* (11), 166704.

Rost, Burkhard, "Twilight zone of protein sequence alignments," *Protein Engineering, Design and Selection*, 1999, *12* (2), 85–94.

Saitou, Naruya and Masatoshi Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, 1987, *4* (4), 406–425.

Salentin, Sebastian, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder, "PLIP: fully automated protein–ligand interaction profiler," *Nucleic Acids Research*, 2015, *43* (W1), W443–W447.

Senior, Andrew W, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W R Nelson, Alex Bridgland et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, 2020, *577* (7792), 706–710.

Shaw, David E, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, and Willy Wriggers, "Atomic-level characterization of the structural dynamics of proteins," *Science*, 2010, *330* (6002), 341–346.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal, "AI models collapse when trained on recursively generated data," *Nature*, 2024, *631* (8022), 755–759.

Slabinski, Lukasz, Lukasz Jaroszewski, Leszek Rychlewski, Ian A Wilson, Scott A Lesley, and Adam Godzik, "XtalPred: a web server for prediction of protein crystallizability," *Bioinformatics*, 2007, *23* (24), 3403–3405.

Steinegger, Martin and Johannes Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, 2017, *35* (11), 1026–1028.

_ and _ , "Clustering huge protein sequence sets in linear time," *Nature Communications*, 2018, *9* (1), 2542.

Sun, Mengyi, Sukwoong Choi, and Yian Yin, "AI predictions and the expansion of scientific frontiers: Evidence from structural biology," 2026. bioRxiv preprint.

Suzek, Baris E, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, 2015, *31* (6), 926–932.

Terwilliger, Thomas C, David Stuart, and Shigeyuki Yokoyama, "Lessons from structural genomics," *Annual Review of Biophysics*, 2009, *38*, 371–383.

_ , Dorothee Liebschner, Tristan I Croll, Christopher J Williams, Airlie J McCoy, Billy K Poon, Pavel V Afonine, Robert D Oeffner, Jane S Richardson, Randy J Read, and Paul D Adams, "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination," *Nature Methods*, 2024, *21* (1), 110–116.

_ , Pavel V Afonine, Dorothee Liebschner, Tristan I Croll, Airlie J McCoy, Robert D Oeffner, Christopher J Williams, Billy K Poon, Jane S Richardson, Randy J Read, and Paul D Adams, "Accelerating crystal structure determination with iterative AlphaFold prediction," *Acta Crystallographica Section D*, 2023, *79* (3), 234–244.

_ , Randy J Read, Paul D Adams, Axel T Brunger, Pavel V Afonine, and Li-Wei Hung, "Can I solve my structure by molecular replacement? A new method to evaluate models for MR," *Acta Crystallographica Section D*, 2014, *70* (8), 2111–2119.

The Royal Swedish Academy of Sciences, "The Nobel Prize in Chemistry 2024," Press release October 2024. 9 October 2024.

The UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2025," *Nucleic Acids Research*, 2025, *53* (D1), D609–D617.

Tranchero, Matteo, "Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies," 2025. Working Paper.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon et al., "Highly accurate protein structure prediction for the human proteome," *Nature*, 2021, *596* (7873), 590–596.

van Kempen, Michel, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger, "Fast and accurate protein structure search with Foldseek," *Nature Biotechnology*, 2024, *42* (2), 243–246.

Varadi, Mihaly, Damian Bertoni, Pascal Borber, Chuming Chen, Davide Clementel, Corin Croda, Nadia Eddaoudi, Jasmine Gupta, Andras Hatos et al., "The impact of AlphaFold Protein Structure Database on the fields of life sciences," *Proteomics*, 2023, *23* (11), 2200128.

Velankar, Sameer, Jose M Dana, Julius Jacobsen, Glen van Ginkel, Paul J Gane, Jie Luo, Thomas J Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J Kleywegt, "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource," *Nucleic Acids Research*, 2013, *41* (D1), D483–D489.

Wang, Renxiao, Xueliang Fang, Yipin Lu, and Shaomeng Wang, "The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures," *Journal of Medicinal Chemistry*, 2004, *47* (12), 2977–2980.

Ward, Andrew B, Andrej Sali, and Ian A Wilson, "Integrative structural biology," *Science*, 2013, *339* (6122), 913–915.

Weitzman, Martin L, "Recombinant Growth," *Quarterly Journal of Economics*, 1998, *113* (2), 331–360.

Xu, Jinrui and Yang Zhang, "How significant is a protein structure similarity with TM-score = 0.5?," *Bioinformatics*, 2010, *26* (7), 889–895.

Yang, Jianyi, Ambrish Roy, and Yang Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Research*, 2013, *41* (D1), D1096–D1103.

Zhang, Yang, Ingrid A Hubner, Andriy K Arakaki, Eugene Shakhnovich, and Jeffrey Skolnick, "Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative," *Proceedings of the National Academy of Sciences*, 2006, *103* (8), 2920–2925.

Tables and Figures

Table 4: Notation guide

Symbol(s)	Meaning
Main text	
$\mathcal{P}, \mathcal{T}, D_{\mathcal{T}}, \mu$	Protein universe, sequence-similarity tree, tree distance, and density measure on the tree.
$x, y, z, p, t, \ell, m, b$	Protein targets (z is the downstream prediction target), date, laboratory, method, and component-bundle indices; $b \subseteq \mathcal{A}$, whose components map to scaffold (R/Q), mechanistic (S/L), and corroboration (C).
ν	Generic landscape measure, $\nu \in \{D, G, \rho^{cl}\}$.
$D_{xt}, G_{xt}, \rho^{cl}_{xt}, F_{xt}, \chi(F_{xt})$	Nearest-solved distance, two-anchor gap width, within-cluster PDB share, frontier position, and frontier-difficulty multiplier.
$J_t(x), I_{\ell xt}$	Public-solved indicator and lab-incumbency indicator.
A_x	Latent AF2 scaffold informativeness; empirical AF coverage proxies for A_x.
$s_b(A_x, F_{xt}), \delta^{sc}_m(A_x), \delta^{g}_{b,m}(A_x)$	AF-induced information substitution, scaffold-input relief, and variable or co-production cost relief.
$\eta_m, h^{sc}_{\ell xt}(b), \mathcal{R}_{b,m}$	Method scale, scaffold-input indicator, and total AF-induced cost relief for a bundle–method pair.
$\Delta\pi^{cur}_{\ell t}(x,b,m), \Delta^{E}_{\ell t}(x,b,m)$	Current AF payoff shock and direct AF payoff shock on a first-ever bundle containing R.
$\mathbf{T}^{exp}_{g+1}(z), M^{b}_{g}$, $\mathbf{\Gamma}_b, \mathcal{K}_b(z,x)$	Experimental input stock (scaffold/range, mechanistic, and validation), deposits by bundle, data-supply loadings, and propagation kernel.
$\boldsymbol{\omega}(z), \tau_{b,g}(x)$	Marginal social weights on directional experimental inputs, and marginal downstream data-supply value of bundle b at x.
$\pi^{cur}_{\ell t}(x,b,m), \pi^{priv}_{\ell t}(x,b,m)$, $\pi^{soc}_{\ell t}(x,b,m)$	Current post-AF payoff, private ranking payoff, and social ranking payoff.
θ_ℓ	Laboratory or funder weight on successor-system data-supply value.
$D^{2016}_p, G^{2016}_p, \mathrm{AF\,cov}_p$, $\mathrm{Post}_t, Y_{p,t}$	Frozen empirical nearest-solved distance, gap width, AF2-coverage proxy, post-release status, and the protein–year outcome.
Appendix (microfoundations and proofs)	
$\mathcal{S}_t, \mathcal{B}(x,t)$	Solved-protein set and feasible component-bundle set.
$\vartheta(x), \varsigma_x$	Latent tree attribute (Brownian) and AF2 scaffold-signal noise.
$y^{-}_t(x), y^{+}_t(x), W_t(x), \bar{u}_{xt}, u_{xt}$	Flanking solved anchors, within-cluster solved-template path, and local uncertainty measures.
$H_x, \mathcal{K}^{A}(x,y), \Phi$	Sequence-based information, local homology kernel, and AF2 information aggregator.
$B^{pre}_{\ell t}, c^{pre}_{\ell t}, \phi^{0}_{m}, \phi^{sc}_{m}$, $c^{var}_{b,m}(x), \kappa^{pre}_{\ell t}$	Pre-AF benefit and cost of an experiment, generic lab–protein entry cost, scaffold-input cost, variable cost, and pre-AF cost exposure.
$\mathcal{W}^{exp}_{g+1}$	Downstream value of the inherited experimental corpus.
$r_{\ell t}(x,b), V^{cont}_{t}(x,b), \zeta_\ell$	Future reward, continuation value for later human research, and the weight a laboratory or funder places on it.

Table 5: Protein features

Variable	Mean	Std. Dev.	Min	Max
Protein features ($N = 530,495$ proteins, cross-section)				
Year UniProt entry created	2005.427	5.302	1986	2016
Length (residues)	351.647	269.849	16	13,477
# Pfam domains	1.523	0.981	0	13
Share membrane	0.138	0.344	0	1
Share disease-associated	0.011	0.102	0	1
Scientific output pre-2017				
Year of first PDB deposition (if any)	2011.416	8.557	1976	2025
Share with any PDB (ever)	0.063	0.243	0	1
# PDB depositions	0.248	4.124	0	684
Share with PDB deposition	0.043	0.202	0	1
# Cryo-EM PDB depositions	0.019	0.812	0	87
# pub.	6.486	43.991	0	8,423
Structural landscape (as of 2016)				
Tree path distance to nearest PDB	0.935	0.442	0	1.773
Gap in tree	2.326	0.586	0	2.734
Share isolated	0.141	0.348	0	1
Share PDB in cluster	0.042	0.123	0	1
AF2 prediction				
AF2 coverage (pLDDT ≥ 90)	0.693	0.274	0	1
AF2 coverage (pLDDT ≥ 70)	0.868	0.19	0	1
AF2 coverage (pLDDT ≥ 50)	0.933	0.129	0	1
Mean pLDDT	87.673	10.511	26.181	98.754

Notes. This table reports the descriptive statistics of a cross-section of $N = 530{,}495$ proteins deposited in SwissProt before 2017 with AF coverage. Panel A describes protein-level features: the UniProt entry creation year, amino-acid length, number of Pfam domains, a membrane-protein indicator, and a disease-association indicator. Panel B describes the pre-2017 scientific-output measures: year of the first PDB deposition (conditional on having any), share of proteins with any PDB (ever) and with any pre-2017 PDB, count of pre-2017 PDB depositions, and count of pre-2017 PubMed citations. Panel C describes the position in the structural landscape as of 2016: the protein-level distance to the nearest protein based on NJ-tree sequence, the corresponding gap, and the share of isolated proteins and the protein's cluster members with a PDB. Panel D describes the AlphaFold2 prediction: coverages at various pLDDT thresholds and mean pLDDT.

Table 6: Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
Protein-year panel ($N = 530,495$ proteins between 2017 and 2025)				
# depositions	0.019	0.379	0	132
# first-ever depositions	0.002	0.047	0	1
# follow-up depositions	0.017	0.374	0	132
# publications	0.178	2.048	0	556
PDB depositions ($N = 93,522$ protein–deposition events between 2017 and 2025)				
Share follow-up	0.885	0.319	0	1
Share Scaffold	0.15	0.357	0	1
Share Mechanistic	0.789	0.408	0	1
Share Refinement	0.227	0.419	0	1
Share Corroborating	0.149	0.356	0	1
Share new conformation	0.182	0.386	0	1
Share first in cluster	0.063	0.242	0	1
Share novel structure	0.015	0.12	0	1
Share bio ligand	0.598	0.49	0	1
Share binding affinity	0.03	0.171	0	1
Share within pipeline	0.438	0.496	0	1
Unique structures ($N = 38,526$ structures between 2017 and 2025)				
Share X-ray	0.672	0.47	0	1
Share Cryo-EM	0.297	0.457	0	1
Share NMR	0.029	0.169	0	1
Share academic institution	0.623	0.397	0	1
Share company institution	0.052	0.203	0	1
Share government institution	0.007	0.069	0	1
Share facility	0.098	0.238	0	1
Share other/unknown	0.001	0.03	0	1
Labs ($N = 9,119$ labs)				
Share incumbent (≥ 1 pre-2017 deposit)	0.367	0.482	0	1
# depositions	9.087	44.319	1	2,220
# proteins	5.352	16.815	1	481
Share within pipeline	0.202	0.302	0	1

Notes. Sample restricted to SwissProt proteins created before 2017 and with non-missing AF2 coverage. Panel A reports statistics for the protein $\times$ year panel covering 2017–2025. Panel B reports statistics at the level of protein-PDB depositions with release year between 2017 and 2025. Panel C reports statistics at the unique PDB deposition level. Panel D describes labs.

Table 7: Main results

	(1)	(2)	(3)	(4)	(5)	(6)
	# depositions		First-deposition		# follow-up	
Post × AF cov	0.342***	0.445***	-0.001***	0.004***	0.494***	0.339***
	(0.094)	(0.120)	(0.000)	(0.001)	(0.123)	(0.122)
Post × Dist.		0.361***		0.004***		-0.204
		(0.097)		(0.001)		(0.171)
Post × AF cov × Dist.		-0.235**		-0.004***		0.370*
		(0.112)		(0.001)		(0.209)
Year FE	X	X	X	X	X	X
Protein FE	X	X			X	X
Family FE			X	X		
Pre-AF features			X	X		
Pre-AF features × Post	X	X	X	X	X	X
N	171,135	171,135	4,365,071	4,365,071	72,261	72,261
R^2 / $\tilde{R}^2$	0.460	0.460	0.021	0.021	0.532	0.532
N proteins	530,495	530,495	507,917	507,917	22,578	22,578
N structures	92,930	92,930	14,370	14,370	68,001	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.017	0.002	0.002	0.305	0.305
Effect (full range)	+40.8%	+56.0%	-0.1 p.p.	+0.4 p.p.	+63.8%	+40.3%
Effect (1 SD)	+6.7%	+8.8%	-0.0 p.p.	+0.1 p.p.	+9.8%	+6.6%
Effect (Q4 vs Q1)	+14.3%	+19.0%	-0.0 p.p.	+0.1 p.p.	+21.3%	+14.2%

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Cols (1)–(2) PPML on the annual number of depositions on the balance panel; cols (3)–(4) LPM/discrete-time hazard for the first-deposition indicator on the at-risk sample on proteins without any PDB before 2017; cols (5)–(6) PPML on the annual number of follow-up depositions on the intensive sample (proteins with at least one PDB before 2017). All columns include year FE, plus the pre-AF features interacted with Post_t. Cols (1)–(2) and cols (5)–(6) add protein fixed effects; cols (3)–(4) add the pre-AF features in levels (dropping log pre-2017 PDB count and log pre-2017 cryo-EM PDB count as both are mechanically zero on the at-risk sample) and primary-Pfam family FE. Cols (2), (4), and (6) additionally include a triple interaction with the pre-2017 distance to the nearest solved protein, all corresponding lower-level interaction in addition to an isolated $\times \text{Post}_t$ interaction (a control entered so the distance triple identifies off non-isolated proteins, not itself a triple). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8: Composition of depositions by content components

	(1) # Scaffold	(2) # Mechanistic	(3) # Corrob.
AF cov × Post	0.209*	0.682***	1.014**
	(0.117)	(0.163)	(0.484)
Year FE	X	X	X
Protein FE	X	X	X
Pre-AF features × Post	X	X	X
N	64,935	65,556	23,859
$\tilde{R}^2$	0.134	0.582	0.563
N proteins	22,578	22,578	22,578
N structures	68,001	68,001	68,001
$\bar{Y}_{t\leq 2021}$	0.156	0.528	0.086
Effect (full range)	+23.2%	+97.8%	+175.7%
Effect (1 SD)	+4.3%	+14.6%	+22.4%
Effect (Q4 vs Q1)	+10.1%	+36.8%	+59.3%

Notes. This table reports the coefficients from Equation (12) estimated by PPML on the annual PDB depositions by action class between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. The outcomes are the annual number of scaffold, mechanistic and corroboration follow-up depositions on the intensive sample (proteins with at least one PDB before 2017). AF cov_p is the share of a protein's residues predicted at pLDDT $\geq$ 70 and Post_t is an indicator for years after 2021. All columns include year FE, plus the pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post_t. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 9: Training value

# depositions by type:	(1)	(2)	(3)	(4)	(5)
	Novelty			Biological richness	
	New conformation	First in cluster	Novel structure	Binding affinity	Bio ligand
AF cov × Post	0.346***	−0.670***	−0.968***	0.162	0.198*
	(0.103)	(0.137)	(0.320)	(0.365)	(0.113)
Year FE	X	X	X	X	X
Protein FE	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X
N	115,596	52,587	8,532	7,872	108,090
$\tilde{R}^2$	0.123	0.005	0.084	0.260	0.407
N proteins	530,495	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.003	0.001	0.000	0.001	0.010
Effect (full range)	+41.3%	-48.8%	-62.0%	+17.5%	+21.9%
Effect (1 SD)	+6.8%	-11.9%	-16.8%	+3.1%	+3.8%
Effect (Q4 vs Q1)	+14.5%	-23.0%	-31.5%	+6.5%	+8.0%

Notes. This table reports the coefficients from Equation (12) on counts of depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by training value: Col (1) new conformation - TM $\geq$ 0.85 cluster never previously observed for this protein, Col (2) first in cluster based on MMseqs2, Col (3) novel structure = 1 - best TM to any prior PDB above 0.5, Col (4) binding affinity $\geq$ 1 Kd/Ki/IC50 record, and Col (5) Bio ligand $\geq$ 1 non-additive ligand. AF cov_p is the share of a protein's residues predicted at pLDDT $\geq$ 70 and Post_t is an indicator for years after 2021. All columns are estimated by PPML and include protein and year FE in addition to the pre-AF features interacted with Post. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 10: Heterogeneity by laboratory pipeline

	(1) First-deposition	(2) # New to lab	(3) # Inside
AF cov × Post	-0.001^{***}	0.458^{***}	0.865^{***}
	(0.000)	(0.137)	(0.174)
Year FE	X	X	X
Protein FE		X	X
Family FE	X		
Pre-AF features	X		
Pre-AF features × Post	X	X	X
N	4,365,071	59,706	28,548
R^2 / $\tilde{R}^2$	0.021	0.464	0.417
N proteins	507,917	22,578	22,578
N structures	14,370	68,001	68,001
$\bar{Y}_{t\leq 2021}$	0.002	0.182	0.099
Effect (full range)	-0.1 p.p.	+58.1%	+137.5%
Effect (1 SD)	-0.0 p.p.	+9.1%	+17.8%
Effect (Q4 vs Q1)	-0.0 p.p.	+19.6%	+40.2%

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. AF cov_p is the share of a protein's residues predicted at pLDDT $\geq$ 70 and Post_t is an indicator for years after 2021. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Col (1) LPM/discrete-time hazard for the first-deposition indicator on the at-risk sample on proteins without any PDB before 2017; col (2) / (3) PPML on the annual number of depositions new to the lab's pipeline / inside the lab's pipeline on the intensive sample (proteins with at least one PDB before 2017). All columns include year FE, plus the pre-AF features interacted with Post_t. Cols (2)–(3) add protein fixed effects, while col (1) adds the pre-AF features in levels and primary-Pfam family FE (dropping log pre-2017 PDB count and log pre-2017 cryo-EM PDB count as both are mechanically zero on the at-risk sample). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 11: Heterogeneity by laboratory type

	(1)	(2)	(3)	(4)	(5)	(6)
	Academic			Industry		
	# depositions	First-deposition	# follow-up	# depositions	First-deposition	# follow-up
AF cov × Post	0.446***	−0.001***	0.545***	0.891**	−0.000	1.248***
	(0.137)	(0.000)	(0.180)	(0.355)	(0.000)	(0.393)
Year FE	X	X	X	X	X	X
Protein FE	X		X	X		X
Family FE		X			X	
Pre-AF features		X			X	
Pre-AF features × Post	X	X	X	X	X	X
N	140,508	4,365,071	61,425	13,752	4,365,071	10,224
R^2 / $\bar{R}^2$	0.533	0.019	0.591	0.313	0.007	0.328
N proteins	530,495	507,917	22,578	530,495	507,917	22,578
N structures	92,930	14,370	68,001	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.024	0.002	0.445	0.001	0.000	0.031
Effect (full range)	+56.2%	-0.1 p.p.	+72.4%	+143.8%	-0.0 p.p.	+248.3%
Effect (1 SD)	+8.8%	-0.0 p.p.	+10.9%	+18.4%	-0.0 p.p.	+26.7%
Effect (Q4 vs Q1)	+19.0%	-0.0 p.p.	+23.7%	+41.7%	-0.0 p.p.	+62.9%
$\hat{\beta}_{\text{acad}} - \hat{\beta}_{\text{ind}}$ (full, PPML)			−0.445 (0.370); $p = 0.229$			
$\hat{\beta}_{\text{acad}} - \hat{\beta}_{\text{ind}}$ (at-risk, LPM)			−0.001*** (0.000); $p = 0.002$			
$\hat{\beta}_{\text{acad}} - \hat{\beta}_{\text{ind}}$ (intensive, PPML)			−0.703* (0.421); $p = 0.095$			

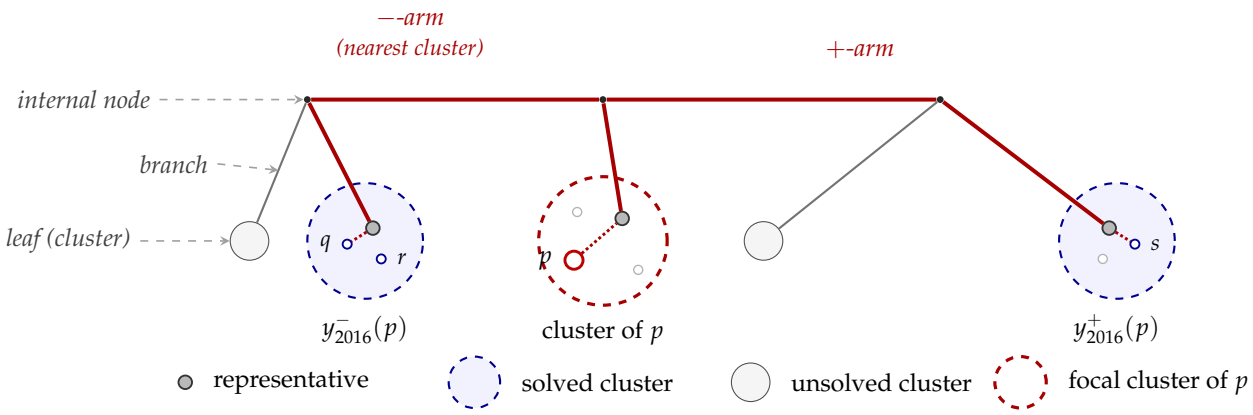
Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by lab type (academic cols (1)–(3) and industry cols (4)–(6)). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Cols (1) and (4) PPML on the annual number of depositions; Cols (2) and (5) LPM/discrete-time hazard for the first-deposition indicator on the at-risk sample on proteins without any PDB before 2017; Cols (3) and (6) PPML on the annual number of follow-up depositions on the intensive sample (proteins with at least one PDB before 2017). All columns include year FE, plus the pre-AF features interacted with Post_t. Cols (1), (3), (4), and (6) add protein fixed effects, while cols (2) and (5) adds the pre-AF features in levels and primary-Pfam family FE (dropping log pre-2017 PDB count and log pre-2017 cryo-EM PDB count as both are mechanically zero on the at-risk sample). The bottom rows report three sample-specific stacked tests of column equality between the academic and industry coefficients (stacked PPML on the full and intensive samples; stacked LPM on the at-risk sample with the labtype-specific binary indicator): each row uses the long (protein-year-labtype) panel for the corresponding sample with labtype-specific FE and labtype-specific slopes on every right-hand-side term; the labtype interaction on $\text{AF cov}_p \times \text{Post}_t$ recovers $\hat{\beta}_{\text{ind}} - \hat{\beta}_{\text{acad}}$ (log scale for PPML rows, probability scale for the LPM row; reported with sign flipped, protein-clustered SE). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 12: Heterogeneity by experimental method

