How AI Impacts the Quest for Knowledge: Evidence from AlphaFold2

Myra Mohnen and Joshua S. Gans*

26 March 2026

Abstract

We study how advances in artificial intelligence reshape scientific knowledge production. We first develop a model of knowledge acquisition on a metric space in which experiments generate local spillovers, laboratories face setup costs, and AlphaFold2 (AF2) arrives as an exogenous signal that lowers baseline information costs unevenly across proteins. The model predicts reallocation rather than simple expansion: substitution away from experiments AF2 already informs, alongside increased search on frontier proteins and shifts in follow-up work. We test these predictions using a new dataset linking AF2 outputs to roughly 553,000 proteins and Protein Data Bank activity. Exploiting cross-protein variation in AF2 prediction quality after the July 2021 release, we find that AF2 substantially reduces the probability that a well-predicted protein receives any experimental attention, yet total depositions and follow-up work increase sharply—effort narrows onto fewer proteins rather than disappearing. The composition of new depositions shifts toward first-ever structure determinations for previously unresolved proteins, and within follow-up work, toward resolving previously uncharacterised regions and away from pure corroboration. The net effect is a reallocation of effort toward proteins where AI predictions and prior experimental knowledge are jointly available, driven primarily by academic rather than industry laboratories. AI thus reorganises where, how, and by whom science is produced.

*University of Ottawa (Mohnen) and University of Toronto and NBER (Gans). Thanks to seminar participants at Toronto for helpful comments, the SSHRC for funding, and to ChatGPT 5.4 Pro and Claude Opus 4.6 for research assistance. Responsibility for all errors remains our own.

1 Introduction

How does a major advance in artificial intelligence reshape the production of scientific knowledge? Economic theory offers competing predictions. AI may substitute for human effort, reducing research activity; it may lower costs and thereby expand the feasible set of research problems; or it may redirect effort across the knowledge landscape without expanding total volume. Adjudicating among these possibilities requires a setting where AI delivers a large, measurable, and heterogeneous shock to scientific production—and where the research choices of individual scientists can be observed in detail.

This paper studies such a setting: the release of AlphaFold2 (AF2), a deep-learning system that predicts protein three-dimensional structures with near-experimental accuracy (Jumper et al., 2021). In July 2021, DeepMind publicly released predicted structures for nearly all known proteins, providing the structural biology community with a universal but heterogeneous information endowment. The release was universal because AF2 generated predictions for essentially every protein; it was heterogeneous because predictions varied substantially in confidence across proteins, depending primarily by evolutionary conservation rather than prior experimental effort. This combination of breadth and cross-protein variation in informativeness provides a quasi-experimental setting for studying how a cost-reducing AI tool reshapes the direction, composition, and organization of scientific research.

We construct a novel dataset linking approximately 553,000 proteins to their experimental structures in the Protein Data Bank (PDB), the public archive where research teams deposit solved structures. For each protein, we measure the quality of AF2's predictions and the protein's position on the structural knowledge frontier—a sequence-similarity tree that organises proteins by evolutionary relatedness. Our empirical strategy exploits cross-protein variation in AF2 coverage, defined as the share of each protein's residues predicted above a usable confidence threshold, interacted with a post-release indicator. Because AF2 coverage reflects intrinsic biological properties, prior experimental effort and template availability at the time of prediction, protein fixed effects absorb the level of treatment. Identification therefore comes from differential changes in experimental activity after 2021 across proteins receiving more versus less informative AF2 predictions. This approach relies on the assumption that, absent the release of AF2, proteins with different levels of AF2 coverage would have followed parallel trends in experimental activity. While AF2 coverage is predetermined, it may be correlated with underlying biological characteristics that also shape research trajectories; we therefore interpret our estimates as capturing differential responses conditional on these fixed characteristics, rather than as

arising from random assignment of treatment intensity.

We develop a theoretical framework, building on Carnehl and Schneider (2025), to discipline the empirical analysis. The model embeds proteins in a sequence-similarity tree where experiments generate informational spillovers for nearby proteins, and laboratories face protein-specific setup costs that create persistent incumbency advantages. AF2 enters as an independent signal on the dominant structural component of each protein, attenuating the marginal value of experiments that produce baseline coordinate information while leaving the value of conformational-state and ligand-binding experiments unchanged. The framework generates predictions about the direction of reallocation across extensive and intensive margins, the role of frontier position, the composition of follow-up work, and heterogeneity across methods and institutions.