	(1) # X-ray	(2) # Cryo-EM	(3) # NMR	(4) # Other
AF cov × Post	0.296	0.519***	−0.422	1.933
	(0.196)	(0.119)	(0.347)	(1.278)
Year FE	X	X	X	X
Protein FE	X	X	X	X
Pre-AF features × Post	X	X	X	X
N	95,409	95,139	8,199	432
$\tilde{R}^2$	0.490	0.602	0.122	0.401
N proteins	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.016	0.017	0.000	0.000
Effect (full range)	+34.4%	+68.0%	-34.4%	+591.1%
Effect (1 SD)	+5.8%	+10.4%	-7.7%	+44.3%
Effect (Q4 vs Q1)	+12.3%	+22.5%	-15.2%	+112.9%

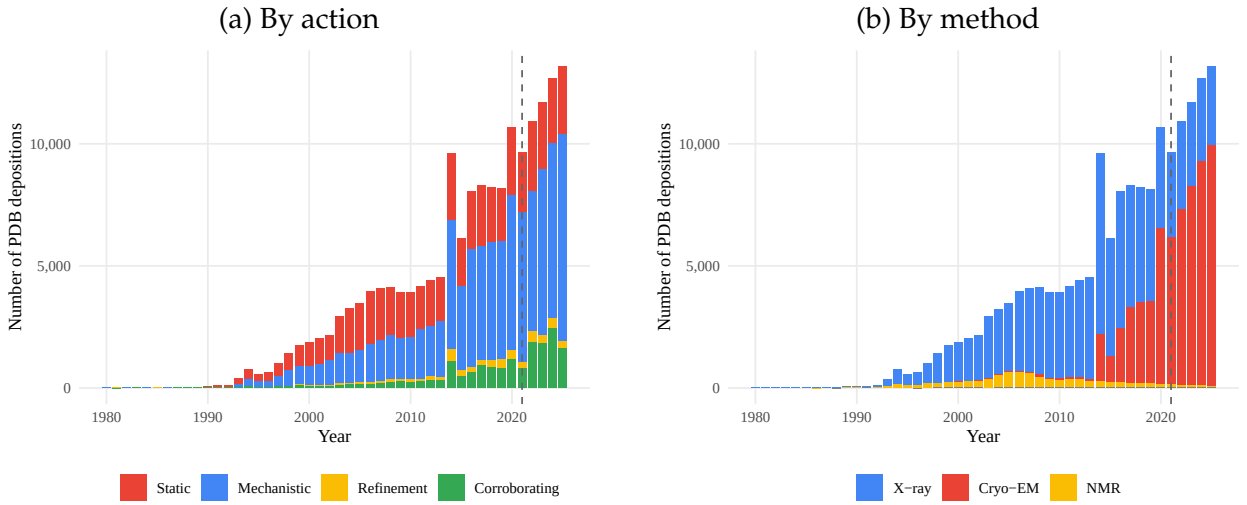
Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by method: X-ray (col 1), cryo-EM (col 2) and NMR (col 3). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. All columns are estimated by PPML and include protein and year FE, plus the pre-AF features interacted with Post_t. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure 2: Tree-path distance and gap to the nearest pre-2017-solved proteins: an illustration



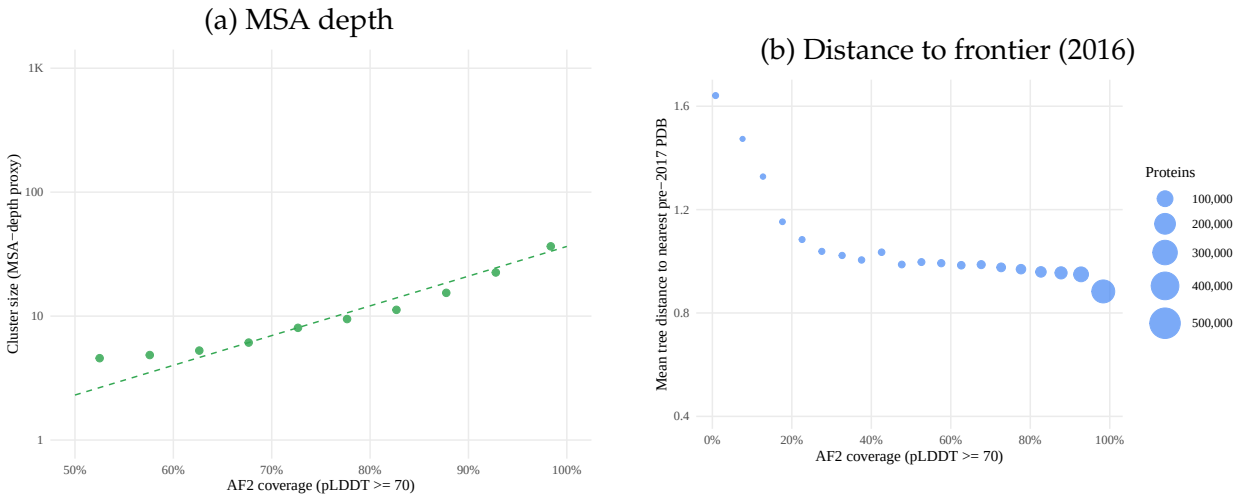
Notes. Stylised illustration of the protein-level NJ tree distance and gap. **Clusters.** Each leaf of the tree is an MMseqs2 cluster of proteins, grouped at $\geq$ 50% identity and $\geq$ 80% coverage of the shorter sequence (Steinegger and Söding, 2017). Within each cluster, the representative is the longest sequence, returned by MMseqs2's greedy set-cover step; longer sequences contain more k-mers and recruit short homologues more sensitively, the same convention used by UniRef and Linclust (Suzek et al., 2015; Steinegger and Söding, 2018). Solved members appear in blue, unsolved members in white, and the focal protein p in red. The focal cluster is unsolved (otherwise p would itself belong to a solved cluster). **Tree.** Neighbour-joining tree (Saitou and Nei, 1987) on cluster representatives. Each branch has length $D_{ij} = 1 - \text{pident}_{ij}/100$ (e.g. $D = 0.30$ corresponds to 70% identity), and each internal (Steiner) node is an inferred joining point under the binary NJ topology. $y^-_{2016}(p)$ and $y^+_{2016}(p)$ are the nearest pre-2017-solved clusters flanking p's cluster, with labelled solved members q and s. **Measures.** Solid red lines mark cluster-level tree branches, dotted red lines mark within-cluster spokes from a member to its representative ($d_{\text{intra}}(x) = 1 - \text{pident}_{x,\text{rep}}/100$). Each arm runs from p to a solved protein — the minus-arm to q, the plus-arm to s — beginning with the intra-cluster spoke $p \to \text{rep}$, traversing the cluster-level tree, and ending with the intra-cluster spoke from the destination representative to its solved member. The distance to the nearest solved protein is $D_p^{2016} = \text{minus-arm} = d_{\text{intra}}(p) + d_-^{\text{tree}} + d_{\text{intra}}(q)$; the gap is $G_p^{2016} = \text{minus-arm} + \text{plus-arm} = 2\,d_{\text{intra}}(p) + d_-^{\text{tree}} + d_+^{\text{tree}} + d_{\text{intra}}(q) + d_{\text{intra}}(s)$, with $d_{\text{intra}}(p)$ entering twice because each arm begins at p. If the focal cluster itself contains a solved member, the minus-arm short-circuits to that member.

Figure 3: Annual PDB depositions by action category and method



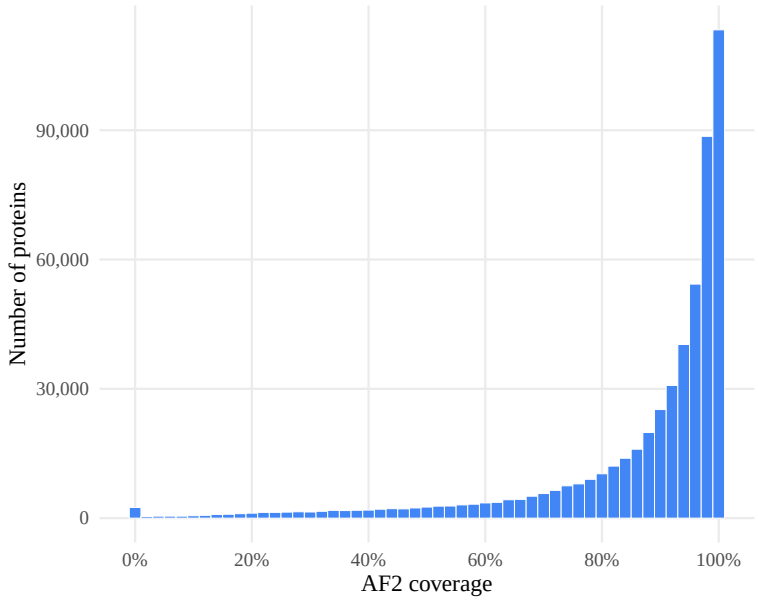
Notes. Figure (a) presents the number of PDB depositions by action category, and Figure (b) presents the number of PDB depositions by method (X-ray, cryo-EM and NMR).

Figure 4: AF2 coverage and its biological inputs



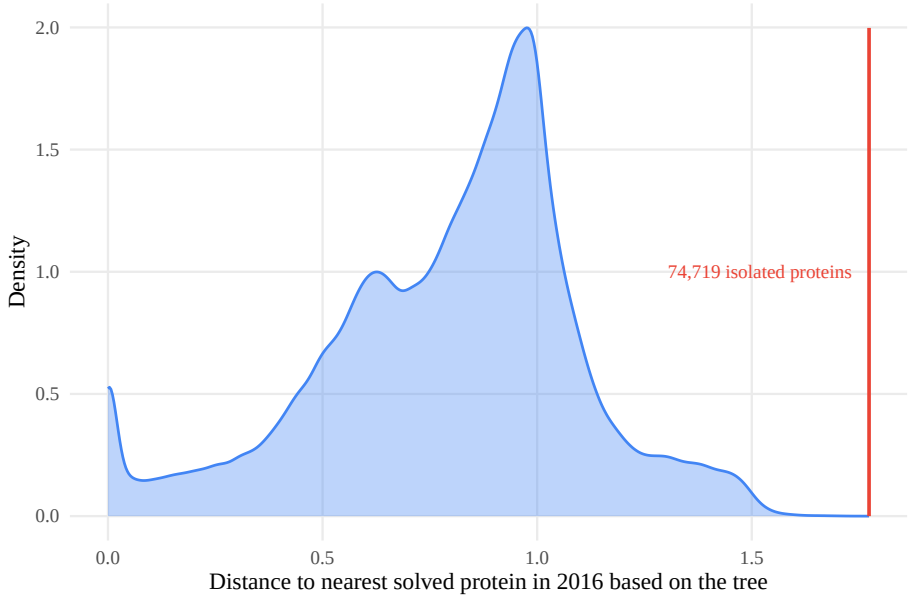
Notes. Figure (a) plots the protein-level cross-sectional scatters of AF coverage at pLDDT $\geq$ 70 against MSA depth, proxied by the number of SwissProt entries in the protein's MMseqs2 cluster ($\geq$ 50% identity, $\geq$ 80% coverage). Figure (b) plots the protein-level cross-sectional scatters of AF coverage at pLDDT $\geq$ 70 against 2016 sequence-tree distance to the nearest cluster with a pre-2017 PDB.

Figure 5: Distribution of AF2 coverage



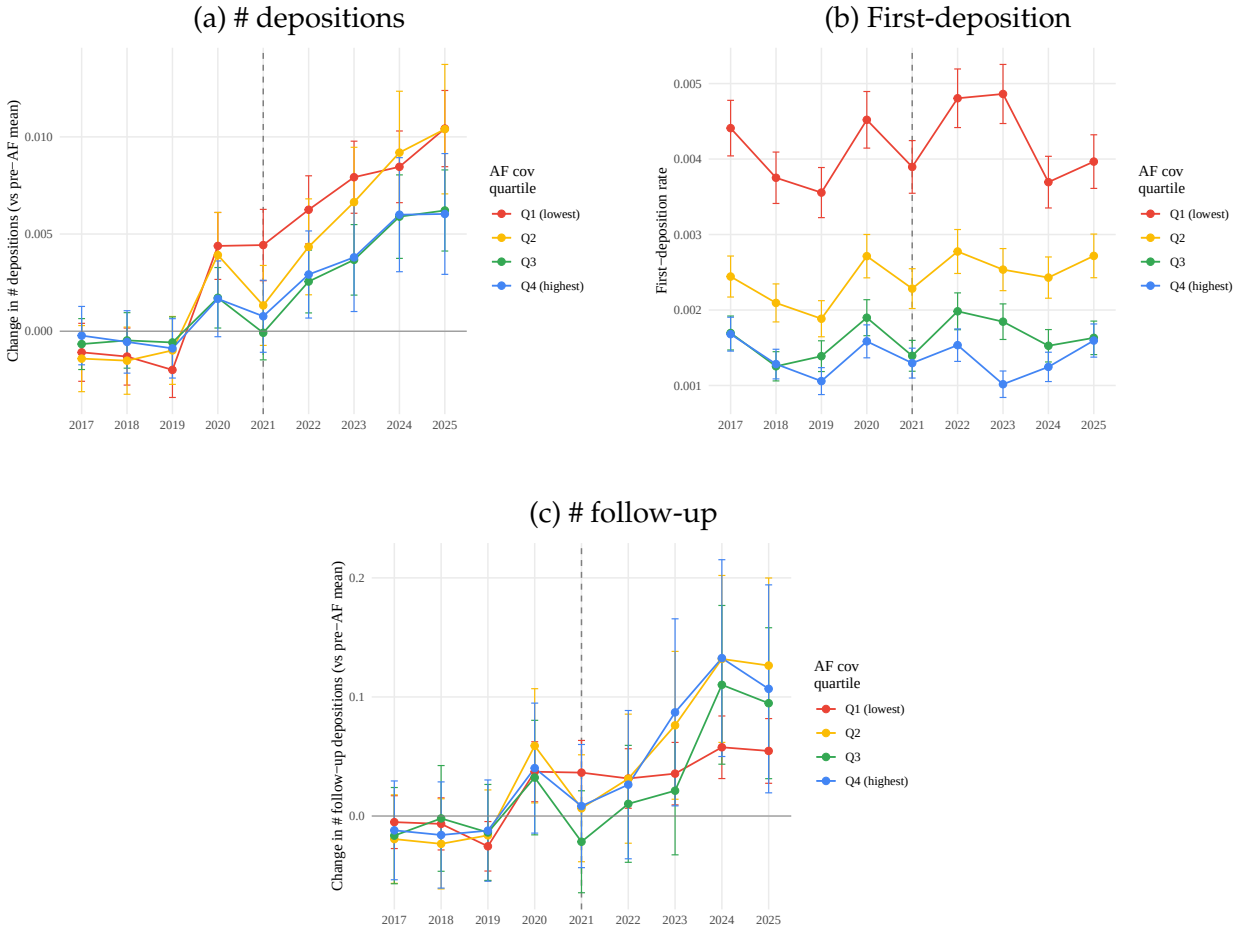
Notes. The figure shows the cross-protein distribution of AF2 coverage, defined as the share of residues with pLDDT ≥ 70.

Figure 6: Distribution of distance to nearest pre-2017 solved protein



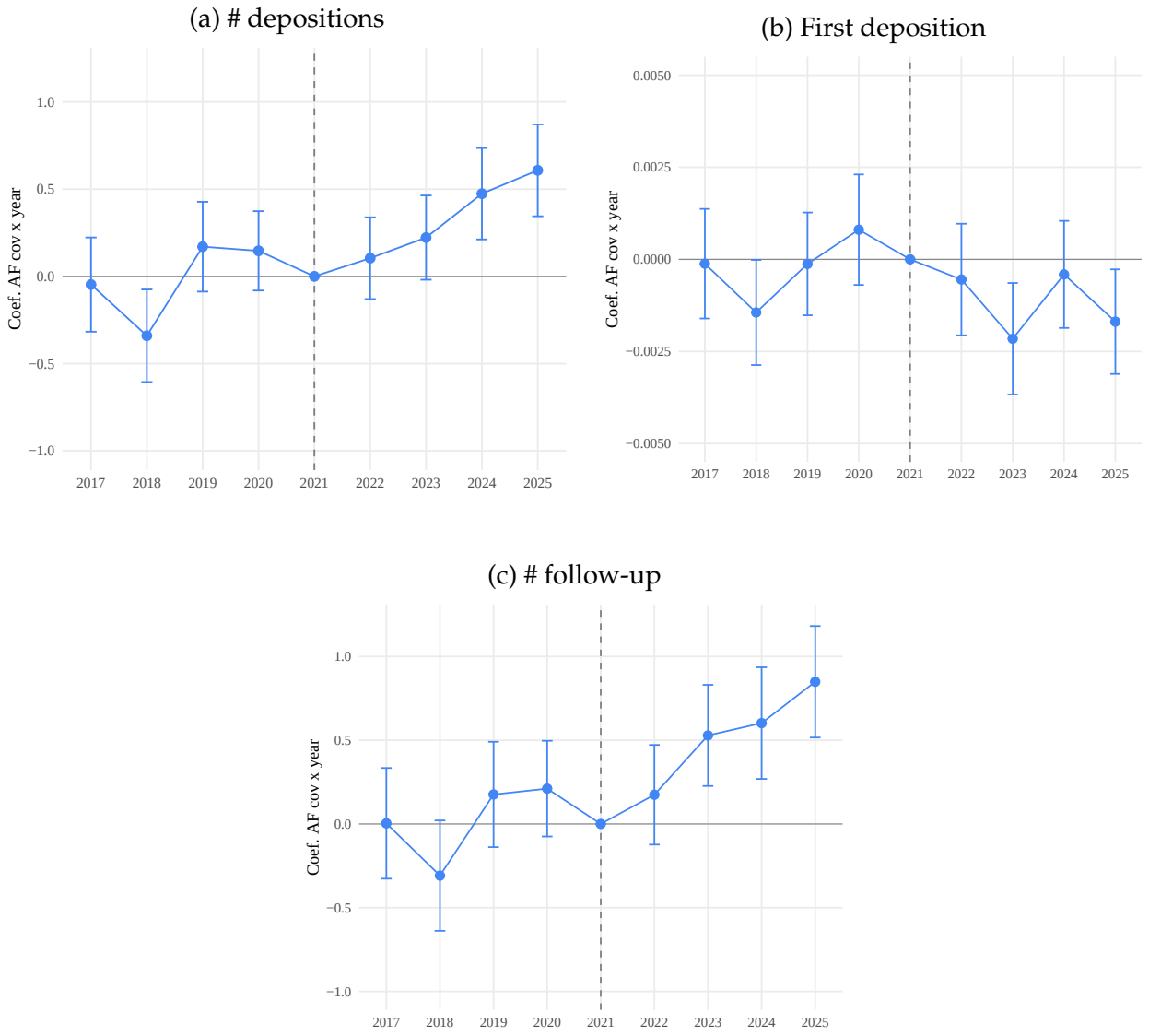
Notes. Each observation is a SwissProt protein in the 2017 cross-section ($N = 530{,}495$). The horizontal axis is D_p^{2016}, the tree-path distance from p's MMseqs2 cluster to the nearest cluster containing a protein with at least one PDB structure deposited on or before 31 December 2016, computed on the 2016 neighbour-joining sequence-similarity tree. The blue density covers the 455,776 non-isolated proteins, for which a tree path to a pre-2017-solved cluster exists. The red vertical line marks the 74,719 isolated proteins, whose cluster shares no detectable sequence similarity ($\geq 30\%$ pairwise identity) with any other cluster; their distance is max-imputed to the observed sample maximum. Lower D_p^{2016} corresponds to densely-mapped sequence space and higher D_p^{2016} to the unsolved frontier.

Figure 7: Deposition rates by AlphaFold coverage quartile



Notes. The figure plots outcome means per protein–year by AlphaFold coverage quartile (Q1 = lowest, Q4 = highest), where AF coverage is the share of a protein's residues predicted at pLDDT $\geq$ 70. Panel (a) reports the mean number of PDB depositions on the full sample of SwissProt proteins. Panel (b) reports the first-deposition rate on the survival-filtered at-risk sample of proteins without any PDB before 2017, kept until and including the year of first deposition. Panel (c) reports the mean number of follow-up depositions on the intensive sample of proteins with at least one PDB before 2017. Panels (a) and (c) subtract each quartile's pre-AF (2017–2020) mean from every year so that within-quartile post-AF deviations are comparable across quartiles; panel (b) is plotted in levels to display the cross-sectional gap between high- and low-AF coverage proteins. The vertical dashed line marks AlphaFold2's 2021 public release. Error bars are $\pm 1.96 \cdot$ SE.

Figure 8: Event study



Notes. Year-by-year coefficients $\hat{\delta}_t$ on $\text{AF cov}_p \times \mathbf{1}\{\text{year} = t\}$, with 2021 as the omitted reference. Panel (a): total depositions per protein–year (PPML, with protein and year fixed effects). Panel (b): first-deposition indicator on the survival-filtered at-risk sample (SwissProt proteins with no PDB before 2017, kept until and including the year of first deposition; LPM with primary-Pfam family and year fixed effects, AF coverage and the seven pre-AF features in level form alongside their year-by-year interactions; log pre-2017 PDB count and log pre-2017 cryo-EM PDB count are mechanically zero on at-risk and dropped by collinearity). Panel (c): follow-up depositions on the intensive sample (proteins with at least one PDB before 2017; PPML, protein and year fixed effects). All three panels include year-by-year interactions of the seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, log pre-2017 cryo-EM PDB count). Coefficients are in probability-point units per unit of AF coverage in panel (b) and in log-rate units in panels (a) and (c). Error bars are 95% confidence intervals from standard errors clustered at the protein level. The dashed vertical line marks AF2's July 2021 public release.

Appendix

Table of contents

A.1 Model Microfoundations and Proofs

This appendix gives a foundation for the reduced-form model in Section 3. The purpose is not to make a single latent state stand in for protein science. Rather, it shows how the frontier measures arise from local learning on a tree, how multifaceted experimental outputs can be represented as component bundles, how AF2 changes private experimental choice, and how the resulting experiments alter the directional data supply inherited by later prediction systems. Each formal result is followed immediately by its proof.

A.1.1 A local tree foundation

Let $(\mathcal{T}, D_{\mathcal{T}})$ be a rooted protein-similarity tree with edge lengths, and embed each protein $x \in \mathcal{P}$ at a node of that tree. For any two proteins $x, y \in \mathcal{P}$, let $d(x, y) = D_{\mathcal{T}}(x, y)$ be the length of the unique path joining them. A latent protein attribute $\vartheta(x)$ follows a Brownian process on the tree: conditional on a value at a parent node, increments on distinct branches are independent and an edge of length h contributes variance $\sigma^2 h$. The attribute can be interpreted as a scalar summary of one local dimension of structural knowledge.

At date t, experiments have produced structural anchors $\mathcal{S}_t \subset \mathcal{P}$. Consider the canonical two-anchor case in which a target x lies on the path between two solved anchors p and c, the anchor values are observed without error, and no additional anchor or root information is informative for x after conditioning on those two anchors. Let $X = d(p, c)$ and $d = d(p, x)$. Conditional on the anchor values, the remaining variance at x is

$$u^{\text{scaf}}(d, X) = \sigma^2 \frac{d(X - d)}{X}. \tag{13}$$

This is the variance of a Brownian bridge. To see it, represent the path from p to c as a Brownian motion on $[0, X]$ with observed endpoints. Conditional on the endpoint at X, the variance at location d is

$$\sigma^2 d - \frac{\text{Cov}(B(d), B(X))^2}{\text{Var}(B(X))} = \sigma^2 d - \frac{(\sigma^2 d)^2}{\sigma^2 X} = \sigma^2 \frac{d(X - d)}{X}.$$

The expression is exact only in this local on-path two-anchor setting.

The exact two-anchor formula differs when the target sits on a side branch. Let a be the attachment point of x to the path between p and c, let $r = d(x, a)$, let $d_p = d(p, a)$, let $d_c = d(a, c)$, and let $X = d_p + d_c = d(p, c)$. Brownian increments on the side branch are independent of the bridge on the anchor path, so

$$\text{Var}(\vartheta(x) \mid \vartheta(p), \vartheta(c)) = \sigma^2 r + \sigma^2 \frac{d_p d_c}{d_p + d_c}. \tag{14}$$

The first term is the variance of the branch from the attachment point to the target; the second is the Brownian-bridge variance at the attachment point. With more than two informative anchors, the posterior variance is obtained by Gaussian conditioning on the full vector of anchors and depends on the complete anchor geometry. The empirical frontier index is therefore a local reduced-form index, not the exact posterior variance of an arbitrary multi-anchor tree.

Lemma 1 (Local gaps)**.** *For a fixed gap length X, the posterior variance in (13) is increasing in d for $d < X/2$, decreasing in d for $d > X/2$, and maximised at $d = X/2$. Holding a target's relative position fixed, posterior variance increases with the length of its unresolved gap.*

Proof. The derivative of $u^{\text{scaf}}(d, X)$ with respect to d is $\sigma^2(X - 2d)/X$. For a fixed relative position $\alpha = d/X$, $u^{\text{scaf}}(\alpha X, X) = \sigma^2 \alpha(1 - \alpha)X$, which is increasing in X for $\alpha \in (0, 1)$. □

To connect the bridge to the empirical variables, let $\mathcal{S}_t = \{y \in \mathcal{P} : J_t(y) = 1\}$ be the

set of publicly solved proteins. For a non-isolated target x, let $y_t^-(x)$ and $y_t^+(x)$ denote the nearest solved external anchors reached by leaving the target's local branch in its two directions. Let $W_t(x)$ denote the distance to the nearest solved member of x's own sequence cluster, with $W_t(x) = +\infty$ when the cluster has no solved member. Define

$$D_{xt} \equiv \begin{cases} W_t(x), & \text{if } x\text{'s cluster contains a solved member,} \\ \min\{D_{\mathcal{T}}(x, y_t^-(x)), D_{\mathcal{T}}(x, y_t^+(x))\}, & \text{otherwise,} \end{cases} \tag{15}$$
$$G_{xt} \equiv D_{\mathcal{T}}(x, y_t^-(x)) + D_{\mathcal{T}}(x, y_t^+(x)).$$

The first object is the distance to a usable solved structure. Following the empirical construction, it uses the within-cluster path whenever the cluster already contains a solved member and otherwise uses the nearest external anchor. The second is the width of the unsolved external gap surrounding the target. It uses two flanking anchors and therefore distinguishes a protein with one nearby anchor but a poorly mapped surrounding region from a protein embedded in a dense structural neighbourhood.

Additional anchors matter as well. Let ρ_{xt}^{cl} be the share of proteins in x's sequence cluster that have a solved structure. The cluster share records depth within a closely related sequence family. Its non-negative precision loading is λ_{cl}.

In the canonical case in which the closest usable anchor is one of the two flanking anchors, let $D = \min\{d, X - d\}$ denote the distance to the nearest endpoint and let $G = X$ denote the total external gap width. Since $D \leq G/2$, the bridge variance can be rewritten as $\bar{u}(D, G) = \sigma^2 D(G - D)/G$. In the protein tree, a solved protein in the same cluster may instead supply the relevant usable template path. Additional solved proteins in the same cluster also provide local information whose exact contribution depends on where those proteins sit. We therefore use the reduced-form frontier index

$$F_{xt} = f\left(\Big[\underbrace{\left(\sigma^2 \tfrac{D_{xt}(G_{xt} - D_{xt})}{G_{xt}}\right)^{-1}}_{\text{flanking-anchor precision}} + \underbrace{\lambda_{cl}\rho_{xt}^{cl}}_{\text{within-cluster precision}} \Big]^{-1} \right).^{26} \tag{16}$$

The reduced-form precision stack should be read as a tractable local index motivated by the Brownian bridge and the exact off-path decomposition in (14), not as the unique posterior variance of an arbitrary multi-anchor Brownian tree.

Assumption 1 (Regularity)**.** *On the local non-isolated region used for derivative statements, $G_{xt} > 0$ and $0 \leq D_{xt} \leq G_{xt}/2$. The functions f and χ are continuously differentiable, $f'(u) > 0$, $\chi(F) > 0$, and $\chi'(F) > 0$. The payoff, substitution, cost-relief, training-value, and continuation-value functions are differentiable at the relevant points. Cost relief is bounded so that post-AF experimental costs are non-negative. Whenever a strict comparative static is stated, the relevant derivative of F is non-zero and the target is away from the boundary at which local uncertainty is zero.*

The assumption restricts attention to a local region in which greater uncertainty makes

[26] At $D_{xt} = 0$, the expression is interpreted by continuity: flanking-anchor precision is infinite and residual local uncertainty is zero.

experiments more difficult and permits signed comparative statics. It does not require the tree index to summarise every source of experimental tractability.

Lemma 2 (Tree geometry and the empirical frontier measures)**.** *For a non-isolated target in the local region represented by* (16)*,*

$$F_D \equiv \frac{\partial F_{xt}}{\partial D_{xt}} \geq 0, \qquad F_G \equiv \frac{\partial F_{xt}}{\partial G_{xt}} \geq 0, \qquad F_{\rho^{cl}} \leq 0.$$

Thus, nearest-solved distance and gap width are measures of remoteness, while the cluster PDB share is an inverse measure of remoteness. The nearest-distance inequality is strict for $0 < D_{xt} < G_{xt}/2$*; the gap-width inequality is strict whenever* $D_{xt} > 0$*; and the density inequality is strict whenever its loading is positive and local posterior uncertainty is positive.*

Proof. For $G > 0$ and $0 \leq D \leq G/2$,

$$\frac{\partial \bar{u}}{\partial D} = \sigma^2 \frac{G - 2D}{G} \geq 0, \qquad \frac{\partial \bar{u}}{\partial G} = \sigma^2 \frac{D^2}{G^2} \geq 0.$$

The first derivative is strictly positive for $0 < D < G/2$, and the second is strictly positive whenever $D > 0$. Let

$$q = \bar{u}^{-1} + \lambda_{cl}\rho^{cl}, \qquad u = q^{-1}.$$

For $\bar{u} > 0$,

$$\frac{\partial u}{\partial \bar{u}} = \frac{u^2}{\bar{u}^2} > 0, \qquad \frac{\partial u}{\partial \rho^{cl}} = -\lambda_{cl} u^2 \leq 0.$$

At $D = 0$, the result follows by the continuous extension $\bar{u} = u = 0$. Since $F = f(u)$ and $f'(u) > 0$, the same signs apply to the frontier index. The density derivative is strict whenever its loading is positive and $u > 0$. □

Proteins in isolated clusters have no measurable path to a solved neighbour. For those proteins, the local interpolation formula is not available. This is why the empirical analysis max-imputes distance and uses a separate isolated indicator. The construction is local, but it gives the sign restrictions used by the empirical specifications.

A.1.2 Multifaceted knowledge and experimental payoffs

Protein knowledge is not one-dimensional. For each target x, let

$$\mathbf{u}_t(x) = \left(u_t^{\text{scaf}}(x), u_t^{\text{rel}}(x), u_t^{\text{conf}}(x), u_t^{\text{lig}}(x), u_t^{\text{val}}(x)\right)$$

collect marginal uncertainties about the dominant scaffold, the reliability of an available structure, conformational states, ligand-bound interactions, and validation. A deposition chooses a component bundle $b \subseteq \mathcal{A}$.

A fully correlated multidimensional representation is a precision-matrix update. Let $\Sigma_t(x)$ be the posterior covariance matrix for the component vector at x. Suppose bundle b

generates a Gaussian signal

$$Z_b = E_b \vartheta(x) + \varepsilon_b, \qquad \varepsilon_b \sim N(0, Q_b^{-1}),$$

where E_b selects or loads on the components informed by the bundle and Q_b is the signal-precision matrix. Then

$$\Sigma_{t+1}(x)^{-1} = \Sigma_t(x)^{-1} + E_b' Q_b E_b. \tag{17}$$

The componentwise update used in the main text is the diagonal or orthogonalised special case of (17). If components are conditionally independent, or have been rotated into independent local dimensions, and if $e_{b,j} \geq 0$ indicates whether bundle b informs component j with precision $q_{b,j} > 0$, then

$$u_{j,t+1}(x) = \left[u_{j,t}(x)^{-1} + e_{b,j} q_{b,j} \right]^{-1}. \tag{18}$$

Information can also propagate locally through the tree with a non-negative kernel $\mathcal{K}_b(x, y)$.

AF2 supplies an additional signal about the dominant scaffold,

$$Z_x^{AF} = \vartheta^{\text{scaf}}(x) + \varepsilon_x^{AF}, \qquad \varepsilon_x^{AF} \sim N(0, \varsigma_x^2), \tag{19}$$

where the signal noise may depend on multiple-sequence-alignment depth, local training-data density, disorder, domain architecture, and other features of the prediction problem. Conditional on the pre-AF posterior scaffold variance $u_t^{\text{scaf}}(x)$, the post-AF scaffold variance is

$$u_t^{\text{scaf},AF}(x) = \left[u_t^{\text{scaf}}(x)^{-1} + \varsigma_x^{-2} \right]^{-1}. \tag{20}$$

The reduction in scaffold uncertainty is

$$u_t^{\text{scaf}}(x) - u_t^{\text{scaf},AF}(x) = \frac{u_t^{\text{scaf}}(x)^2}{u_t^{\text{scaf}}(x) + \varsigma_x^2}, \tag{21}$$

which is increasing in prior scaffold uncertainty and decreasing in AF2 noise. The derivatives are

$$\frac{\partial}{\partial u} \frac{u^2}{u + \varsigma^2} = \frac{u(u + 2\varsigma^2)}{(u + \varsigma^2)^2} > 0, \qquad \frac{\partial}{\partial \varsigma^2} \frac{u^2}{u + \varsigma^2} = -\frac{u^2}{(u + \varsigma^2)^2} < 0.$$

This provides a posterior-learning interpretation of A_x in Section 3.2. The empirical variable $\text{AF}\,\text{cov}_p$ is a proxy for this latent scaffold informativeness. It is useful because pLDDT coverage summarises the share of residues predicted with usable local confidence, but it is not a complete measure of ς_x^{-2} for every downstream experiment.

Let a laboratory of type ℓ have a downstream value function

$$V_\ell(\mathbf{u}_t) = -\int_{\mathcal{P}} \sum_j w_{\ell j}(y) u_{j,t}(y)\, d\mu(y), \qquad w_{\ell j}(y) \geq 0. \tag{22}$$

The gross scientific or commercial value of experiment b at x is the improvement in (22), including propagated benefits to nearby targets. AF2 reduces the marginal value of experiments that reproduce information in components it already informs; in the main text this loss is $s_b(A_x, F_{xt})$. For scaffold-dependent experiments, AF2 can also lower the cost of selecting constructs, interpreting density maps, identifying candidate mechanisms, or designing ligand and conformation experiments. Those cost reductions are represented by $\delta^{sc}_m(A_x)$ and $\delta^{g}_{b,m}(A_x)$.

The distinction between substitution and complementarity is therefore component- and method-specific. A first-ever structure bundle has a direct substitution margin because it contains an R component. A ligand-bound or conformational component can become cheaper once a scaffold prediction is available even though the AF2 prediction does not itself observe the downstream state. A validating component can have both effects: its value as a generic measurement falls, but its value for establishing that a consequential prediction is reliable may remain high.

A.1.3 A general laboratory-choice model

For each protein x, let $J_t(x) \in \{0, 1\}$ indicate whether any experimental structure has been deposited publicly. Let $I_{\ell xt} \in \{0, 1\}$ indicate whether laboratory ℓ has previously built protein-specific capital for x. These states are cumulative. If any public deposition on x is released at date t, then $J_{t+1}(x) = 1$; otherwise $J_{t+1}(x) = J_t(x)$. If laboratory ℓ conducts any experiment on x at t, then $I_{\ell,x,t+1} = 1$; otherwise $I_{\ell,x,t+1} = I_{\ell xt}$. The comparative statics below condition on the state at the moment the laboratory chooses an experiment. A protein may therefore be familiar to the field but new to a particular laboratory.

An experiment selects a target x, a method $m \in \{\text{X-ray}, \text{cryoEM}, \text{NMR}\}$, and a non-empty component bundle $b \subseteq \mathcal{A}$. The feasible set is

$$\mathcal{B}(x,t) = \begin{cases} \{b \subseteq \mathcal{A} : R \in b,\ C \notin b\}, & J_t(x) = 0, \\ \{b \subseteq \mathcal{A} : b \neq \emptyset\}, & J_t(x) = 1. \end{cases} \tag{23}$$

A first-ever public structure must contain an R component because it creates the first public residue-level structural anchor for the protein. It need not be only R: it may also contain static refinement, conformational, or ligand-bound information. A pure corroboration component requires an existing public scaffold and is therefore a follow-up component in this taxonomy.

Let $B^{pre}_{\ell t}(x, b, m)$ denote the current scientific and commercial value that laboratory ℓ assigns to experiment (x, b, m) before AF2. This value includes own-target knowledge and local spillovers on the tree. Define

$$h^{sc}_{\ell xt}(b) = \mathbf{1}\{I_{\ell xt} = 0,\ J_t(x) = 1,\ b \cap \{S, L, C\} \neq \emptyset\}.$$

This indicator captures the scaffold-input cost faced by a laboratory that enters an already solved protein in order to conduct a scaffold-dependent follow-up experiment. For a first deposition, the R component generates the baseline scaffold within the same experimental episode; any additional difficulty of co-producing S or L information in that first deposition

is included in the variable cost of the bundle. Before AF2, the cost is

$$c_{\ell t}^{pre}(x,b,m) = \eta_m \Big[\underbrace{\phi_m^0 \mathbf{1}\{I_{\ell xt}=0\}}_{\text{lab–protein entry cost}} + \underbrace{\phi_m^{sc} h_{\ell xt}^{sc}(b)}_{\text{scaffold-input cost}} + \underbrace{c_{b,m}^{var}(x)}_{\text{bundle-specific variable cost}} \Big] \underbrace{\chi(F_{xt})}_{\text{frontier difficulty}}, \tag{24}$$

where $\phi_m^0 > 0$ is a generic lab–protein entry cost, $\phi_m^{sc} \geq 0$ is a scaffold-input cost, $c_{b,m}^{var}(x) \geq 0$ is a bundle-specific variable cost, $\eta_m > 0$ is a method scale, and χ is positive and increasing.

Distance affects experimental choice even before AF2. Define the pre-AF cost exposure

$$\kappa_{\ell t}^{pre}(x,b,m) = \eta_m \left[\phi_m^0 \mathbf{1}\{I_{\ell xt}=0\} + \phi_m^{sc} h_{\ell xt}^{sc}(b) + c_{b,m}^{var}(x) \right].$$

Holding other target characteristics fixed, write $B_{\ell t}^{pre}(x,b,m;F)$ for the gross value of an experiment at frontier position F. The derivative B_F^{pre} captures a knowledge-expansion benefit: a remote experiment can be valuable because it supplies information that is poorly inferred from existing structures. The derivative $\kappa_{\ell t}^{pre}\chi'(F)$ captures the opposing difficulty effect: remote experiments are costlier because structural templates and local knowledge are thinner.

Assumption 2 (Local choice mapping)**.** *For each feasible experiment $e = (x,b,m)$, the observed selection probability or experimental intensity is a differentiable function $\Lambda_{\ell t,e}(\boldsymbol{\pi}_{\ell t})$ of the deterministic payoff vector over feasible experiments. The map is focal-monotone: holding rival deterministic payoffs fixed,*

$$\frac{\partial \Lambda_{\ell t,e}}{\partial \pi_{\ell t,e}} > 0.$$

The assumption signs local changes in focal selection when only the focal payoff is perturbed. It does not sign arbitrary simultaneous changes in the payoffs of competing experiments.

Proposition 2 (Distance and experimental choice before AF2)**.** *Fix laboratory capital, method, component bundle, and non-landscape target characteristics. For any landscape measure $\nu \in \{D, G, \rho^{cl}\}$,*

$$\frac{\partial \pi_{\ell t}^{pre}(x,b,m)}{\partial \nu} = \left[\underbrace{B_F^{pre}(x,b,m;F_{xt})}_{\textit{knowledge-expansion benefit}} - \underbrace{\kappa_{\ell t}^{pre}(x,b,m)\chi'(F_{xt})}_{\textit{frontier difficulty effect}} \right] F_\nu. \tag{25}$$

The deterministic payoff rises with remoteness measures D and G, and falls with the solved-share measure ρ^{cl}, if and only if the knowledge-expansion benefit exceeds the difficulty effect:

$$B_F^{pre}(x,b,m;F_{xt}) > \kappa_{\ell t}^{pre}(x,b,m)\chi'(F_{xt}).$$

The signs reverse when the inequality is reversed. Under Assumption 2, a local perturbation that changes only this focal payoff produces the same sign for the focal selection probability or intensity.

Proof. Before AF2, deterministic payoff is

$$\pi^{pre}_{\ell t}(x,b,m) = B^{pre}_{\ell t}(x,b,m;F_{xt}) - \kappa^{pre}_{\ell t}(x,b,m)\chi(F_{xt}).$$

Holding laboratory capital, method, bundle, and non-landscape target characteristics fixed, differentiation with respect to a landscape measure ν gives (25). Lemma 2 gives $F_D, F_G \geq 0$ and $F_{\rho^{cl}} \leq 0$, which yields the stated sign conditions. Under Assumption 2, if only the focal payoff changes, then

$$\frac{\partial \Lambda_{\ell t,e}}{\partial \nu} = \frac{\partial \Lambda_{\ell t,e}}{\partial \pi_{\ell t,e}} \frac{\partial \pi_{\ell t,e}}{\partial \nu},$$

and the first factor is strictly positive. □

The proposition explains why distance is not simply a control for target difficulty. It is an economic state variable. For a first-ever bundle containing R, the value of opening a new structural region may be large. For scaffold-dependent S/L components, the benefit may be more target-specific and the cost gradient may dominate. The empirical distance coefficients reveal the net balance of these forces before AF2; the AF2 interactions reveal how the balance changes when a predicted scaffold becomes available.

AF2 provides a public signal with heterogeneous usable scaffold informativeness $A_x \in [0,1]$. To allow sequence information to matter separately from prior PDB geography, write

$$A_x = \Phi\Bigg(\underbrace{H_x}_{\text{sequence-based information}} + \underbrace{\int_{\mathcal{T}} \mathcal{K}^A(x,y)\mathbf{1}\{y \text{ solved by 2018-04-30}\}\, d\mu(y)}_{\text{inherited experimental anchors}} \Bigg), \tag{26}$$

where H_x summarises sequence-based information, including MSA depth; $\mathcal{K}^A(x,y) \geq 0$ decreases with tree distance; and Φ is weakly increasing. The release of AF2 is exogenous to an individual laboratory. The informativeness of its prediction at a particular protein is not: it reflects both evolutionary information and the inherited geography of experimental knowledge. Conditional on sequence information H_x, impose the local monotonicity of the inherited-anchor component:

$$A_D, A_G \leq 0, \qquad A_{\rho^{cl}} \geq 0.$$

These inequalities are maintained assumptions about the inherited experimental-anchor contribution to AF2 informativeness. They do not require sequence-based information and experimental geography to coincide, and they do not identify the structural cross-partials below from empirical AF coverage without additional controls.

Let $s_b(A_x, F_{xt}) \geq 0$ denote the loss in the current informational value of bundle b caused by AF2. The terms s_R, s_Q, and s_C are expected to be material because those components load directly on static scaffold information or its validation. For S and L, allow partial substitution rather than imposing zero substitution:

$$0 \leq s_S(A_x, F_{xt}), s_L(A_x, F_{xt}),$$

with the maintained interpretation that these terms are often smaller than static-scaffold substitution because AF2 does not itself observe the experimentally realised conformation or ligand-bound state. For a bundle b, s_b may be non-additive across components.

Let $\delta^{sc}_m(A_x)$ denote AF2 relief on the scaffold-input cost and let $\delta^{g}_{b,m}(A_x)$ denote bundle-specific variable-cost relief, including relief in construct design, model building, density interpretation, or co-producing mechanistic information with a first scaffold. Both are non-negative and weakly increasing in A_x. The cost-relief index $\mathcal{R}_{b,m}$ is defined in (1). The post-AF cost is

$$c^{AF}_{\ell t}(x,b,m) = c^{pre}_{\ell t}(x,b,m) - \eta_m \mathcal{R}_{b,m}(A_x, I_{\ell xt}, J_t(x))\chi(F_{xt}), \tag{27}$$

with non-negativity guaranteed by Assumption 1. The direct post-AF payoff shock is therefore the substitution-versus-complementarity expression in (2).

Proposition 3 (Distance and the AF2 treatment effect)**.** *Fix laboratory capital, method, component bundle, and non-landscape target characteristics. Let $\nu \in \{D, G, \rho^{cl}\}$. Allowing AF2 informativeness to vary with experimental geography, the AF2 payoff shock satisfies*

$$\frac{d\Delta\pi^{cur}_{\ell t}}{d\nu} = \underbrace{\left[\eta_m \mathcal{R}_{b,m}\chi' - s_{b,F}\right] F_\nu}_{\textit{frontier-position channel}} + \underbrace{\left[\eta_m \mathcal{R}_{b,m,A}\chi - s_{b,A}\right] A_\nu}_{\textit{inherited-quality channel}}. \tag{28}$$

Holding AF2 informativeness fixed gives the structural cross-partial in (3). At local interior points where $F_\nu > 0$, a more informative AF2 prediction has a larger payoff effect along either remoteness measure $\nu \in \{D, G\}$ if and only if

$$\eta_m \mathcal{R}_{b,m,A}(A_x, I_{\ell xt}, J_t(x))\chi'(F_{xt}) > s_{b,AF}(A_x, F_{xt}). \tag{29}$$

The interaction is negative when the inequality is reversed. For the cluster PDB share, the sign reverses at interior points because $F_{\rho^{cl}} < 0$ whenever its loading and local uncertainty are positive.

Proof. Using the cost-relief index in (1), the AF2 payoff shock is

$$\Delta\pi^{cur}_{\ell t} = -s_b(A_x, F_{xt}) + \eta_m \mathcal{R}_{b,m}(A_x, I_{\ell xt}, J_t(x))\chi(F_{xt}).$$

Taking the total derivative with respect to landscape measure ν, allowing A_x to vary with geography, gives (28). Taking the partial derivative first with respect to A_x and then with respect to ν, holding A_x fixed when differentiating the landscape, gives (3). For $\nu \in \{D, G\}$ and $F_\nu > 0$, the cross-partial has the sign of $\eta_m \mathcal{R}_{b,m,A}\chi' - s_{b,AF}$. For the density measures, the sign is reversed when $F_\nu < 0$. □

Equation (28) describes how the overall AF2 shock varies across the realised landscape. Equation (3) instead describes whether the marginal payoff effect of a higher-quality AF2 prediction changes with distance, holding that quality fixed. The empirical AF coverage-by-distance coefficient is a reduced-form analogue of this cross-partial. It need not equal the structural object unless the empirical design isolates variation in latent scaffold informativeness from other post-AF changes associated with distance. For S/L components,

the interaction with distance is positive when AF2's cost-relief effect strengthens with informativeness more than any partial substitution does. For R, Q, and C components, the sign can reverse if a usable AF2 prediction substitutes especially strongly for static-scaffold experiments in remote regions.

Proposition 4 (Direct payoff effects by experiment type)**.** *Fix protein x, method m, and laboratory ℓ. The direct AF2 payoff shock for bundle b is given by (2). The direct payoff from the experiment rises after AF2 if and only if $\Delta_{\ell t}(x, b, m) > 0$. It falls if $\Delta_{\ell t}(x, b, m) < 0$ and is unchanged under equality. Under Assumption 2, the same sign applies to the focal selection probability or intensity for a local perturbation that holds rival payoffs fixed. In particular:*

(i) For a first-ever structure, $R \in b$, $C \notin b$, and $h^{sc}_{\ell x t}(b) = 0$. The payoff rises if and only if

$$\underbrace{\eta_m \delta^g_{b,m}(A_x)\chi(F_{xt})}_{\text{variable and co-production cost relief}} > \underbrace{s_b(A_x, F_{xt})}_{\text{substitution for first-ever scaffold information}}.$$

(ii) For a scaffold-dependent follow-up experiment conducted by an outside-pipeline laboratory, $h^{sc}_{\ell x t}(b) = 1$. The payoff rises if and only if

$$\underbrace{\eta_m \left[\delta^{sc}_m(A_x) + \delta^g_{b,m}(A_x)\right]\chi(F_{xt})}_{\text{scaffold-input relief and variable-cost relief}} > \underbrace{s_b(A_x, F_{xt})}_{\text{information substitution}}.$$

(iii) For a follow-up experiment conducted by an inside-pipeline laboratory, $h^{sc}_{\ell x t}(b) = 0$. The payoff rises if and only if

$$\underbrace{\eta_m \delta^g_{b,m}(A_x)\chi(F_{xt})}_{\text{variable-cost relief}} > \underbrace{s_b(A_x, F_{xt})}_{\text{information substitution}}.$$

Proof. From (2),

$$\pi^{AF}_{\ell t}(x, b, m) - \pi^{pre}_{\ell t}(x, b, m) = -s_b(A_x, F_{xt}) + \eta_m \mathcal{R}_{b,m}(A_x, I_{\ell x t}, J_t(x))\chi(F_{xt}).$$

Substituting the definition of $\mathcal{R}_{b,m}$ gives (2). The sign of the direct payoff shock is therefore the sign of $\Delta_{\ell t}(x, b, m)$. Assumption 2 transfers this sign to a focal selection probability or intensity when rival payoffs are held fixed. □

The distinction between the first and third cases is economic rather than algebraic. A first-ever experiment receives no scaffold-input relief because it must establish the first public scaffold itself. An inside-pipeline laboratory receives no such relief because it has already built protein-specific capital. By contrast, an outside-pipeline laboratory can use AF2 to avoid part of the setup cost that previously made scaffold-dependent follow-up work expensive. This creates a natural reallocation towards mechanistic experiments, such as new conformational states or ligand-bound structures, for which AF2 supplies an input without supplying the experiment's main output.

Proposition 5 (Corroboration and validation)**.** *Suppose a corroborating experiment adds an independent scalar scaffold signal to an already public scaffold. Its marginal informational value is positive, decreases with the number of prior corroborations, and falls after AF2 supplies additional scaffold precision. The direct payoff from corroboration rises whenever*

$$\underbrace{\eta_m \delta^{sc}_m(A_x)\mathbf{1}\{I_{\ell xt}=0, J_t(x)=1\}\chi(F_{xt})}_{\text{scaffold-input relief}} + \underbrace{\eta_m \delta^{g}_{\{C\},m}(A_x)\chi(F_{xt})}_{\text{bundle-specific cost relief}} + \underbrace{v^{AF}_{\ell}(x)}_{\text{incremental validation payoff}} > \underbrace{s_{\{C\}}(A_x, F_{xt})}_{\text{information substitution}}.$$

Here $v^{AF}_{\ell}(x)$ is the incremental validation payoff created by the availability or use of the AF2 prediction, and $s_{\{C\}}$ is measured in the same payoff units as the value lost when AF2 supplies prior scaffold information.

Proof. Let $P_n > 0$ be prior scaffold precision after n legacy measurements, let $q > 0$ be the precision of another validating experiment, and let $a > 0$ be the precision provided by AF2. Before AF2, the marginal variance reduction from an additional corroborating signal is

$$\frac{1}{P_n} - \frac{1}{P_n + q} = \frac{q}{P_n(P_n + q)}.$$

After AF2, the corresponding marginal variance reduction is

$$\frac{1}{P_n + a} - \frac{1}{P_n + a + q} = \frac{q}{(P_n + a)(P_n + a + q)}.$$

Both expressions are positive. Differentiating $q/[P(P+q)]$ with respect to P gives $-q(2P + q)/[P^2(P+q)^2] < 0$, so marginal informational value decreases as prior precision rises. Since $a > 0$ raises the prior precision from P_n to $P_n + a$, the AF2 expression is smaller. If the laboratory values a unit of variance reduction at weight $w_{\ell C}(x)$, the information-substitution term $s_{\{C\}}$ is the value-weighted loss in marginal variance reduction. Adding incremental validation payoff $v^{AF}_{\ell}(x)$ and the cost-relief term gives the inequality in the proposition. □

The validation term is likely to differ across institutions. Industry laboratories may place greater weight on ligand-bound and corroborating experiments when a validated structure has commercial value. Academic laboratories may place broader weight on scientific spillovers, while public consortia may deliberately reward work on underexplored proteins. These are contingent institutional predictions rather than primitive theorems. They motivate the empirical comparisons between academic and industry laboratories.

For a first-ever structure bundle $b \in \mathcal{B}(x,t)$ with $R \in b$ and $C \notin b$, define the direct AF2 shock

$$\Delta^{E}_{\ell t}(x, b, m) = \underbrace{\eta_m \delta^{g}_{b,m}(A_x)\chi(F_{xt})}_{\text{variable and co-production relief}} - \underbrace{s_b(A_x, F_{xt})}_{\text{information substitution}}, \tag{30}$$

where the scaffold-input indicator is zero because $J_t(x) = 0$.

Proposition 6 (Frontier selection conditional on first deposition)**.** *Fix method m, a feasible first-ever bundle b with $R \in b$ and $C \notin b$, and a remoteness measure $\nu \in \{D, G\}$. Conditional on AF2 informativeness,*

$$\frac{\partial^2 \Delta^E_{\ell t}(x, b, m)}{\partial A \, \partial \nu} = \left[\underbrace{\eta_m \delta^g_{b,m,A}(A_x)\chi'(F_{xt})}_{\substack{\textit{frontier-amplified} \\ \textit{cost relief}}} - \underbrace{s_{b,AF}(A_x, F_{xt})}_{\substack{\textit{frontier-amplified} \\ \textit{substitution}}} \right] F_\nu. \tag{31}$$

At local interior points where $F_\nu > 0$, a more informative AF2 prediction shifts the payoff gradient for first deposition outwards if variable or co-production cost relief is more strongly amplified by remoteness than information substitution. It shifts the payoff gradient inwards when the inequality is reversed.

For target selection, suppose the candidate first-ever targets under consideration have the same non-landscape characteristics and the same b and m, and suppose the AF2 shock can be written as a function $\delta(z(x))$ of remoteness alone over that candidate set. If δ is strictly increasing, every post-AF maximiser has weakly larger z than every pre-AF maximiser. If δ is strictly decreasing, every post-AF maximiser has weakly smaller z than every pre-AF maximiser. If δ is only weakly monotone, the same point-selection statement requires a tie-breaking rule that is monotone in z; without such a rule the payoff comparison implies only the corresponding weak ordering of the AF2 shocks.

Proof. For a first-ever public structure, $J_t(x) = 0$ and every feasible bundle contains R and excludes C. The scaffold-input indicator is zero, so $\mathcal{R}_{b,m} = \delta^g_{b,m}(A_x)$. Substituting this identity into (3) gives (31). Since $F_D, F_G \geq 0$, and strictly positive at the relevant interior points, its sign is determined by the comparison in the proposition.

For the selection statement, let x^{pre} and x^{AF} be optimal first-ever targets before and after AF2 within the candidate set, holding b and m fixed. Write the AF2 shock as $\delta(z(x))$ over this candidate set. Optimality implies

$$\pi^{pre}(x^{pre}) \geq \pi^{pre}(x^{AF})$$

and

$$\pi^{pre}(x^{AF}) + \delta(z(x^{AF})) \geq \pi^{pre}(x^{pre}) + \delta(z(x^{pre})).$$

Adding the two inequalities gives $\delta(z(x^{AF})) \geq \delta(z(x^{pre}))$. If δ is strictly increasing, then $z(x^{AF}) \geq z(x^{pre})$ for any chosen pre- and post-AF maximisers. If δ is strictly decreasing, then $z(x^{AF}) \leq z(x^{pre})$. If δ is weakly monotone, the same conclusion for point selections follows when ties are broken monotonically in z; otherwise, only the weak ordering of the AF2 shocks follows. □

The sign is not mechanical. For first deposition, an informative AF2 prediction can lower the cost of residue-level and co-produced mechanistic work, but it can also replace discovery value in regions where pre-AF templates were scarce. Equation (31) separates

those forces. Nearest-solved distance and gap size are therefore not merely descriptive variables. They are remoteness measures governing the local payoff comparative static. The cluster PDB share provides an inverse measure of the same local frontier state.

A.1.4 Multifaceted experimental data supply

Current experiment choice affects the corpus inherited by later systems. The model represents this as directional experimental data supply rather than as a prediction of realised successor-model performance. Let g index prediction-system generations and let M_g^b denote the measure of experimental deposits with bundle b generated while system g is available. Let $\mathbf{T}_{g+1}^{exp}(z)$ be the vector of experimental input stocks available for a later model at protein z. Its components include the scaffold-range, mechanistic, and corroboration stocks introduced in (4). Assume each M_g^b is a finite non-negative measure on $\mathcal{T}$, μ is finite or σ-finite, $\mathbf{\Gamma}_b \in \mathbb{R}_+^3$, and $\mathcal{K}_b(z,x)$ is measurable. When the welfare weights $\boldsymbol{\omega}(z)$ are introduced below, assume $\boldsymbol{\omega}(z)'\mathbf{\Gamma}_b\mathcal{K}_b(z,x)$ is absolutely integrable with respect to $\mu \times M_g^b$ for every bundle b. These regularity conditions make the following objects finite and justify the exchange of integration used below:

$$\mathbf{T}_{g+1}^{exp}(z) = \sum_{b\subseteq\mathcal{A},\, b\neq\varnothing} \int_{\mathcal{T}} \underbrace{\mathbf{\Gamma}_b}_{\text{data-supply content}} \underbrace{\mathcal{K}_b(z,x)}_{\text{local propagation}} \, dM_g^b(x). \tag{32}$$

Here $\mathcal{K}_b$ allows experimental value to propagate locally and $\mathbf{\Gamma}_b$ records which data-supply components each experimental bundle improves. A remote first-ever bundle containing R can broaden fold and sequence-family coverage. A bundle containing S or L supplies mechanistic observations. A bundle containing C supplies validation and error-correction observations.

A later system maps these experimental stocks, together with evolutionary and other non-PDB information, pseudo-data and self-distillation, model architecture, curation choices, and proprietary data, into realised capability. The empirical analysis measures changes in the experimental input stock $\mathbf{T}_{g+1}^{exp}$, not this full mapping.

To connect directional data supply to the welfare discussion in Section 3.4, let $\boldsymbol{\omega}(z)$ be a vector of marginal social weights on the experimental input stocks available at z. These weights can be read as a first-order approximation to the downstream scientific and innovation value of a richer inherited experimental corpus around the current state. Define

$$\mathcal{W}_{g+1}^{exp} = \int_{\mathcal{T}} \boldsymbol{\omega}(z)'\mathbf{T}_{g+1}^{exp}(z)\, d\mu(z). \tag{33}$$

Substituting (32) into (33) and, under the integrability conditions above, exchanging the order of integration gives

$$\mathcal{W}_{g+1}^{exp} = \sum_{b\subseteq\mathcal{A},\, b\neq\varnothing} \int_{\mathcal{T}} \tau_{b,g}(x)\, dM_g^b(x), \qquad \tau_{b,g}(x) \equiv \int_{\mathcal{T}} \boldsymbol{\omega}(z)'\mathbf{\Gamma}_b\mathcal{K}_b(z,x)\, d\mu(z). \tag{34}$$

Thus, $\tau_{b,g}(x)$ is not an additional primitive attached to the model. It is the marginal downstream data-supply value implied by the same bundle loadings and propagation kernels that generate $\mathbf{T}^{exp}_{g+1}$. A bundle may have high value because it supplies mechanistic observations, scaffold-range anchors, corroboration observations, or several at once.

For a component k of the experimental input stock and any target set $H \subseteq \mathcal{T}$, define the AF2-induced weighted change

$$\Delta T^{exp}_{k,H}(z) = \sum_{b \subseteq \mathcal{A},\, b \neq \emptyset} \int_H \Gamma_{b,k} \mathcal{K}_b(z, x)\, d\Delta M^b_g(x),$$

where ΔM^b_g is the change in the deposit measure induced by AF2. This object, not a raw deposit count, is what enters successor-model data supply when bundles load on multiple components.

Lemma 3 (Componentwise interpretation of experimental data-supply changes)**.** *Order experimental input stocks by the product partial order on $\mathbb{R}^3$. Let H be a high-frontier region. Suppose the AF2-induced weighted change in the scaffold/range component satisfies $\Delta T^{exp}_{k_R,H}(z) < 0$ for the relevant successor target z, where k_R denotes scaffold-range observations. Suppose also that the weighted change in the mechanistic component is positive. Then AF2 shifts experimental data supply away from broadening scaffold-range observations in H and towards deepening mechanistic observations. Raw increases or decreases in deposit counts imply these signs only when the relevant cross-component loadings are absent or are dominated in the corresponding weighted measure. The sign of the net effect on realised successor-model performance is not determined by this lemma.*

Proof. For component k of the experimental input stock and frontier region H, the AF2-induced weighted change is

$$\Delta T^{exp}_{k,H}(z) = \sum_b \int_H \Gamma_{b,k} \mathcal{K}_b(z, x)\, d\Delta M^b_g(x).$$

The lemma assumes $\Delta T^{exp}_{k_R,H}(z) < 0$ for the scaffold/range component and a positive weighted change for the mechanistic component. These are exactly the componentwise inequalities defining a shift away from broadening scaffold-range observations in H and towards deepening mechanistic observations. Because bundles can load on multiple components, raw deposit counts sign these inequalities only under additional restrictions on $\mathbf{\Gamma}_b$ and $\mathcal{K}_b$. The statement concerns $\Delta \mathbf{T}^{exp}$. The mapping from $\mathbf{T}^{exp}$ to realised model performance also depends on $\mathbf{\Phi}$, architecture, curation, pseudo-data, non-PDB data, and evaluation tasks, so the lemma does not determine the net performance effect. □

The lemma rules out a scalar interpretation of "more training data". A larger deposition stream can supply less frontier-expanding data and more mechanistic data at the same time. This distinction is measured using new conformations, biologically relevant ligands, binding-affinity data, first-in-cluster deposits, and structurally novel folds.

A.1.5 Proof of Proposition 1

Let $\pi^{cur}_{\ell t}(x, b, m)$ denote the current payoff after AF2, and let a laboratory internalise share θ_ℓ of the marginal downstream data-supply value in (34). Its private ranking payoff and the corresponding social ranking payoff are

$$\pi^{priv}_{\ell t}(x, b, m) = \pi^{cur}_{\ell t}(x, b, m) + \theta_\ell \tau_{b,g}(x), \qquad \pi^{soc}_{\ell t}(x, b, m) = \pi^{cur}_{\ell t}(x, b, m) + \tau_{b,g}(x).$$

Their difference is

$$\pi^{soc}_{\ell t}(x, b, m) - \pi^{priv}_{\ell t}(x, b, m) = (1 - \theta_\ell)\tau_{b,g}(x).$$

When $\theta_\ell < 1$ and $\tau_{b,g}(x) > 0$, private choice underweights the experiment relative to the social ranking. By assumption, AF2 raises selected follow-up activity enough that total activity rises, while reducing selected first-ever R-containing bundles in the high-frontier region H enough that the weighted range component satisfies $\Delta T^{exp}_{k_R,H}(z) < 0$. The additional mechanistic-component assumption and Lemma 3 then imply a shift away from scaffold-range expansion in H and towards mechanistic deepening even though total activity rises.

If the total downstream data-supply value $\tau_{b,g}(x)$ increases with remoteness for the relevant frontier-expanding bundles, differentiating the wedge with respect to frontier position gives

$$\frac{\partial}{\partial F_{xt}}\left[\pi^{soc}_{\ell t}(x, b, m) - \pi^{priv}_{\ell t}(x, b, m)\right] = (1 - \theta_\ell)\frac{\partial \tau_{b,g}(x)}{\partial F_{xt}} > 0.$$

The total missing reward is then largest at the frontier. If only the range-expansion component is known to increase with remoteness, define

$$\tau^{scaf}_{b,g}(x) \equiv \int_{\mathcal{T}} \omega_{k_R}(z)\Gamma_{b,k_R}\mathcal{K}_b(z, x)\, d\mu(z).$$

The same calculation applied to $(1 - \theta_\ell)\tau^{scaf}_{b,g}(x)$ establishes only the component-specific claim: the range-expansion part of the missing reward is largest at the frontier. □

A.1.6 Downstream incentives and moonshots

Equation (34) derives the downstream experimental-data value $\tau_{b,g}(x)$ from the bundle content and propagation structure of the model. The baseline private-choice model does not require laboratories to internalise that value. Section 3.4 isolates the resulting successor-system wedge. This subsection allows a laboratory or its funder to reward both future AI data supply and continuation value for later human research. Let

$$r_{\ell t}(x, b) = \underbrace{\theta_\ell \tau_{b,g}(x)}_{\text{future AI data-supply reward}} + \underbrace{\zeta_\ell V^{cont}_t(x, b)}_{\text{continuation value for human research}}. \tag{35}$$

The first term is the reward laboratory ℓ receives for an experiment's contribution to future AI data supply or model usefulness. It may arise from commercial returns, contracts, grants, citations, reputation, or an explicit data-generation programme. The second term rewards the continuation value created because the experiment guides later human research. The parameters θ_ℓ and ζ_ℓ measure how strongly the laboratory or its funder internalises these downstream effects. When a laboratory internalises these effects, it adds $r_{\ell t}(x, b)$ to the current payoff when ranking feasible experiments.

The continuation term connects the model to Carnehl and Schneider (2025). A remote anchor may have low immediate value but open a gap that later scientists can fill. Their moonshot logic is not that arbitrarily remote work is valuable. An anchor that lies too close to existing knowledge adds little, while one that is too disconnected may fail to inspire productive follow-on work. The continuation value can therefore be hump-shaped in frontier distance.

The training term connects the model to Gans (2026). If downstream AI value is concentrated within a reliable operational range, scientists may "work to the AI" by choosing experiments near that boundary. Such experiments are immediately useful but need not provide the out-of-range data required to expand later systems. By contrast, a patient institution may reward deliberately remote anchors because they broaden the experimental data supply available for future systems.

Proposition 7 (Training incentives and moonshots)**.** *Let $\mathcal{E}$ be a candidate set of feasible first-ever experiments containing R, and write each candidate as $e = (x_e, b_e, m_e)$. Let $e^\star = (x^\star, b^\star, m^\star)$ maximise current payoff $\pi^{cur}_{\ell t}(x_e, b_e, m_e)$ over $\mathcal{E}$, and let $e_M = (x_M, b_M, m_M) \in \mathcal{E}$ be a candidate moonshot with $F_{x_M t} > F_{x^\star t}$. With future rewards, the moonshot is selected over the current-payoff maximiser if and only if*

$$\underbrace{\theta_\ell \left[\tau_{b_M,g}(x_M) - \tau_{b^\star,g}(x^\star)\right]}_{\text{incremental future AI value}} + \underbrace{\zeta_\ell \left[V^{cont}_t(x_M, b_M) - V^{cont}_t(x^\star, b^\star)\right]}_{\text{incremental continuation value}} \geq \underbrace{\pi^{cur}_{\ell t}(x^\star, b^\star, m^\star) - \pi^{cur}_{\ell t}(x_M, b_M, m_M)}_{\text{current-payoff cost of the moonshot}}.$$

It is selected from the full feasible set if and only if its augmented payoff is at least as large as that of every feasible alternative, with strict inequality giving uniqueness and weak inequality allowing ties. If training value peaks near the current reliable AI boundary, higher θ_ℓ can induce boundary-following research when the induced future-AI term dominates offsetting current-payoff and continuation-value differences. If the continuation value is hump-shaped, an intermediate moonshot may be selected while a sufficiently remote "marsshot" is not.

Proof. For any feasible experiment $e = (x_e, b_e, m_e)$, define the augmented payoff

$$\Pi_{\ell t}(e) = \pi^{cur}_{\ell t}(x_e, b_e, m_e) + \theta_\ell \tau_{b_e,g}(x_e) + \zeta_\ell V^{cont}_t(x_e, b_e).$$

The researcher selects e_M over $e^\star$ if and only if $\Pi_{\ell t}(e_M) \geq \Pi_{\ell t}(e^\star)$. Substituting the definition of $\Pi_{\ell t}$ and rearranging yields the displayed inequality. Selection from the full feasible set is equivalent to $\Pi_{\ell t}(e_M) \geq \Pi_{\ell t}(e)$ for every feasible alternative e, with strict inequality for uniqueness. The boundary-following and moonshot statements are conditional comparative-static implications of the same inequality: they follow when the stated shapes of $\tau_{b,g}$ or V^{cont}_t dominate the relevant current-payoff differences and the

other downstream term. □

AF2 is, therefore, not itself a moonshot programme. It is an infrastructure shock that changes the return to different experiments. Whether it encourages frontier expansion depends on the rewards faced by laboratories and on whether future AI value is attached to immediately usable structures, genuinely out-of-range structures, or both.

A.1.7 Relationship to Carnehl–Schneider and Gans

The tree model contains local ingredients of the linear knowledge space in Carnehl and Schneider (2025) under additional restrictions. Restrict the tree to a line, retain first-ever scaffold-anchor experiments, and suppose a solved anchor creates an unresolved interval of length X. Equation (13) then gives the same local Brownian learning object: nearby experiments are cheaper or more predictable, while a deliberately distant anchor can create a region that subsequent scientists find valuable to fill in. The full dynamic search environment in Carnehl and Schneider (2025) includes additional structure beyond the local bridge used here. The reduced-form term $V_t^{cont}(x, b)$ records the continuation value associated with that logic.

The model also provides a protein-science analogue of the endogenous-training mechanism in Gans (2026). Collapse the experimental input stock to a scalar reliable range R_g and let first-ever frontier experiments beyond that range expand the next model's data frontier:

$$R_{g+1} = (1 - \delta_R)R_g + \gamma_R \int_{\mathcal{P}} (F_{xg} - R_g)_+ \, dM_g^R(x). \tag{36}$$

Equation (36) is an illustrative one-dimensional restriction of (32). If current AI makes work inside its range more attractive, experiments can bunch near the work-to-AI boundary and produce too little frontier data supply. A training-aware scientist, funder, or firm places positive weight on $\tau_{b,g}(x)$ and may instead select an experiment with greater successor data value.

The protein setting adds two features. First, later systems can benefit from several types of observations: novel structural anchors, interaction data, conformational-state data, and validation data may expand different components of the experimental input stock. Second, the institution conducting the experiment matters. Academic laboratories may place relatively more weight on scientific spillovers, citations, and grant missions, captured by ζ_ℓ. Commercial laboratories may place more weight on proprietary downstream value and on models they can use or control, captured by $w_{\ell j}$ and θ_ℓ. The signs of the local comparative statics are common across institutions, while their magnitudes and chosen experiment types may differ.

A.1.8 Alternative knowledge-space foundation

The Brownian-tree formulation makes unresolved gaps transparent. A smooth process with protein-specific uncertainty gives related reduced-form objects. Let

$$\vartheta(x) = g(x) + \xi_x,$$

where

$$\mathrm{Cov}(g(x), g(y)) = \sigma_g^2 \exp\{-d(x,y)/\ell\}, \qquad \xi_x \sim N(0, \sigma_\xi^2)$$

independently across proteins. If K_t is the covariance matrix among solved proteins and $k_t(x)$ is the covariance vector between x and those proteins, the posterior variance is

$$u_t(x) = \sigma_g^2 + \sigma_\xi^2 - k_t(x)'(K_t + \sigma_\xi^2 I)^{-1} k_t(x).$$

In one-anchor or symmetric local settings, this posterior variance falls as a solved anchor moves closer and falls as additional nearby independent anchors are added. In arbitrary multi-anchor geometries, global monotonicity in a single distance statistic need not hold because the covariance vector and the inverse covariance matrix change jointly. The empirical nearest-solved distance, gap length, and solved-neighbour density should therefore be interpreted jointly as observable proxies for local posterior uncertainty, rather than as a complete ordering of all targets. Applying the AF2 signal in (19) again yields (20) and (21).

A.2 Data Construction

A.2.1 Data Sources

Universal Protein Resource Knowledgebase (UniProtKB) is the central hub for functional information on proteins, with accurate, consistent, and rich annotation. Each entry records core data such as the amino-acid sequence, protein name or description, taxonomic information, and citations. UniProtKB consists of manually annotated records with information extracted from the literature and curator-evaluated computational analysis (UniProtKB/Swiss-Prot), and computationally analysed records that await full manual annotation (UniProtKB/TrEMBL). Both are available to download here (downloaded 2026-01-05).

The Protein Data Bank (PDB) is a freely available repository for structural coordinates of biological macromolecules, primarily proteins. The rcsb-api package provides a Python interface to the RCSB PDB Search and Data APIs (Piehl et al., 2025; Rose et al., 2021).

Structure integration with function, taxonomy and sequence (SIFTS) is a project in the PDBe-KB resource for residue-level mapping between UniProt and PDB entries. Three data tables are useful for this project, available here (downloaded on 2025-09-01) or from the FTP. The first is *uniprot_segments_observed*. It contains many rows per protein: each row is an observed interval for a specific PDB chain. It provides the detailed residue-level mapping from individual PDB chains to UniProt residue numbering. Specifically, it shows the start and end residues of the mapping using SEQRES, PDB sequence, and UniProt numbering. This is used to determine experimental coverage. The second is *pdb_chain_uniprot*. It contains one row per PDB chain with the UniProt link. This is used to count chains and build the set of PDB entries per UniProt. The last table, *pdb_pubmed*, summarises the PubMed identifier(s) associated with each PDB entry.

The AlphaFold Protein Structure Database (AFDB) is a publicly available repository created by DeepMind and EMBL–EBI to disseminate protein structures predicted by the AlphaFold2 algorithm. Launched in July 2021, the database initially contained

high-confidence predictions for the complete human proteome and 20 model organisms, and was later expanded to cover over 200 million proteins from the UniProt sequence collection—effectively providing structural models for nearly every protein known to science. Each entry in AFDB is indexed by a UniProt accession identifier and includes downloadable coordinate files in PDB and mmCIF formats, a per-residue confidence score (pLDDT) that quantifies the reliability of the prediction, and a predicted aligned error (PAE) matrix that indicates uncertainty in the relative positions of different regions of the protein. AFDB distributes PDB files for each predicted model, where the B-factor column stores pLDDT. In the PDB format, atom field columns 61–66 normally hold experimental temperature factors; AlphaFold repurposes that field to carry the confidence score. We parse the AF PDB to take the CA atom B-factor per residue as pLDDT (falling back to the average of atoms if CA is missing).

OpenAlex is a free, open catalogue of the world's scholarly works, authors, institutions, and their connections (Priem et al., 2022). It indexes over 250 million works from Crossref, PubMed, institutional repositories, and other sources, linking each publication to its authors and their institutional affiliations. We use OpenAlex to enrich our PDB deposition data with author-level metadata. Each PDB entry has an associated primary citation (a DOI or PubMed ID); we match these to OpenAlex works records, which provide the full author list, each author's institutional affiliation at the time of publication, and career-level bibliometric information. This linkage allows us to characterise *who* is doing the experimental work—whether depositions come from incumbent laboratories that have previously worked on a protein or from new entrants, and whether the institutional composition of the research workforce shifts after AlphaFold2.

A.2.2 Identifying Lab

The match uses a priority chain: PMID first, then DOI, then a title-and-year fallback that we keep only when the OpenAlex candidate is unique, and the first structure-author surname on the PDB entry appears among the OpenAlex authors. From the matched work, we take the disambiguated OpenAlex ID of the last listed author as the laboratory identifier, falling back to the corresponding author and then the first author when the last-author ID is missing. After this fallback chain, the laboratory identifier is missing for 16.3% of PDB entries; we drop these rows from any specification that uses the lab identifier. We retain a surname-only alternative built from the PDB list as a robustness identifier, but our preferred specifications use the disambiguated OpenAlex ID because surname keys both merge unrelated researchers sharing a surname and split a single researcher across name-format variants. Each lab is classified by the affiliation type of its last author from the OpenAlex institutions array, using the priority ordering education $\succ$ facility $\succ$ company $\succ$ healthcare $\succ$ government $\succ$ nonprofit.

A.2.3 Deposition Categories

Most PDB depositions (88.5%) are follow-up structures for proteins with at least one prior entry. These intensive-margin depositions serve scientifically distinct purposes, which

motivate our five-category classification. For each deposition d of protein p, released in year t, the following flags are evaluated:

- ***E* (First Structure).** The deposition is the first-ever PDB entry for protein p. Formally, $E_d = 1$ if d is an extensive-margin deposition (no prior PDB entry maps to p via SIFTS). A first structure creates a new protein-level anchor: it generates positive externalities for sequence neighbours via homology modelling and opens the protein to structure-based drug design, mechanistic annotation, and comparative analysis across species. Its contribution to frontier expansion depends on its distance from previously solved neighbours.

- **Completeness *R* (or Improved Coverage).** The deposition resolves residues not covered by any prior deposition for the same protein. We link experiments to UniProtKB proteins using the Structure Integration with Function, Taxonomy and Sequences (SIFTS) resource, which provides residue-level alignments between PDB depositions and UniProtKB accessions (Velankar et al., 2013). Residues—the individual amino acids in a protein chain—are indexed according to the UniProtKB reference sequence. For each deposition, SIFTS identifies which residues are experimentally observed. Each deposition d can contribute to the coverage of the protein:

$$\Delta e_d = \frac{1}{L_p} \sum_{i=1}^{L_p} \mathbf{1}\{\text{residue } i \text{ observed in } d \text{ but not in any prior deposition}\}. \tag{37}$$

  Coverage is inherently continuous—a protein may have select domains resolved across multiple depositions spanning many years before the full chain is characterised. Extending the experimentally observed fraction of a protein is scientifically valuable because many functional sites—active sites, allosteric pockets, regulatory loops—may lie in regions absent from earlier, partial structures. Coverage extension is also common when a protein is re-solved in a different crystal form, space group, or by a complementary method (e.g., cryo-EM after an initial X-ray structure), each of which may stabilise different regions of the polypeptide chain.

- **Mechanistic *SL* (New Conformational State or Biologically Relevant Ligand).** The deposition reveals new mechanistic information about the protein, either by capturing it in a previously unobserved three-dimensional conformation or by resolving a biologically relevant bound ligand. Formally, $SL_d = 1$ if $S_d = 1$ or $L_d = 1$, or both, where the two constituent flags are defined as follows. $S_d = 1$ if d is the first deposition in a new Foldseek conformational cluster—all PDB entries for the same protein are compared pairwise using TM-scores, structures with TM-score ≥ 0.85 are grouped via single-linkage clustering, and a deposition triggers S when its 3D shape is sufficiently distinct (TM-score < 0.85) from all previously deposited structures. $L_d = 1$ if the deposition contains at least one non-polymer ligand of biological interest after excluding crystallographic additives (solvents, cryoprotectants, buffer components, and common ions such as HOH, SO_4, GOL, PEG, EDO, glycerol, and common metal ions). The exclusion list and the underlying "biologically relevant

ligand" convention follow established curation practice in the structural-biology and cheminformatics databases that apply the same crystallographic-additive filter when assembling protein–ligand datasets from the PDB (Yang et al., 2013; Wang et al., 2004; Salentin et al., 2015). These two flags are combined because they share a common scientific rationale: they capture target-specific mechanistic value that goes beyond extending the structural map of a protein. Proteins are not static objects—they adopt open and closed forms, active and inactive states, and pre- and post-catalytic snapshots that illuminate mechanism. Conformational diversity is shaped by pH, temperature, post-translational modifications (phosphorylation, glycosylation), and the presence of binding partners; when two or more structures of the same protein are compared, RMSD values can range from 0.1–0.4 Å under similar conditions to tens of angstroms when conditions differ (Burra et al., 2009). Fraser et al. (2011) showed that cryocooling alone remodels the conformational distributions of more than 35% of side chains, and Keedy et al. (2023) demonstrated that temperature can redirect ligand binding to entirely different sites. Ligand-bound structures are likewise central to structure-based drug design: comparing drug-bound and apo forms reveals how cofactors, metal ions, and substrates reshape the binding pocket. High-throughput fragment screening campaigns are a major modern driver of such depositions, with pharmaceutical and academic groups routinely depositing hundreds of structures of the same protein with different fragments bound (Jaskolski et al., 2022).

- **Refinement Q (Improved Resolution).** The deposition achieves higher reported resolution than any prior deposition for the same protein. Depositions are ordered chronologically for each protein, and the cumulative best (minimum) resolution is tracked. $Q_d = 1$ if d is an intensive-margin deposition with a reported resolution strictly lower (better) than the prior cumulative best. Depositions without a reported resolution (common for NMR structures) cannot trigger this flag. Within the empirical coding, Q is reserved for improved reported resolution; a change in experimental method that does not improve resolution is not sufficient on its own. Higher resolution is scientifically important because going from, say, 3.5 Å to 1.8 Å reveals side-chain conformations, ordered water molecules, and hydrogen positions invisible at lower resolution. Improved resolution also yields better B-factors and lower R_{free}, increasing confidence in the deposited coordinates. Orthogonal methods—such as neutron diffraction, which directly places hydrogen atoms critical for understanding enzyme mechanisms—and re-refinement of existing data with newer software (Terwilliger et al., 2014) are recurrent sources of resolution-improving depositions.

- **Corroboration (Corr.).** The deposition triggers none of the four action flags above: it is an intensive-margin entry with no new residue coverage, no new conformational state, no biologically relevant ligand, and no resolution improvement relative to prior depositions of the same protein. Corroborating depositions account for roughly one quarter of the PDB. Typical examples include re-refinements of existing data with newer software, alternative crystal forms or space groups, NMR structures (which often lack a resolution field and therefore cannot register as Q), structures containing

only crystallographic additives (water, sulfate, glycerol), and independent validations by a second laboratory. While these depositions do not advance the structural frontier by any of our measurable criteria, they nonetheless contribute to science by confirming prior results, providing alternative crystal packing environments, and validating that deposited coordinates are reproducible.

A.2.4 Type of deposition by training value

The action-based taxonomy above classifies depositions by what they reveal about a specific protein. A complementary question is what each deposition contributes to the training corpus of the next generation of structure predictors. Recent and forthcoming models—AlphaFold3, AF-Multimer, RoseTTAFold2NA—extend coverage to protein–nucleic-acid complexes, ligand binding, large assemblies, and conformational ensembles—domains in which AlphaFold2 either has a weak signal or no template. We construct ten binary deposition flags, grouped into three thematic axes, that capture distinct dimensions of training-data value. Each flag is defined at the level of a single (protein, PDB entry) deposition and is summed within protein-year cells to form the count outcomes in Table 9.

Novelty. These flags identify depositions that add structural information previously absent from the public corpus, applied at three nested scopes: within the protein (*New conformation*), within its sequence family (*First in cluster*), and across the entire archive (*Novel structure*).

- **New conformation.** The deposition is the first entry in a new Foldseek conformational cluster for the same protein. All PDB entries mapped to the same UniProtKB accession are pairwise compared by TM-score, grouped via single-linkage clustering at TM-score ≥ 0.85, and a deposition triggers *New conformation* when its three-dimensional shape is sufficiently distinct (TM < 0.85) from all previously deposited structures of that protein. New conformational states are central training data for predictors that model conformational ensembles rather than a single static structure: a protein observed in only one conformation cannot teach a model how it transitions between states. This flag corresponds to the S component in the deposition-category taxonomy.

- **First in cluster.** The deposition is the first PDB entry for any member of the protein's MMseqs2 sequence cluster (50% identity, 80% coverage of the shorter sequence), and it is the protein's own extensive-margin deposition. Such depositions extend the experimental structural coverage of sequence space: they bring structural information to families of proteins that previously had none. Sequence families without a representative structure are precisely where a sequence-to-structure model has the least information to learn from, so a single new representative provides a usable starting model for every other member of the family.

- **Novel structure.** The deposited structure has a Foldseek TM-score below 0.5 to every prior PDB structure released in any earlier year, computed via the all-vs-all

structural-alignment search described in Section A.2.5 (van Kempen et al., 2024). A TM-score below 0.5 is the conventional threshold below which two structures are considered to adopt different folds (Xu and Zhang, 2010). Novel structures are the dominant source of new fold-level experimental training data: each newly observed fold expands the region of structural space represented in the corpus.

Biological richness. These flags identify depositions containing types of information that AlphaFold2 cannot derive from sequence alone but that next-generation models explicitly aim to predict. The two flags capture complementary facets of ligand work: *Binding affinity* isolates the subset for which the strength of binding is measured (a number), while *Biological ligand* isolates the subset for which a biologically meaningful ligand is present (an identity), after stripping out crystallographic additives.

- **Binding affinity.** The deposition has at least one quantitative binding-affinity measurement (K_d, K_i, or IC_{50}) recorded in the RCSB `rcsb_binding_affinity` field. The $K_d/K_i/IC_{50}$ trio is the canonical pharmacology measure of binding strength and is the same channel through which the PDBbind (Wang et al., 2004) and BindingDB (Liu et al., 2007b) databases aggregate experimentally determined protein–ligand affinities from the PDB. Quantitative ligand-binding pairs are the central training input for predictors that model small-molecule binding: a structure deposited without a measured affinity teaches a model what the ligand-bound complex looks like but not what binding strengths look like.

- **Biological ligand.** The deposition contains at least one non-polymer ligand of biological interest, after excluding crystallographic additives (solvents, cryoprotectants, buffer components, common ions, and heavy-atom phasing labels). The exclusion list comprises water, sulfate, glycerol, polyethylene-glycol fragments, ethylene glycol, acetate, dimethyl sulfoxide, common buffer molecules (Tris, MES, HEPES, imidazole, β-mercaptoethanol, DTT, formate, tartrate), heavy-atom labels (Br, I, Hg), and common ions (Cl, Na, Mg, Zn, Ca, K, F, NH_4, NO_3, PO_4). The filtered ligand set isolates biological substrates, cofactors, inhibitors, and drug-like molecules: the substrate-bound structures that future models of protein–ligand interaction must learn from. The exclusion logic parallels the curation conventions of BioLiP (Yang et al., 2013) and PLIP (Salentin et al., 2015), which apply analogous filters to separate biologically relevant ligands from crystallographic additives in PDB-derived protein–ligand datasets.

Relation to the action-based taxonomy. The training-value classification overlaps with the action-based taxonomy in Section A.2.3 but is not redundant. The mechanistic action *SL* is, by construction, the disjunction of *New conformation* (its *S* component) and the broader RCSB ligand flag (a superset of *Biological ligand*), so the aggregated mechanistic-count result in our action-based regressions reflects the sum of the conformation and ligand training-value channels.

A.2.5 Tree Construction

Proteins evolve by mutation, so two proteins that share a common ancestor will have similar amino acid sequences. Percent identity (pident) counts the fraction of aligned residue positions that match; higher values indicate more recently diverged, closely related proteins. The structural knowledge frontier described in Section 4.3 rests on an estimated tree that organises all SwissProt proteins by sequence similarity. This object is used as a reduced-form geometry for local homology spillovers; it is not intended to recover a literal gene tree, species tree, or complete evolutionary history. Construction proceeds in several steps.

Step 1: Clustering. We cluster the ∼573K SwissProt proteins using MMseqs2 (Steinegger and Söding, 2017) at 50% sequence identity and 80% coverage of the shorter sequence, producing ∼127K clusters. The threshold matches the boundary of reliable homology modelling: within a cluster, resolving any one member provides a usable structural template for all others, so the cluster is the natural unit at which structural knowledge transfers. MMseqs2 selects representatives via a greedy set-cover algorithm that processes proteins longest-first, so the representative is typically the longest sequence in the cluster.

Step 2: Inter-cluster distances. We run an MMseqs2 all-vs-all search among the ∼127K cluster representatives. For each pair with detectable similarity ($\geq$ 30% identity), we record the per cent identity, yielding ∼1.4M inter-cluster pairs out of ∼8 billion possible. Sequence identity values are converted to distances $D_{ij} = 1 - \text{pident}_{ij}/100$, so that a pair at 90% identity has $D = 0.10$ and a pair at the detection threshold of 30% has $D = 0.70$. Clusters with no detectable similarity receive no edge—they are *isolated* in the sequence similarity network.

The practical thresholds:

Tree distance	Pident equivalent	modelling quality
$d \approx 0$	Same cluster ($\geq$50%)	High-quality homology model
$d \in (0, 0.3)$	70–100% via path	Good model, minor adjustments needed
$d \in (0.3, 0.5)$	50–70% via path	Moderate model, loops unreliable
$d \in (0.5, 0.7)$	30–50% via path	"Twilight zone"—fold correct, details wrong
$d > 0.7$ or isolated	<30%	"Midnight zone"—homology modelling fails

Step 3: Tree construction. We organise the clusters into a tree using the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987), which constructs an additive tree from a distance matrix by iteratively joining the two closest nodes (by the Q-criterion) to minimise total tree length. Leaves are clusters, branch lengths reflect relative sequence dissimilarity, and internal nodes (Steiner nodes) are algorithmic intermediaries that make the similarity geometry additive. Unmeasured pairwise distances (pairs below 30% identity) are set to

$D = 1.0$. Formally, the input distance matrix is

$$D_{ij} = \begin{cases} 1 - \text{pident}_{ij}/100 & \text{if } (i,j) \text{ has a direct MMseqs2 measurement } (\geq 30\%), \\ 1 & \text{if no measurement is available,} \\ 0 & i = j, \end{cases}$$

so every branch length the NJ algorithm assigns is a pident-based dissimilarity, and the flat $D = 1$ fallback for undetectable pairs caps—rather than path-inflates—the distance between sequences too diverged for MMseqs2 to align. The cluster-level frontier distance used below is then a *patristic* distance: the sum of these pident-weighted branch lengths along the unique tree path from a cluster to the nearest cluster containing a pre-2017-solved protein, traversing the intervening Steiner nodes. Because the distance matrix is sparse—only ∼1.4M of ∼8 billion possible pairs have detectable similarity—the resulting structure is a forest of connected components rather than a single tree.

The NJ tree has ∼210K nodes (127K cluster leaves + ∼83K Steiner nodes) and ∼175K edges:

- *Giant component* (∼48K clusters): NJ applied to all clusters in the largest connected component.
- *Small components* (2–530 clusters each): NJ applied to each component independently.
- *Isolated clusters* (∼28K): No tree edges; these appear in the node table but have no connections.

Step 4: From clusters to proteins. The tree is defined over clusters, but the analysis requires protein-level distances. Each cluster has a hub-and-spoke topology: the representative protein sits at the hub, and all other members connect to it through a spoke of length

$$d_{\text{intra}}(p) = 1 - \text{pident}(p, \text{rep})/100,$$

where $\text{pident}(p, \text{rep})$ is protein p's sequence identity to its cluster representative. For representatives and singleton clusters, $d_{\text{intra}} = 0$.

Step 5: Dynamic overlay of PDB. Sequence similarity is time-invariant—determined by biology, not by experiments. What changes over time is which nodes have been experimentally resolved. The tree provides the fixed geography; the time-varying information is the set of clusters with PDB coverage in each year $t \in \{1980, \ldots, 2025\}$. A two-pass dynamic programming algorithm propagates nearest-solved distances through the tree: the first pass (bottom-up) finds the nearest solved descendant at each node, the second pass (top-down) finds the nearest solved node outside each subtree, and the two are combined to yield the global nearest-solved distance. Structural knowledge is cumulative—depositions are never removed—so the solved set expands monotonically.

Step 6: Protein-level distance measures. To reach any point outside the cluster, protein p first traverses its spoke to the representative, then follows the inter-cluster tree. The protein-level distance is *selected*—not minimised—according to whether p's own cluster already contains a solved member by year t:

$$D_{p,t} = \begin{cases} \underbrace{d_{\text{intra}}(p) + d_{\text{within}}(p,t)}_{\text{within-cluster path}}, & \text{if } p\text{'s cluster has a solved member by year } t, \\ \underbrace{d_{\text{intra}}(p) + d^{\text{cluster}}_{\text{tree}}(t)}_{\text{inter-cluster path}}, & \text{otherwise,} \end{cases}$$

where $d_{\text{within}}(p,t)$ is the spoke distance from p's representative to the nearest solved member of the same cluster, and $d^{\text{cluster}}_{\text{tree}}(t)$ is the tree-path distance from the representative to the nearest solved cluster in year t. We use the within-cluster path whenever the cluster is itself solved because a solved cluster has cluster-level tree distance $d^{\text{cluster}}_{\text{tree}}(t) = 0$; the inter-cluster expression would then collapse to $d_{\text{intra}}(p)$ and understate how far p actually sits from the nearest resolved sequence.

The gap measure captures the width of the unsolved region around a protein:

$$G_{p,t} = d^{\text{cluster}}_{\text{left}}(t) + d^{\text{cluster}}_{\text{right}}(t) + 2 \cdot d_{\text{intra}}(p),$$

where d_{left} and d_{right} are the tree distances from p's cluster to the nearest solved cluster reached by exiting p's parent edge in each of the two directions: d_{left} is the nearest solved cluster on one side of p in the tree, d_{right} the nearest on the other side. By construction, the two anchors sit on different sides of p, so the gap captures the width of the unsolved sequence space around the protein rather than double-counting a single direction. When p's own cluster has no solved member, the smaller external arm determines the inter-cluster component of $D_{p,t}$. When the cluster is already solved, the headline nearest-distance measure instead uses the within-cluster path. A wide gap means a large span of sequence space around the protein lacks a usable structural template, even when one local anchor is close.

For proteins in isolated clusters—those with no detectable sequence similarity ($\geq 30\%$ identity) to any other cluster—no tree path to a solved neighbour exists. We max-impute these at the largest nearest-solved tree-path distance observed in the year-t cross-section (the sample maximum, $D_{p,t} = 1.773$ in 2016), so the isolated point mass sits at the frontier edge of the distribution rather than in its interior.[27] The gap is undefined for isolated clusters (no second flanking neighbour), and tree-derived features such as bridging and betweenness are set to zero. When a non-isolated cluster belongs to a tree component in which no cluster has yet been solved in year t, the nearest-solved distance is set to the component diameter, and the gap is undefined until a second solved cluster appears.

We also construct a local knowledge density measure. The *cluster PDB share* is the fraction of cluster members with any PDB structure by year t. Even when a cluster's

[27] The tree-path distance sums branch lengths along the route between clusters, so it can exceed 1; the value 1.0 imputed elsewhere applies instead to the individual *pairwise* distances between clusters (the inputs to the tree), which are each capped at 1.

representative has been solved, individual members may lack coverage for their specific sequence variants (insertions, deletions, mutations), so a higher share indicates better coverage of the cluster's sequence diversity. The share is recomputed for each year as the solved set expands.

Step 7: Direct pairwise distance measures at the 2016 snapshot. The path-based $D_{p,t}$ of Step 6 is the headline frontier measure used in Section 4.3 and in the binned and continuous frontier specifications of Section 7.2. We complement it with a family of direct pairwise distance measures, fixed at the 2016 cutoff, which serve as robustness alternatives. The construction runs a single MMseqs2 search of all ∼574K SwissProt query proteins against the target set $\mathcal{S}_{2016}$ of ∼38K SwissProt proteins with at least one PDB structure released on or before 31 December 2016, at MMseqs2 default sensitivity $s = 5.7$ and `-max-seqs 100`. The raw output is a BLAST-tab `.m8` table giving, for each query–target alignment that passes the MMseqs2 prefilter, the percent identity (pident), alignment length (alnlen), and bitscore (b). For each query p, we retain a single best alignment, ranked by pident with b as a tiebreaker, and write the per-protein file `distance_measures2016.parquet`.

Three pairwise distance *families* are then derived from each query's best hit. The **percent-identity distance** is

$$\widetilde{d}_p^{\text{pid}} \;=\; 1 \;-\; \tfrac{1}{100}\,\text{pident}_p^{\star},$$

recorded as `pid_dist_2016`, where $\text{pident}_p^{\star} = \max_{q \in \mathcal{S}_{2016}} \text{pident}_{p,q}$ is the percent identity of p's best hit. The **bitscore distance** is

$$\widetilde{d}_p^{\text{bit}} \;=\; \frac{1}{\log(2 + b_p^{\star})},$$

recorded as `bitscore_dist_2016`, where $b_p^{\star}$ is the bitscore of p's best hit. The MMseqs2 bitscore corrects pident for alignment length and database size, so $\widetilde{d}_p^{\text{bit}}$ ranks queries differently from $\widetilde{d}_p^{\text{pid}}$ whenever short or low-complexity alignments inflate pident; the $1/\log(2+\cdot)$ transform compresses the long upper tail of bitscores into the unit interval. The **coverage-weighted percent-identity distance** is

$$\widetilde{d}_p^{\text{align}} \;=\; \widetilde{d}_p^{\text{pid}} \cdot (1 - f_p), \qquad f_p \;=\; \min\Big(1,\; \text{alnlen}_p^{\star}/L_p\Big),$$

recorded as `pid_align_dist_2016`, where L_p is p's sequence length. A 90%-identity hit covering only 20% of the query is treated as substantially more distant than one with the same per cent identity covering the full sequence: the coverage term penalises short partial alignments, which is otherwise a known weak point of raw pident at protein scale. Queries whose MMseqs2 search returns no hit above the prefilter against $\mathcal{S}_{2016}$ are flagged `isolated_pid_2016` $= 1$ and max-imputed at $\widetilde{d}_p^{\text{pid}} = 1$, $\widetilde{d}_p^{\text{bit}} = 1$, and $\widetilde{d}_p^{\text{align}} = 1$.

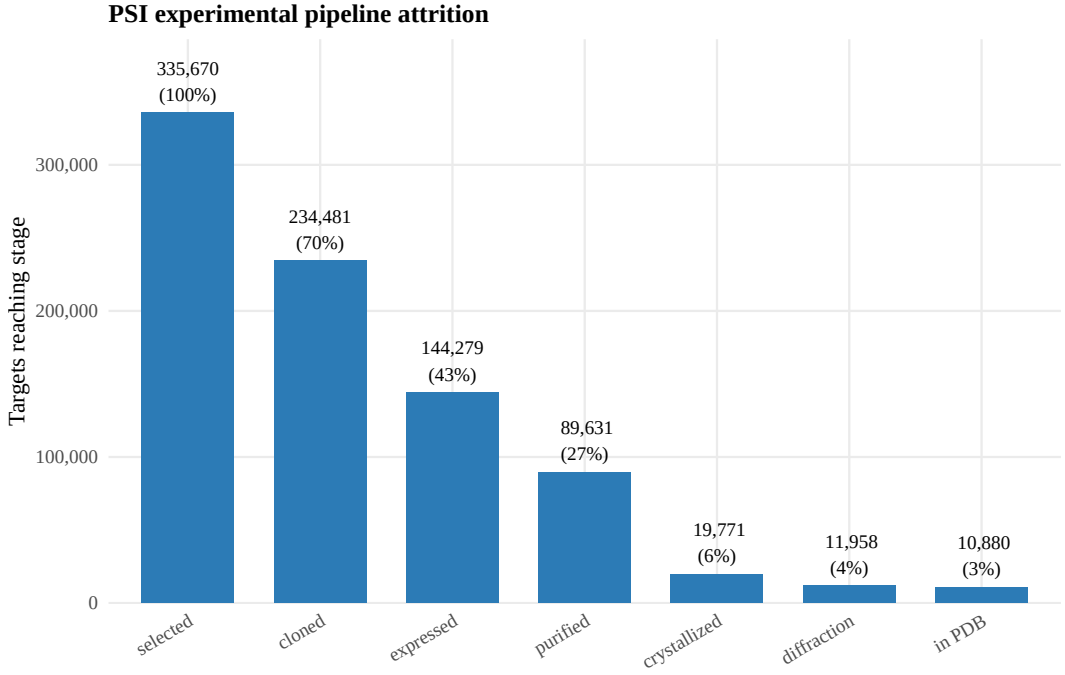
Each of the three families has a corresponding tree-anchored *gap* measure obtained by replacing the single nearest-solved target by the two tree-flanking anchors of p on the 2016 neighbour-joining tree (Step 6), $y^{-}_{2016}(p)$ and $y^{+}_{2016}(p)$. For each anchor, the per-arm

distance is computed from the best MMseqs2 hit between p and the anchor in the same `.m8` table, under the family's transform: $1 - \text{pident}/100$ for `pid`, $1/\log(2+b)$ for `bitscore`, and the coverage-weighted variant $(1 - \text{pident}/100) \cdot (1 - f_{\text{arm}})$ for `pid_align`. The two arms are summed to yield the gap, recorded as `pid_gap_2016`, `bitscore_gap_2016`, and `pid_align_gap_2016`. When the tree assigns p an anchor on a given side but MMseqs2 reports no hit between p and that anchor, the arm is max-imputed at the family's worst value, and the indicator `anchor_left_unobserved_2016` or `anchor_right_unobserved_2016` is set to one; every regression that uses a gap measure controls for these two flags.

The three direct pairwise families serve distinct robustness roles relative to the tree-path headline. `pid_dist_2016` guards against any inflation of distance from path summation when the nearest pre-2017-solved protein is reached only through intermediate clusters on the tree, replacing the path with the direct query–target identity; `bitscore_dist_2016` guards against the dependence on raw pident as the alignment-quality summary; `pid_align_dist_2016` guards against the leniency of pident on short partial alignments; and the three tree-anchored gap measures collectively guard against the choice of a single nearest-solved target as the frontier proxy by integrating over the protein's two-sided sequence-space neighbourhood. The headline and all alternative measures are constructed once, at the 2016 cutoff, and held fixed throughout the panel.

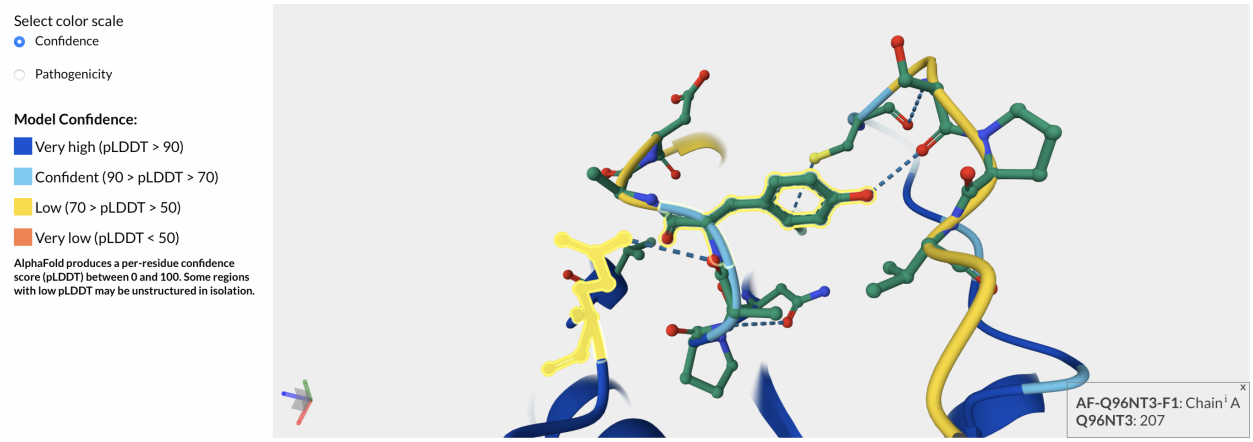
A.3 Additional Descriptives

Figure 9: PSI experimental funnel



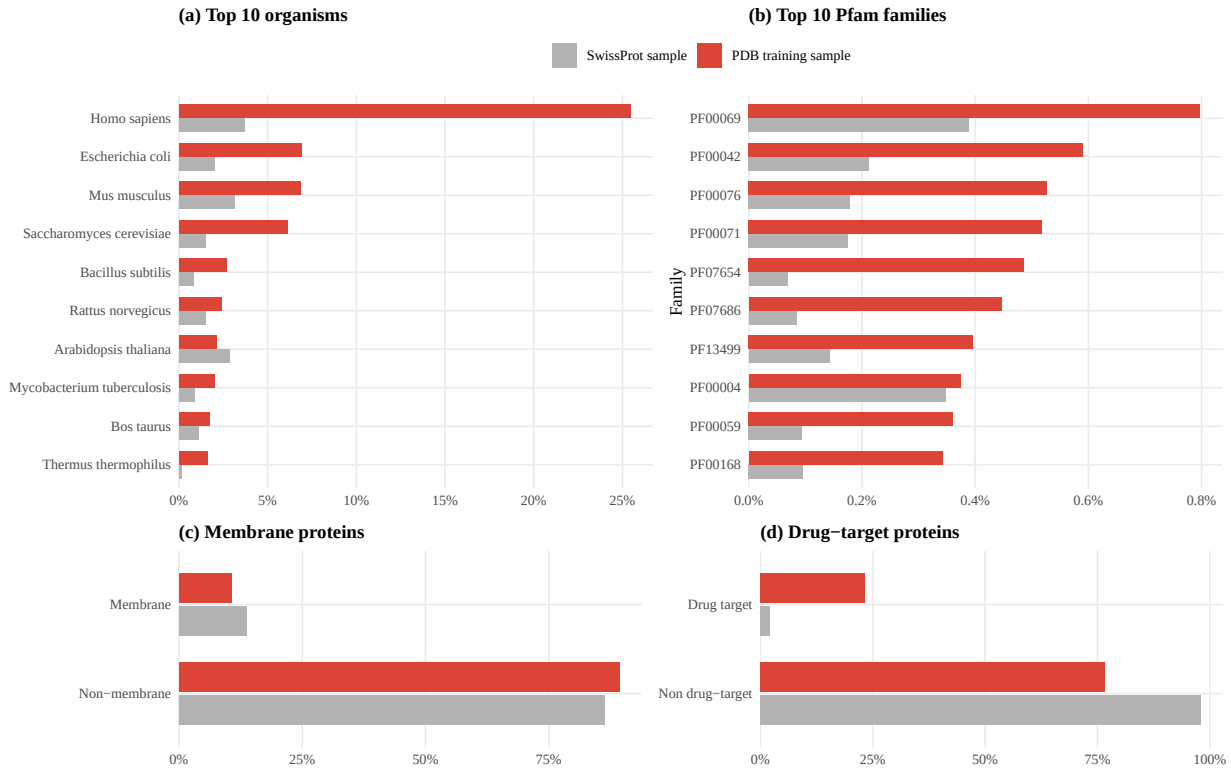
Notes. Number of targets reaching each pipeline stage. Percentages show cumulative survival relative to selected targets.

Figure 10: AlphaFold2 confidence



Notes. Screenshot of AlphaFold2 confidence zoomed on the residues.

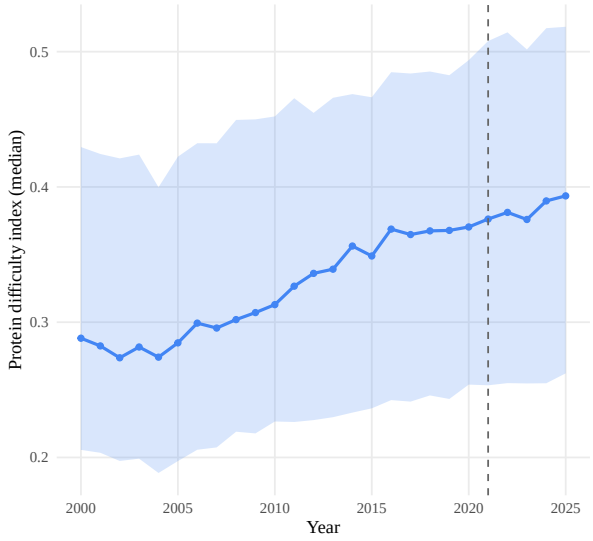
Figure 11: Non-random set of proteins in the PDB training sample



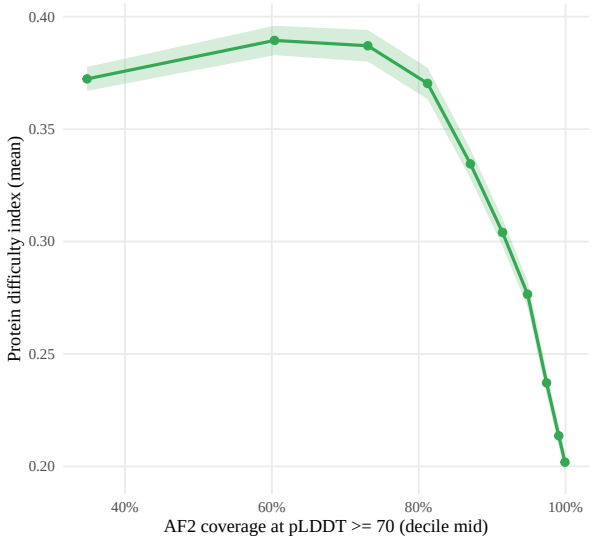
Notes. Each figure plots the share of SwissProt proteins (in grey) and the share of proteins in the pre-2018 PDB (in red) across different dimensions: top 10 organisms (figure a), top 10 Pfam families (figure b), membrane proteins (figure c), and drug-target proteins (figure d).

Figure 12: Protein difficulty: time trend and relationship to AF coverage

(a) Median difficulty of new depositions over time

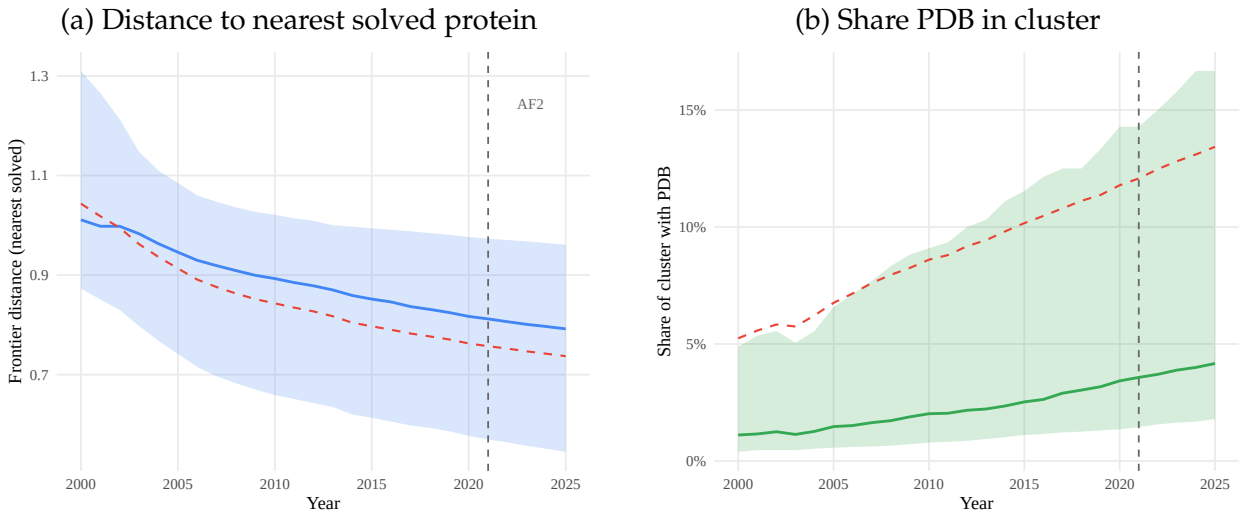


(b) Mean difficulty by AF coverage decile



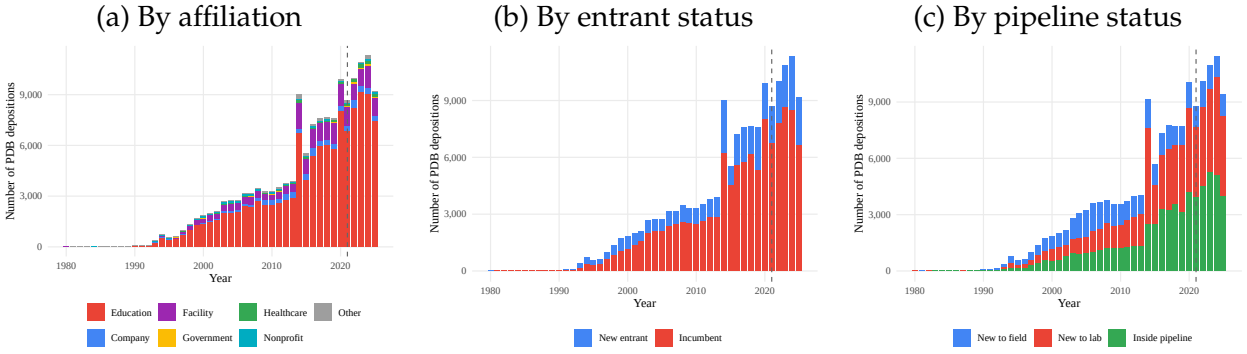
Notes. Panel (a): median value of the protein-difficulty index across PDB depositions in each release year, with the inter-quartile range shaded; the dashed vertical line marks the AF2 release in July 2021. Panel (b): mean difficulty across proteins within each AF cov_p^{70} decile (cross-section, 2017 panel), with 95% confidence intervals. AF coverage is systematically lower on harder proteins.

Figure 13: Frontier contraction over time



Notes. Figure (a) plots the average tree distance to the nearest solved protein over time. Figure (b) plots the average share of PDB-solved members within a protein's cluster over time. The solid (dashed) line represents the median (mean), and the shaded ribbon shows the interquartile range. Both panels show steady contraction of the structural knowledge frontier from 2000 onward.

Figure 14: Annual PDB depositions by lab affiliation, entrant status, and pipeline



Notes. Annual PDB depositions decomposed by depositing lab's primary affiliation (academic, industry, government, other) (Figure a), entrant status (incumbent labs with at least one pre-2017 deposit vs. entrant labs) (Figure b), and pipeline status (deposition on a protein the lab had previously worked on vs. a protein new to the lab) (Figure c). The dashed line marks the AF2 release in 2021.

Table 13: Lab-margin concentration

Lab rank	All deposits	First-deposition	Follow-up
Top 1%	35.3%	35.1%	37.3%
Top 5%	59.8%	60.6%	62.2%
Top 10%	70.6%	73.9%	72.5%
Top 25%	84.2%	91.8%	85.5%
Top 50%	93.0%	100.0%	94.1%

Notes. Cumulative share of depositions accounted for by the top X% of labs, where labs are ranked separately within each column by their count in that margin (first depositions or follow-up depositions).

Table 14: Depositions by within-protein margin and neighbourhood density

	Isolated	Sparse	Dense	Total
Panel A: All depositions, 2017–2025 ($N = 93,522$)				
First-deposition	4,583 (4.9%)	2,382 (2.5%)	3,814 (4.1%)	10,779
Follow-up	11,237 (12.0%)	20,885 (22.3%)	50,621 (54.1%)	82,743
Total	15,820	23,267	54,435	93,522
Panel B: Pre-AF, 2017–2021 ($N = 45,010$)				
First-deposition	2,397 (5.3%)	1,266 (2.8%)	2,133 (4.7%)	5,796
Follow-up	4,688 (10.4%)	9,745 (21.7%)	24,781 (55.1%)	39,214
Total	7,085	11,011	26,914	45,010
Panel C: Post-AF, 2022–2025 ($N = 48,512$)				
First-deposition	2,186 (4.5%)	1,116 (2.3%)	1,681 (3.5%)	4,983
Follow-up	6,549 (13.5%)	11,140 (23.0%)	25,840 (53.3%)	43,529
Total	8,735	12,256	27,521	48,512

Notes. The sample consists of all depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. The within-protein margin (rows): first deposition for the protein, number of follow-up depositions, and all depositions. The neighbourhood bin (columns): Each protein is assigned its 2016 distance to the nearest PDB-solved cluster, D_p^{2016}. Non-isolated proteins are split at the 33rd and 67th percentile of distance on the regression-sample panel (*Dense*: $D_p^{2016} \leq 0.70$; *Sparse*: $0.70 < D_p^{2016} \leq 0.95$; *Isolated*: $D_p^{2016} > 0.95$, plus singleton-cluster proteins with no inter-cluster tree edge). Cells report counts and share of the panel total in parentheses.

A.4 Additional Results

A.4.1 Distance to the frontier

Table 15: Composition of follow-up depositions by content components

	(1)	(2)	(3)	(4)	(5)	(6)
	# Static		# Mechanistic		# Corrob.	
AF cov × Post	0.005	0.353***	0.547***	0.748***	0.758*	0.905
	(0.083)	(0.119)	(0.127)	(0.167)	(0.445)	(0.629)
Post × Dist		0.615***		0.414***		1.537**
		(0.110)		(0.144)		(0.686)
Post × AF cov × Dist		−0.561***		−0.416***		−1.063
		(0.128)		(0.158)		(0.761)
Year FE	X	X	X	X	X	X
Protein FE	X	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X	X
N	163,845	163,845	164,466	164,466	36,954	36,954
$\tilde{R}^2$	0.137	0.137	0.537	0.538	0.531	0.532
N proteins	22,578	22,578	22,578	22,578	22,578	22,578
N structures	68,001	68,001	68,001	68,001	68,001	68,001
$\bar{Y}_{t\leq 2021}$	0.156	0.156	0.528	0.528	0.086	0.086
Effect (full range)	+0.5%	+42.3%	+72.8%	+111.3%	+113.3%	+147.1%
Effect (1 SD)	+0.1%	+6.9%	+10.9%	+15.3%	+15.5%	+18.7%
Effect (Q4 vs Q1)	+0.2%	+14.8%	+23.8%	+34.0%	+34.5%	+42.4%

Notes. This table reports the coefficients from Equation (12) estimated by PPML on the annual PDB depositions by content components between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. The outcomes are the annual number of scaffold, mechanistic and corroboration follow-up depositions on the intensive sample (proteins with at least one PDB before 2017). AF cov_p is the share of a protein's residues predicted at pLDDT $\geq$ 70 and Post_t is an indicator for years after 2021. All columns include year FE, plus the pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post_t. We also include a triple interaction with the pre-2017 distance to the nearest solved protein, all corresponding lower-level interaction in addition to an isolated $\times \text{Post}_t$ interaction (a control entered so the distance triple identifies off non-isolated proteins, not itself a triple). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 16: Composition of all depositions by content components (full sample)

	(1) # Static	(2) # Mechanistic	(3) # Corrob.
AF cov × Post	0.005	0.547***	0.758*
	(0.083)	(0.127)	(0.445)
Year FE	X	X	X
Protein FE	X	X	X
Pre-AF features × Post	X	X	X
Sample	All	All	All
N	163,845	164,466	36,954
Eff. proteins	18,205	18,274	4,106
$\tilde{R}^2$	0.137	0.537	0.531
Raw N	4,774,455	4,774,455	4,774,455
N proteins	530,495	530,495	530,495
N structures	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.012	0.028	0.004
Effect (full range)	+0.5%	+72.8%	+113.3%
Effect (1 SD)	+0.1%	+10.9%	+15.5%
Effect (Q4 vs Q1)	+0.2%	+23.8%	+34.5%
Implied dep./yr	+35	+10,878	+2,426

Notes. PPML estimates on the full no-batch, no-SARS-CoV-2 protein × year panel (SwissProt, 2017–2025; not restricted to proteins with a pre-2017 PDB). Outcomes are counts of depositions by action type: Static (R/Q: completeness or resolution refinement), Mechanistic (S/L: new conformational state or ligand complex), Corroboration (C: independent validation of a previously solved scaffold). AF cov is the share of a protein's residues predicted at pLDDT ≥ 70 and Post is an indicator for years after 2021. All columns include protein and year FE, in addition to seven pre-AF features interacted with post (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count). Magnitude rows report $(\exp(\hat{\beta} \cdot x) - 1) \times 100$. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 17: Training value

# depositions by type:	(1)	(2) Novelty	(3)	(4) Biological richness	(5)
	New conformation	First in cluster	Novel structure	Binding affinity	Bio ligand
AF cov × Post	1.008***	0.101	−0.582	−0.643	0.353***
	(0.175)	(0.332)	(0.476)	(0.469)	(0.134)
Post × Dist	0.697***	0.525**	0.318	−1.091*	0.440***
	(0.139)	(0.228)	(0.361)	(0.615)	(0.120)
Post × AF cov × Dist	−0.871***	−0.626**	−0.485	1.712**	−0.336**
	(0.157)	(0.246)	(0.412)	(0.767)	(0.136)
Year FE	X	X	X	X	X
Protein FE	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X
N	115,596	52,587	8,532	7,872	108,090
$\tilde{R}^2$	0.123	0.005	0.085	0.261	0.407
N proteins	530,495	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.003	0.001	0.000	0.001	0.010
Effect (full range)	+174.1%	+10.6%	-44.1%	-47.4%	+42.4%
Effect (1 SD)	+21.1%	+1.9%	-10.5%	-11.5%	+6.9%
Effect (Q4 vs Q1)	+48.3%	+4.0%	-20.3%	-22.2%	+14.8%

Notes. This table reports the coefficients from Equation (12) on counts of depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by training value: Col (1) new conformation - TM $\geq$ 0.85 cluster never previously observed for this protein, Col (2) first in cluster based on MMseqs2, Col (3) novel structure = 1 - best TM to any prior PDB above 0.5, Col (4) binding affinity $\geq$ 1 Kd/Ki/IC50 record, and Col (5) Bio ligand $\geq$ 1 non-additive ligand. AF cov_p is the share of a protein's residues predicted at pLDDT $\geq$ 70 and Post_t is an indicator for years after 2021. D_p^{2016} is the pre-2017 distance to the nearest solved protein. All columns are estimated by PPML and include protein and year FE in addition to the pre-AF features interacted with Post. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. We also include a triple interaction with the pre-2017 distance to the nearest solved protein, all corresponding lower-level interaction in addition to an isolated $\times \text{Post}_t$ interaction (a control entered so the distance triple identifies off non-isolated proteins, not itself a triple). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 18: Heterogeneity by laboratory pipeline

	(1)	(2)	(3)	(4)	(5)	(6)
	First-deposition		# New to lab		# Inside	
AF cov × Post	-0.001^{***}	0.004^{***}	0.458^{***}	0.227^{*}	0.865^{***}	0.859^{***}
	(0.000)	(0.001)	(0.137)	(0.130)	(0.174)	(0.239)
Post × Dist		0.004^{***}		-0.328		0.140
		(0.001)		(0.201)		(0.342)
Post × AF cov × Dist		-0.004^{***}		0.572^{**}		-0.056
		(0.001)		(0.249)		(0.378)
Year FE	X	X	X	X	X	X
Protein FE			X	X	X	X
Family FE	X	X				
Pre-AF features	X	X				
Pre-AF features × Post	X	X	X	X	X	X
N	4,365,071	4,365,071	59,706	59,706	28,548	28,548
R^2 / $\tilde{R}^2$	0.021	0.021	0.464	0.464	0.417	0.417
N proteins	507,917	507,917	22,578	22,578	22,578	22,578
N structures	14,370	14,370	68,001	68,001	68,001	68,001
$\bar{Y}_{t\leq 2021}$	0.002	0.002	0.182	0.182	0.099	0.099
Effect (full range)	-0.1 p.p.	+0.4 p.p.	+58.1%	+25.4%	+137.5%	+136.0%
Effect (1 SD)	-0.0 p.p.	+0.1 p.p.	+9.1%	+4.4%	+17.8%	+17.7%
Effect (Q4 vs Q1)	-0.0 p.p.	+0.1 p.p.	+19.6%	+9.3%	+40.2%	+39.9%

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. Cols (1) and (2) LPM/discrete-time hazard for the first-deposition indicator on the at-risk sample on proteins without any PDB before 2017; cols (3)–(6) PPML on the annual number of depositions new to the lab's pipeline / inside the lab's pipeline on the intensive sample (proteins with at least one PDB before 2017). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. Cols (2), (4), and (6) include a triple interaction with the pre-2017 distance to the nearest solved protein, all corresponding lower-level interaction in addition to an isolated $\times \text{Post}_t$ interaction (a control entered so the distance triple identifies off non-isolated proteins, not itself a triple). Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. All columns include year FE, plus the pre-AF features interacted with Post_t. Cols (3)–(6) add protein fixed effects, while cols (1)–(2) add the pre-AF features in levels and primary-Pfam family FE (dropping log pre-2017 PDB count and log pre-2017 cryo-EM PDB count as both are mechanically zero on the at-risk sample). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 19: Heterogeneity by laboratory type

	(1)	(2)	(3)	(4)
	# academic		# industry	
AF cov × Post	0.446^{***}	0.612^{***}	0.891^{**}	0.568
	(0.137)	(0.185)	(0.355)	(0.460)
Post × Dist		0.471^{***}		-0.107
		(0.162)		(0.624)
Post × AF cov × Dist		-0.390^{**}		0.613
		(0.179)		(0.717)
Year FE	X	X	X	X
Protein FE	X	X	X	X
Pre-AF features × Post	X	X	X	X
N	140,508	140,508	13,752	13,752
$\tilde{R}^2$	0.533	0.533	0.313	0.314
N proteins	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.024	0.024	0.001	0.001
Effect (full range)	+56.2%	+84.4%	+143.8%	+76.6%
Effect (1 SD)	+8.8%	+12.3%	+18.4%	+11.4%
Effect (Q4 vs Q1)	+19.0%	+27.0%	+41.7%	+24.9%

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by lab type (academic cols (1)–(2) and industry cols (3)–(4)). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. All columns are estimated by PPML on the number of depositions by academic labs (cols 1 and 2) and industry labs (cols 3 and 4). All columns include protein and year FE, and seven pre-AF features interacted with Post_t (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count). Cols 2 and 4 include a triple interaction with the pre-2017 distance to the nearest solved protein, all corresponding lower-level interaction in addition to an isolated $\times \text{Post}_t$ interaction (a control entered so the distance triple identifies off non-isolated proteins, not itself a triple). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 20: Heterogeneity by experimental method

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	# X-ray		# Cryo-EM		# NMR		# Other	
AF cov × Post	0.296	0.429	0.519***	0.697***	−0.422	0.240	1.933	1.337
	(0.196)	(0.280)	(0.119)	(0.172)	(0.347)	(0.463)	(1.278)	(2.085)
Post × Dist		0.169		0.715***		0.252		−0.617
		(0.269)		(0.166)		(0.431)		(3.492)
Post × AF cov × Dist		−0.269		−0.485***		−1.103**		2.043
		(0.296)		(0.186)		(0.534)		(4.442)
Year FE	X	X	X	X	X	X	X	X
Protein FE	X	X	X	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X	X	X	X
N	95,409	95,409	95,139	95,139	8,199	8,199	432	432
$\tilde{R}^2$	0.490	0.490	0.602	0.602	0.122	0.123	0.401	0.407
N proteins	530,495	530,495	530,495	530,495	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.016	0.016	0.017	0.017	0.000	0.000	0.000	0.000
Effect (full range)	+34.4%	+53.6%	+68.0%	+100.9%	-34.4%	+27.1%	+591.1%	+280.8%
Effect (1 SD)	+5.8%	+8.5%	+10.4%	+14.2%	-7.7%	+4.7%	+44.3%	+28.9%
Effect (Q4 vs Q1)	+12.3%	+18.3%	+22.5%	+31.3%	-15.2%	+9.8%	+112.9%	+68.7%

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction, split by method: X-ray (cols 1 and 2), cryo-EM (cols 3 and 4) and NMR (cols 5 and 6). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. All columns include protein and year FE, plus the pre-AF features interacted with Post_t. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.4.2 Measuring Deepening

Table 21: Frontier interaction

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	# depositions				First-deposition (at-risk)				# follow-up			
	Dist	Gap	Share	All	Dist	Gap	Share	All	Dist	Gap	Share	All
Post × AF cov	0.445***	0.574***	0.265**	0.574***	0.004***	0.006***	-0.002***	0.005**	0.339***	0.475***	0.360**	0.299
	(0.120)	(0.155)	(0.112)	(0.194)	(0.001)	(0.002)	(0.000)	(0.002)	(0.122)	(0.170)	(0.176)	(0.206)
Post × AF cov × Dist.	-0.235**			-0.188	-0.004***			-0.004***	0.370*			0.489**
	(0.112)			(0.187)	(0.001)			(0.001)	(0.209)			(0.246)
Post × AF cov × Gap		-0.127*		-0.073		-0.003***		-0.001		0.003		-0.102
		(0.073)		(0.103)		(0.001)		(0.001)		(0.107)		(0.102)
Post × AF cov × Share cluster PDB			0.160	-0.127			0.020***	0.007			0.201	0.392
			(0.225)	(0.287)			(0.005)	(0.007)			(0.292)	(0.303)
Year FE	X	X	X	X	X	X	X	X	X	X	X	X
Protein FE	X	X	X	X					X	X	X	X
Family FE					X	X	X	X				
Pre-AF features					X	X	X	X				
Pre-AF features × Post	X	X	X	X	X	X	X	X	X	X	X	X
N	171,135	171,135	171,135	171,135	4,365,071	4,365,071	4,365,071	4,365,071	72,261	72,261	72,261	72,261
R^2 / $\bar{R}^2$	0.460	0.460	0.460	0.460	0.021	0.021	0.021	0.021	0.532	0.532	0.533	0.533
N proteins	530,495	530,495	530,495	530,495	507,917	507,917	507,917	507,917	22,578	22,578	22,578	22,578
$\bar{Y}_{t\leq 2021}$	0.017	0.017	0.017	0.017	0.002	0.002	0.002	0.002	0.305	0.305	0.305	0.305

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction.
AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. All columns include protein and year FE, plus the pre-AF features interacted with Post_t. Pre-AF features comprise log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 22: Frontier interactions: AF effect by neighbourhood density bin

	(1) # depositions	(2) First-deposition	(3) # follow-up
AF cov × Post (isolated)	−0.089	−0.003***	+0.689**
	(0.129)	(0.000)	(0.285)
× sparse	+0.629***	+0.001*	+0.075
	(0.185)	(0.001)	(0.317)
× dense	+0.489***	+0.004***	−0.269
	(0.156)	(0.001)	(0.290)
AF cov (isolated)	—	+0.004***	—
		(0.000)	
× sparse	—	−0.000	—
		(0.000)	
× dense	—	+0.000*	—
		(0.000)	
Post × sparse	−0.528***	−0.001*	−0.016
	(0.164)	(0.001)	(0.290)
Post × dense	−0.505***	−0.004***	+0.223
	(0.138)	(0.001)	(0.267)
Wald χ^2_2 ($\beta_{\text{dense}} = \beta_{\text{sparse}} = \beta_{\text{isolated}}$): p	0.001	<0.001	0.144
Year FE	X	X	X
Protein FE	X		X
Family FE		X	
Pre-AF features		X	
Pre-AF features × Post	X	X	X
N	171,135	4,365,071	72,261
N structures	92,930	14,370	68,001
R^2 / $\tilde{R}^2$	0.460	0.021	0.532

Notes. This table reports the coefficients from Equation (12) on the annual PDB depositions between 2017 and 2025 (excluding batch duplicates and SARS-CoV-2) for the sample of SwissProt proteins created before 2017 with an AF2 prediction. Col (1) PPML on the annual number of depositions on the balance panel; col (2) LPM/discrete-time hazard for the first-deposition indicator on the at-risk sample on proteins without any PDB before 2017; col (3) PPML on the annual number of follow-up depositions on the intensive sample (proteins with at least one PDB before 2017). AF cov_p is the share of a protein's residues predicted at pLDDT ≥ 70 and Post_t is an indicator for years after 2021. Each protein is assigned its 2016 sequence-tree distance to the nearest solved protein, D_p^{2016}. Proteins are split at the 33rd and 67th percentile of D_p^{2016}: dense neighbourhood; sparse neighbourhood; isolated neighbourhood (reference bin). All columns include year fixed effects and the seven pre-AF features interacted with Post_t: log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count. Columns (1) and (3) add protein fixed effects; column (2) adds the pre-AF features in levels and family fixed effects because there are no protein fixed effects. The Wald χ^2_2 row tests the joint null $\beta_{\text{dense}} = \beta_{\text{sparse}} = \beta_{\text{isolated}}$ on the AF $\text{cov}_p \times \text{Post}_t$ effect; rejecting it is evidence that the AF2 effect varies with neighbourhood density. Standard errors are clustered at the protein level. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

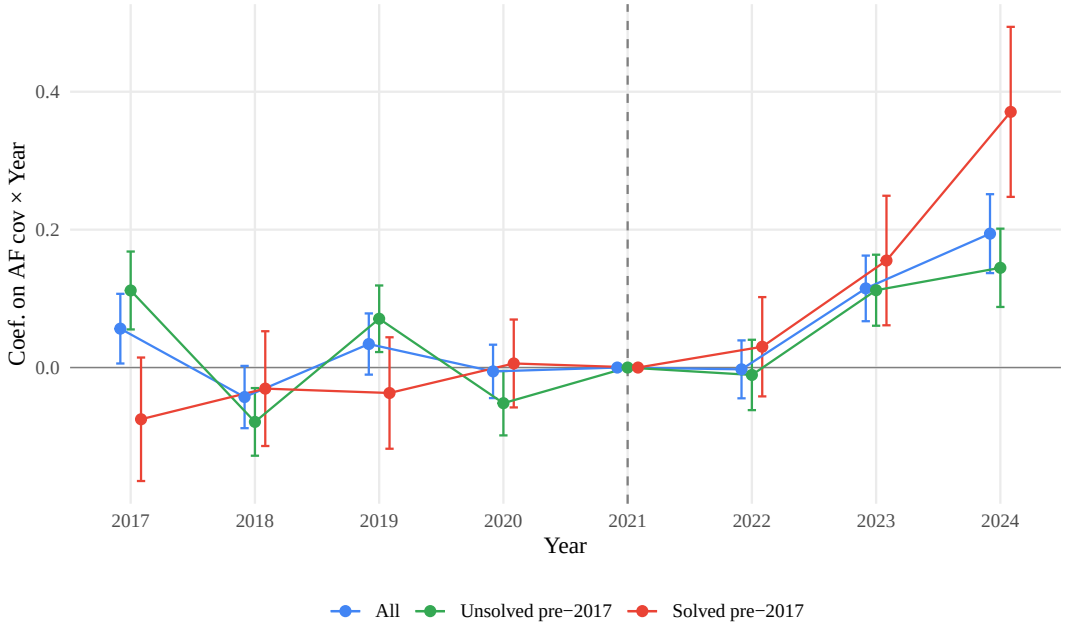
A.4.3 Publication outcomes

Table 23: Annual publication count

	(1)	(2)	(3)	(4)	(5)	(6)
	All proteins		Unsolved pre-2017		Solved pre-2017	
Post × AF cov	0.083***	0.079**	0.065***	0.093**	0.188***	0.068
	(0.021)	(0.034)	(0.020)	(0.041)	(0.042)	(0.053)
Post × Dist.		0.134***		0.088***		-0.226**
		(0.024)		(0.028)		(0.099)
Post × AF cov × Dist.		-0.022		-0.045		0.349***
		(0.035)		(0.040)		(0.128)
Protein FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
N	520,992	520,992	420,248	420,248	100,744	100,744
$\tilde{R}^2$	0.672	0.672	0.589	0.589	0.750	0.750
N proteins	511,845	511,845	489,649	489,649	22,196	22,196
N publications	826,409	826,409	482,354	482,354	344,055	344,055
$\bar{Y}_{t\leq 2021}$	0.227	0.227	0.136	0.136	2.244	2.244
% effect	+8.6%	+8.3%	+6.7%	+9.7%	+20.7%	+7.1%
% effect (SD)	+1.6%	+1.5%	+1.2%	+1.8%	+3.6%	+1.3%

Notes. This table reports the coefficients from Equation (12) on annual UniProt-curated publication counts between 2017 and 2025. Columns (1)–(2) use all SwissProt proteins created before 2017 with an AF2 prediction; columns (3)–(4) use the unsolved subsample, proteins without any PDB before 2017; and columns (5)–(6) use the solved subsample, proteins with at least one PDB before 2017. Columns (2), (4), and (6) add a triple interaction with the pre-2017 distance to the nearest solved protein in addition to an isolated indicator. All columns include protein and year FE. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure 15: AF2 effect on annual publications



Notes. Continuous event-study estimates for annual UniProt-curated publication count. The dashed line marks the AF2 release in 2021.

A.4.4 New lab entrant

Table 24: Annual lab counts

	(1) # lab	(2) # incumbent lab	(3) # entrant lab
AF cov × Post	0.165**	0.156*	0.163
	(0.079)	(0.082)	(0.102)
Year FE	X	X	X
Protein FE	X	X	X
Pre-AF features × Post	X	X	X
Sample	All	All	All
N	158,103	108,387	93,609
Eff. proteins	17,567	12,043	10,401
$\tilde{R}^2$	0.390	0.314	0.298
Raw N	4,774,455	4,774,455	4,774,455
N proteins	530,495	530,495	530,495
N structures	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.012	0.008	0.004
Effect (full range)	+17.9%	+16.9%	+17.8%
Effect (1 SD)	+3.2%	+3.0%	+3.2%
Effect (Q4 vs Q1)	+6.7%	+6.3%	+6.6%
Implied labs/yr	+1,176	+731	+397

Notes. prpfm PPML version of tbl-lab-counts. Same outcomes and sample; protein and year FE, and seven pre-AF features interacted with Post. Magnitude rows report $(\exp(\hat{\beta} \cdot x) - 1) \times 100$. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5 Robustness Checks

A.5.1 Parallel-trends, placebo

Table 25: Long-difference dose-response

	(1) # depositions	(2) First-deposition	(3) # follow-up
Headline TWFE			
AF cov × Post	$+0.342^{***}$	-0.001^{***}	$+0.494^{***}$
	(0.094)	(0.000)	(0.123)
CGS (2024) long-difference dose-response			
$\hat{\beta}_{\text{LD}}$	$+0.015^{***}$	-0.001	$+0.175^{***}$
	(0.003)	(0.001)	(0.052)
Implied % at AF cov $= 1$	$+87.9\%^{***}$	-37.4%	$+57.2\%^{***}$
$\bar{Y}_{\text{pre}}$	0.017	0.002	0.305

Notes. Panel A mirrors the baseline. Panel B's CGS long-difference row implements the Callaway–Goodman-Bacon–Sant'Anna continuous-treatment estimator: the balanced panel is collapsed to one observation per protein, the additive long difference $\Delta Y_i = \bar{Y}_{i,\text{post}} - \bar{Y}_{i,\text{pre}}$ (which differences out the protein effect) is formed, and ΔY_i is regressed by OLS on AF cov_i with the prpfm features in level form and primary-Pfam fixed effects; the AF cov slope is the average causal response. For the count outcomes (cols 1 and 3) the long difference is on the level (per-year deposition-count) scale, so its magnitude is not directly comparable to the multiplicative PPML headline — the relevant check is the sign and significance of the slope; the *Implied %* row rescales each slope by the pre-period mean $\bar{Y}_{\text{pre}}$ to express it as a proportional change at full coverage, comparable to the headline. For first deposition (col 2) the long difference is taken on the balanced at-risk panel (proteins with no pre-2017 PDB, all years) with the first-ever-deposition event indicator, so ΔY_i is the change in the per-year first-deposition rate; a negative slope matches the negative hazard effect in the headline. Survival filtering, which drops a protein after its first deposition, is what previously made this long difference ill-defined. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 26: Pre-trends: AF coverage gradient before AF2

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov $\times$ (Year $-$ 2017)	0.052	0.000	0.049
	(0.034)	(0.000)	(0.042)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features $\times$ Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	61,140	2,436,831	31,690
Eff. proteins	12,228	—	6,338
R^2 / $\tilde{R}^2$	0.417	0.026	0.472
Raw N	2,652,475	2,527,743	112,890
N proteins	530,495	507,917	22,578
N structures	44,588	7,413	34,484
$\bar{Y}_{t\leq 2021}$	0.016	0.003	0.291
$\hat{\beta}/\bar{Y}$	+330.8%	+10.8%	+17.0%
$\hat{\beta} \times \text{SD}/\bar{Y}$	+1.0%	+2.1%	+0.9%

Notes. Pre-trend specification restricting the sample to years $\leq$ 2021 (pre-AF2). The coefficient on AF cov $\times$ (Year $-$ 2017) captures whether the AF-cov gradient already trends during the pre period; small/insignificant coefficients support the parallel-trends assumption of the main DiD specification. Column structure mirrors *tbl-panel-baseline-2*. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 27: Covariate balance across AF coverage quartiles

Variable	Q4 × Trend	SE	SD	Norm. diff.
Length	-16.1736	0.3298	269.8488	-0.0599
# Pfam domains	-0.0151	0.0012	0.9812	-0.0154
Membrane	-0.0290	0.0005	0.3444	-0.0841
Disease assoc.	0.0005	0.0001	0.1023	0.0046
Pharma target	-0.0000	0.0000	0.0151	-0.0021
Secreted	-0.0043	0.0002	0.2202	-0.0193
Mitochondrial	0.0073	0.0002	0.1828	0.0401
# Pathways	0.4402	0.0063	4.2865	0.1027
DrugBank target	0.0036	0.0001	0.0937	0.0386
ChEMBL target	0.0038	0.0002	0.1274	0.0295
# PubMed (pre-2017)	-0.0952	0.0492	43.9909	-0.0022
Has PDB (pre-2017)	0.0084	0.0003	0.2019	0.0418

Note: Differential pre-period trends in each covariate by AF coverage quartile (Q4 vs. Q1). Normalised differences below 0.25 in absolute value indicate that the AF gradient is not capturing pre-trends in observable protein characteristics.

Table 28: Randomisation-inference placebo

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
$\hat{\beta}^{\text{true}}$	0.342^{***}	-0.001^{***}	0.494^{***}
	(0.094)	(0.000)	(0.123)
Placebo mean	-0.0004	0.0000	0.0004
Placebo SD	0.0671	0.0003	0.0847
Placebo 95% CI	[-0.1241, 0.1239]	[-0.0005, 0.0005]	[-0.1499, 0.1670]
RI p-value	0.000	0.000	0.000
# draws	200	200	200
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
Pre-AF features $\times$ Post	X	X	X
Sample	All	At risk	Intensive
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305

Notes. Randomization-inference placebo for *tbl-panel-prpfm-2*. Row $\hat{\beta}^{\text{true}}$ reports the true coefficient (clustered SE at the protein level in parentheses; stars from the clustered t-statistic). The AF coverage vector is permuted across all proteins $B = 200$ times; each draw re-runs the regressions (iid SE, lean fits) and records the placebo coefficient. The RI p-value is the share of draws with $|\hat{\beta}^{\text{placebo}}| \geq |\hat{\beta}^{\text{true}}|$. Protein-level features (log Pfam family size, log sequence length, membrane, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, log pre-2017 cryo-EM PDB count) are held fixed. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 29: 2016-event placebo on the cryo-EM–excluded sample

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov $\times$ Post_{2016}	-0.081	-0.001^{**}	-0.139
	(0.088)	(0.000)	(0.124)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features $\times$ Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	121,572	4,393,355	49,545
Eff. proteins	13,508	—	5,505
R^2 / $\tilde{R}^2$	0.366	0.022	0.415
Raw N	4,746,132	4,555,425	148,851
N proteins	527,348	510,809	16,539
N structures	46,244	10,680	29,437
$\bar{Y}_{t\leq 2021}$	0.010	0.002	0.213
$\hat{\beta}/\bar{Y}$	-802.9%	-36.0%	-65.1%
$\hat{\beta} \times \text{SD}/\bar{Y}$	-1.5%	-6.8%	-2.6%

Notes. Strictest event-2016 falsification: *(i)* prpfm controls rebuilt at end-2011 (as in *tbl-placebo-event2016-2011feat*), and *(ii)* outcomes built from non-EM nbnc depositions only — excluding cryo-EM removes the post-2014 method-revolution channel that disproportionately drove deposits on structurally hard / low-AF-cov targets and produced the negative baseline placebo coefficient. Outcomes are constructed from the deposition-level record (`pdb_deposition_coverage.parquet`) by filtering to `method_bucket != 'em'`, `is_batch_duplicate == 0`, and `is_sars_cov2 == 0`, then aggregating to the protein-year level. Sample splits, the at-risk truncation, and the regression FE structure are identical to *tbl-placebo-event2016*. A residual placebo effect here would survive both controls-leakage and cryo-EM as explanations. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 30: Rambachan–Roth parallel-trends sensitivity bounds

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
$\hat{\beta}_{\text{post,avg}}$	$+0.471$	-0.001	$+0.584$
Original 95% CI	$[+0.247, +0.696]$	$[-0.002, +0.000]$	$[+0.299, +0.868]$
Relative-magnitudes bounds (RR 2023)			
$\bar{M} = 0.5$	$[-0.085, +1.003]$	$[-0.004, +0.003]$	$[-0.137, +1.281]$
$\bar{M} = 1.0$	$[-0.530, +1.440]$	$[-0.008, +0.006]$	$[-0.706, +1.845]$
$\bar{M} = 2.0$	$[-1.458, +2.294]$	$[-0.011, +0.011]$	$[-1.903, +2.902]$

Notes. Rambachan and Roth (2023) parallel-trends sensitivity bounds, prpfm spec. The headline event-study has $\sum_{s \neq 2021} \beta_s \cdot \mathbf{1}\{t = s\} \cdot \text{AF cov}_i + \mu_i + \tau_t + \mathbf{X}_i \times \text{Post}_t \cdot \gamma + \varepsilon_{i,t}$ on the right-hand side (reference year 2021), fit by PPML (fepois) for the count outcomes in cols (1) and (3) and by OLS (linear probability model) for the binary first-deposition indicator in col (2). Reported $\hat{\beta}_{\text{post,avg}}$ is the simple average of the four post-period coefficients for $s \in \{2022, 2023, 2024, 2025\}$ on the log-rate scale in cols (1) and (3) and on the probability scale in col (2). The original 95% confidence interval treats the four pre-period coefficients (2017–2020) as a sharp test of parallel trends. The relative-magnitudes bounds (HonestDiD package) report the worst-case 95% CI for the post-period average under the restriction that any post-period violation is at most $\bar{M}$ times the worst pre-period violation; bounds inherit the underlying coefficient scale. Controls $\mathbf{X}_i$: seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, log pre-2017 cryo-EM PDB count) interacted with Post (the last two are dropped in col (2) because both are mechanically zero on the at-risk sample). Standard errors clustered at the protein level.

A.5.2 Alternative Specifications

Table 31: OLS estimation

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.021***	−0.001***	0.288***
	(0.003)	(0.000)	(0.060)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Estimator	OLS	OLS	OLS
Sample	All	At risk	Intensive
N	4,774,455	4,365,071	203,202
R^2	0.704	0.021	0.729
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305
% effect (SD)	+23.2%	-8.4%	+18.8%

Notes. OLS companion to *tbl-panel-prpfm-2*. All three columns are linear regressions (`feols`) on count outcomes, with the same prpfm specification as the headline (entry + year FE in cols (1) and (3); primary-Pfam family FE + year FE in col (2); seven pre-AF features interacted with Post throughout, plus features in level form on col (2)). The % effect (SD) row reports $100\,\hat{\beta} \times \mathrm{SD}(\mathrm{AF\,cov})/\bar{Y}_{t\leq 2021}$, i.e. the implied effect of a one-SD increase in AF coverage as a percentage of the pre-period mean (linear scale, since estimator is OLS). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 32: Adding last-author lab fixed effects

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.342***	0.003***	0.494***
	(0.095)	(0.000)	(0.126)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Lab FE	X	X	X
Protein features		X	
Sample	All	At risk	Intensive
N	171,135	4,364,820	72,261
Eff. proteins	19,015	—	8,029
R^2 / $\tilde{R}^2$	0.460	0.225	0.532
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305
Effect (full range)	+40.8%	+0.3 p.p.	+63.8%
Effect (1 SD)	+6.7%	+0.1 p.p.	+9.8%
Effect (Q4 vs Q1)	+14.3%	+0.1 p.p.	+21.3%

Notes. Same 3-column prpfm specification as *tbl-panel-prpfm-2* but adding last-author lab fixed effects. `primary_lab` is the `last_author_id` of each protein's first PDB deposition released in 2017–2025; proteins without any deposition in window are assigned a synthetic `_NO_LAB_` category. In columns (1) and (3) protein FE absorb `primary_lab` (it is protein-constant), so the lab FE row is a transparency check rather than identifying variation. In column (2) (at-risk, no protein FE) `primary_lab` adds within-family-lab-year identifying variation. Estimators: PPML (`fepois`) for count outcomes in cols (1) and (3); OLS / LPM (`feols`) on the binary first-deposition indicator in col (2). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 33: Adding organism fixed effects

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.342***	−0.001**	0.494***
	(0.094)	(0.000)	(0.123)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Organism FE	X	X	X
Protein features		X	
Sample	All	At risk	Intensive
N	171,135	4,365,064	72,261
Eff. proteins	19,015	—	8,029
R^2 / $\tilde{R}^2$	0.460	0.030	0.532
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305
Effect (full range)	+40.8%	-0.1 p.p.	+63.8%
Effect (1 SD)	+6.7%	-0.0 p.p.	+9.8%
Effect (Q4 vs Q1)	+14.3%	-0.0 p.p.	+21.3%

Notes. Same 3-column prpfm specification as *tbl-panel-prpfm-2* but adding organism fixed effects (`organism_id`). Organism is protein-constant and time-invariant, so in columns (1) and (3) the protein FE absorb `organism_id` mechanically and coefficients there are identical to *tbl-panel-prpfm-2*. The new content is column (2) (at-risk, no protein FE): the AF cov × Post coefficient is identified from within-organism, within-Pfam variation. A col (2) coefficient that shrinks toward zero relative to *tbl-panel-prpfm-2* col (2) suggests cross-organism composition drives the at-risk effect; one that survives suggests within-organism heterogeneity does the work. Estimators: PPML (`fepois`) for count outcomes in cols (1) and (3); OLS / LPM (`feols`) on the binary first-deposition indicator in col (2). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 34: Cluster

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF $\text{cov}_c \times$ Post	0.333^{***}	-0.003^{***}	0.503^{***}
	(0.112)	(0.001)	(0.137)
Cluster FE	X		X
Year FE	X	X	X
Pre-AF features $\times$ Post	X	X	X
Pre-AF features (level)		X	
Cluster size		X	
N	116,631	798,914	59,067
R^2 / $\tilde{R}^2$	0.555	0.013	0.601
Raw N	963,729	798,914	140,598
N clusters	107,081	91,459	15,622
$\bar{Y}_{t\leq 2021}$	0.083	0.008	0.459
% effect (per unit)	+39.6%	-0.3 p.p.	+65.3%
% effect (SD)	+8.7%	-0.1 p.p.	+10.8%

Notes. Estimates on the cluster-year panel on: (1) deposition counts on the balanced panel (PPML), (2) first-deposition indicator on the at-risk sample (OLS LPM; clusters with no member PDB before 2017), and (3) follow-up deposition counts on the intensive sample (clusters with at least one member PDB before 2017). AF cov is the share of proteins in a cluster predicted at pLDDT ≥ 70 and Post is an indicator for years after 2021. All columns include year fixed effects and the cluster-level analogues of the protein-level prpfm controls interacted with Post: mean log Pfam family size, mean log sequence length, share membrane, share disease-associated, log cluster-sum pre-2017 PubMed citations, log cluster-sum pre-2017 PDB depositions, and log cluster-sum pre-2017 cryo-EM depositions. Columns (1) and (3) add cluster fixed effects; column (2) adds the prpfm features in level form and cluster size. Standard errors clustered at the cluster level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5.3 Alternative AF2 treatment definitions and post-period

Table 35: Mean pLDDT as the treatment variable

	(1) # depositions	(2) First-deposition	(3) # follow-up
mean pLDDT × Post	0.538***	−1.055***	0.956***
	(0.177)	(0.208)	(0.227)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
N	171,135	3,385,281	72,261
$\tilde{R}^2$	0.460	0.201	0.532
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305

Notes. Same 3-column specification as *tbl-panel-baseline-2* but replacing AF coverage with mean pLDDT rescaled to $[0, 1]$. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 36: Quartile of AF coverage

	(1) # depositions	(2) First-deposition	(3) # follow-up
AF $\text{cov}_{Q2} \times$ Post	0.137^{***}	-0.046	0.178^{***}
	(0.034)	(0.051)	(0.044)
AF $\text{cov}_{Q3} \times$ Post	0.158^{***}	-0.121^{*}	0.192^{***}
	(0.045)	(0.062)	(0.060)
AF $\text{cov}_{Q4} \times$ Post	0.169^{***}	-0.329^{***}	0.259^{***}
	(0.058)	(0.069)	(0.069)
Year FE	X	X	X
Protein FE	X		X
Family FE		X	
Pre-AF features		X	
Pre-AF features $\times$ Post	X	X	X
N	171,135	3,385,281	72,261
$\tilde{R}^2$	0.460	0.200	0.532
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.305

Notes. Same prpfm specification as *tbl-panel-prpfm-2* but the continuous AF cov $\times$ Post treatment is replaced by three dummies for the second, third, and fourth quartiles of AF cov_{70}, each interacted with Post; the bottom quartile (Q1) is the omitted reference. Quartiles are computed once on the cross-section of proteins, so each protein carries a fixed quartile across all years. On the at-risk column (no protein FE), the three quartile dummies also enter in level form. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 37: AF coverage at alternative pLDDT thresholds (50, 70, 90)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	# depositions			First-deposition			# follow-up		
Threshold (τ)	50	70	90	50	70	90	50	70	90
AF cov × Post	0.502***	0.342***	0.158**	−0.336**	−0.374***	−0.472***	0.712***	0.494***	0.311***
	(0.121)	(0.094)	(0.064)	(0.157)	(0.113)	(0.082)	(0.156)	(0.123)	(0.082)
Family FE				X	X	X			
Protein FE	X	X	X				X	X	X
Year FE	X	X	X	X	X	X	X	X	X
Family × Post FE	X	X	X	X	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X	X	X	X	X
Pre-AF features (level)				X	X	X			
Protein features				X	X	X			
N	171,135	171,135	171,135	3,385,281	3,385,281	3,385,281	72,261	72,261	72,261
$\bar{Y}_{t\leq 2021}$	0.017	0.017	0.017	0.002	0.002	0.002	0.305	0.305	0.305

Notes. Same 3-column prpfm specification as *tbl-panel-prpfm-2* but applied at three pLDDT thresholds $\tau \in \{50, 70, 90\}$ (share of residues with pLDDT $\geq \tau$). Columns are grouped by outcome; within each group the three columns report the coefficient at the three thresholds. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 38: Defining Post as 2022–2025 (donut)

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post_{2022}	0.396^{***}	-0.364^{***}	0.575^{***}
	(0.088)	(0.118)	(0.116)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	171,135	3,385,281	72,261
Eff. proteins	19,015	—	8,029
$\tilde{R}^2$	0.460	0.200	0.533
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.309
Effect (full range)	+48.6%	-30.5%	+77.7%
Effect (1 SD)	+7.8%	-6.7%	+11.5%
Effect (Q4 vs Q1)	+16.7%	-13.3%	+25.2%

Notes. Same 3-column specification as *tbl-panel-baseline-2* but defining the post period as 2023–2025 instead of 2022–2025, so 2021 is coded as pre-treatment. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

A.5.4 Alternative samples

Table 39: Full deposition record (no batch / SARS-CoV-2 filter)

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.552***	−0.381***	0.724***
	(0.126)	(0.111)	(0.162)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	171,252	3,391,882	72,342
Eff. proteins	19,028	—	8,038
$\tilde{R}^2$	0.567	0.202	0.621
Raw N	4,774,455	4,527,865	203,202
N proteins	530,495	507,917	22,578
N structures	202,400	26,455	152,894
$\bar{Y}_{t\leq 2021}$	0.034	0.002	0.637
Effect (full range)	+73.6%	-31.6%	+106.2%
Effect (1 SD)	+11.0%	-7.0%	+14.7%
Effect (Q4 vs Q1)	+24.1%	-13.8%	+32.7%

Notes. Same 3-column specification as *tbl-panel-baseline-2* but constructing outcomes from the full PDB deposition record (no batch-deposition or SARS-CoV-2 filter). Sample splits (full / at-risk / intensive) and the pre-2017 PDB indicator also use the unfiltered counts. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 40: Trimming proteins with AF coverage < 0.2

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.418***	−0.339***	0.550***
	(0.097)	(0.123)	(0.129)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	168,804	3,365,539	71,361
Eff. proteins	18,756	—	7,929
$\tilde{R}^2$	0.460	0.200	0.532
Raw N	4,695,660	4,452,184	200,718
N proteins	521,740	499,438	22,302
N structures	92,134	14,184	67,447
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.307
Effect (full range)	+52.0%	-28.7%	+73.4%
Effect (1 SD)	+8.3%	-6.2%	+11.0%
Effect (Q4 vs Q1)	+17.8%	-12.4%	+24.0%

Notes. Same 3-column specification as *tbl-panel-baseline-2* but restricted to proteins with AF $\text{cov}_{70} \geq 0.2$, dropping proteins with negligible AlphaFold predictions. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 41: Excluding membrane proteins

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.356***	−0.386***	0.521***
	(0.102)	(0.124)	(0.132)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	138,735	2,928,156	61,470
Eff. proteins	15,415	—	6,830
$\tilde{R}^2$	0.490	0.197	0.554
Raw N	4,117,932	3,902,192	181,458
N proteins	457,548	437,386	20,162
N structures	80,359	11,082	61,438
$\bar{Y}_{t\leq 2021}$	0.017	0.002	0.309
Effect (full range)	+42.8%	-32.0%	+68.4%
Effect (1 SD)	+7.0%	-7.1%	+10.4%
Effect (Q4 vs Q1)	+14.9%	-14.0%	+22.6%

Notes. Same 3-column specification as *tbl-panel-baseline-2* restricted to non-membrane proteins (`is_membrane = 0`). Motivation: *tbl-balance-levels* shows the largest Q4-vs-Q1 AF-cov imbalance for membrane proteins (normalised difference ≈ -0.29); contemporaneous cryo-EM and GPCR/ion-channel advances over 2021–2022 could contaminate the post period for membrane proteins specifically, so dropping them isolates the AF effect. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 42: Human

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.410***	0.219	0.446***
	(0.138)	(0.221)	(0.165)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	56,817	68,148	30,798
Eff. proteins	6,313	—	3,422
$\tilde{R}^2$	0.447	0.204	0.485
Raw N	176,643	113,366	51,705
N proteins	19,627	13,882	5,745
N structures	39,112	4,058	31,205
$\bar{Y}_{t\leq 2021}$	0.178	0.024	0.502
Effect (full range)	+50.8%	+24.5%	+56.2%
Effect (1 SD)	+8.1%	+4.2%	+8.8%
Effect (Q4 vs Q1)	+17.4%	+8.9%	+19.0%

Notes. Same 3-column specification as *tbl-panel-baseline-2* restricted to human proteins (`organism_id = 9606`). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

Table 43: Single protein depositions

	(1) # depositions	(2) First-deposition indicator	(3) # follow-up depositions
AF cov × Post	0.246**	−0.086	0.263*
	(0.122)	(0.201)	(0.151)
Family FE		X	
Protein FE	X		X
Year FE	X	X	X
Pre-AF features × Post	X	X	X
Pre-AF features (level)		X	
Protein features		X	
Sample	All	At risk	Intensive
N	75,393	2,398,693	35,865
$\tilde{R}^2$	0.288	0.173	0.350
Raw N	4,774,455	4,527,884	203,202
N proteins	530,495	507,917	22,578
N structures	92,930	14,370	68,001
$\bar{Y}_{t\leq 2021}$	0.005	0.001	0.083
$\hat{\beta}/\bar{Y}$	+5011.3%	-8585.4%	+316.9%
$\hat{\beta} \times \text{SD}/\bar{Y}$	+4.8%	-1.6%	+5.4%

Notes. Same 3-column specification as *tbl-panel-baseline-2* but restricting the PDB counts used to build outcomes to single-protein depositions (`n_uniprot_ids == 1`), with the nbnc filter applied. Sample composition (which proteins are at risk vs intensive) is preserved. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. The prpfm specification adds seven pre-AF features (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, and log pre-2017 cryo-EM PDB count) interacted with Post. Cols (1) and (3) include protein and year FE; col (2) (no protein FE because each protein has at most one first-ever event) uses primary-Pfam family FE plus the features in level form (log pre-2017 PDB count and log pre-2017 cryo-EM PDB count dropped because both are mechanically zero on the at-risk sample).

A.5.5 Alternative distance measures

Table 44: Distances to nearest solved proteins

Distance	(1) pident	(2) pident, cov.adj.	(3) bitscore	(4) NJ-flat	(5) cw NJ-flat
Post × AF cov	0.491***	0.479***	0.588***	0.435***	0.544***
	(0.122)	(0.107)	(0.125)	(0.119)	(0.114)
Post × Dist.	0.615***	0.727***	0.844***	0.278***	0.516***
	(0.150)	(0.141)	(0.168)	(0.088)	(0.099)
Post × AF cov × Dist.	-0.450**	-0.752***	-0.875***	-0.187*	-0.595***
	(0.192)	(0.182)	(0.217)	(0.110)	(0.126)
Year FE	X	X	X	X	X
Protein FE	X	X	X	X	X
Pre-AF features × Post	X	X	X	X	X
N	171,135	171,135	171,135	171,135	171,135
$\tilde{R}^2$	0.460	0.460	0.460	0.460	0.460
Raw N	4,774,455	4,774,455	4,774,455	4,774,455	4,774,455
N proteins	530,495	530,495	530,495	530,495	530,495
N structures	92,930	92,930	92,930	92,930	92,930
$\bar{Y}_{t\leq 2021}$	0.017	0.017	0.017	0.017	0.017

Notes. PPML estimates of n_{dep} on the full no-batch, no-SARS-CoV-2 protein × year panel. Each column uses a different distance variable Z (column headers; see below for the definitions) and reports AF cov × Post (level), Z × Post, and the heterogeneity triple AF cov × Post × Z. All columns include protein FE, year FE, and seven pre-AF features interacted with Post (log Pfam family size, log sequence length, membrane indicator, disease-association indicator, log pre-2017 publication count, log pre-2017 PDB count, log pre-2017 cryo-EM PDB count). Distance definitions: (1) pident = MMseqs pident (1 - max pident/100); (2) pident, cov.adj. = MMseqs pident, coverage-adjusted; (3) bitscore = MMseqs bitscore; (4) NJ-flat = NJ-flat tree distance (headline); (5) cw NJ-flat = coverage-weighted NJ-flat tree distance. Z columns are NA-filled to 0 (max-imputation lives upstream in the distance-measure construction). Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.