Our main findings are as follows. First, AF2 partially substitutes for experimental work: the probability that any given protein receives experimental attention declines substantially for well-predicted proteins. Yet effort concentrates rather than disappears — among proteins that remain experimentally active, output and follow-up work increase sharply, consistent with AF2 predictions serving as structural templates that make deeper experimental programs worthwhile. The composition of new experiments shifts toward first-ever structure determinations for proteins that had never been experimentally resolved. Second, within follow-up work on already-solved proteins, AF2 shifts the composition of experiments toward resolving previously uncharacterised regions and away from corroborating work. Third, AF2's effect is strongest for proteins located near existing experimental knowledge, consistent with its predictions being more complementary to existing experimental infrastructure in these regions. Fourth, the response is driven primarily by academic laboratories rather than industry teams, consistent with institutions holding broader research objectives being more responsive to a general-purpose informational shock. Fifth, we document patterns consistent with dynamic spillovers, whereby newly solved proteins are associated with increased follow-up activity (conformational and ligand-binding experiments) on nearby proteins thus propagating experimental opportunities through the knowledge landscape. Overall, AF2 reduces the breadth of experimental attention but deepens work on fewer proteins, fundamentally reorganising where, how, and by whom experimental effort is directed.

This paper contributes to several literatures. Most directly, it provides the first integrated theoretical and empirical analysis of how AI reshapes knowledge production in a specific scientific field. A growing literature studies the economics of AI and scientific discovery (Agrawal et al., 2026). Gans (2025b) develops a model in which AI provides informational priors that redirect the search for knowledge, and Gans (2025a) extends this

framework to general equilibrium, showing that AI's impact on long-run growth depends on whether it promotes incremental or exploratory research. Our theoretical framework adapts the knowledge-acquisition model of Carnehl and Schneider (2025) to the specific institutional features of structural biology—tree-structured spillovers, multidimensional experiments, and laboratory capital—and our empirical results provide a concrete test of the qualitative predictions that emerge from these models.

Within the empirical literature on AlphaFold specifically, contemporaneously to our own work Hill and Stein (2026) study how AF2 affects the direction of experimental effort by estimating changes in PDB depositions as a function of AF2's "homology score." Our paper complements theirs through several extensions: a theoretical model that generates testable predictions about the margins of adjustment, a richer characterization of the knowledge frontier using tree-based distance and density measures, a decomposition of intensive-margin depositions into economically distinct action types, and analysis of dynamic spillovers through which extensive entry propagates experimental opportunities to neighbouring proteins. Cavalli (2024) examines how AF2 affects the organization of academic laboratories, finding specialist–generalist responses that align with our finding that the response to AF2 is mediated by laboratory-specific capital. Kim (2025) studies how data-extrapolation technologies affect knowledge production more broadly, providing complementary evidence on the mechanisms through which AI tools redirect scientific search. Tranchero (2024) examines how data-driven methods reshape pharmaceutical innovation, documenting a parallel reallocation toward novel targets enabled by computational tools.

More broadly, the paper connects to the literature on AI as a general-purpose technology and its implications for economic growth. Aghion et al. (2019) and Jones (2025) examine AI's potential to accelerate the production of ideas, with implications that depend critically on whether AI substitutes for or complements human research effort. Acemoglu (2024) develops a macroeconomic framework that emphasizes the limited scope of current AI applications, while Acemoglu and Restrepo (2018, 2019) analyse the task-level implications of automation for labor markets. Our evidence on AlphaFold speaks to these frameworks by documenting a concrete case in which AI neither uniformly expands nor contracts research effort, but instead reallocates it—reallocating activity across the knowledge landscape toward proteins where AI predictions complement existing experimental infrastructure. The finding that AI redirects rather than expands total activity resonates with Bloom et al. (2020), who document that research productivity has been declining; AI may address this not by generating more ideas per researcher, but by improving the allocation of effort across problems.

Finally, our emphasis on the knowledge frontier connects to models of endogenous

growth and recombinant innovation. Romer (1990) and Weitzman (1998) model growth as driven by the recombination of existing ideas, a process that depends on how densely the knowledge space is populated. Agrawal et al. (2019) formalise this in the AI context, arguing that prediction machines help identify valuable combinations from large search spaces. Mullainathan and Rambachan (2025) provides a broader perspective on how algorithmic tools are reshaping the scientific process. The pattern we document — experimental attention narrowing to fewer proteins while new structures raise the value of conformational and ligand-binding work on neighbouring proteins—illustrates a concrete mechanism through which AI may sustain recombinant growth by expanding the set of solved problems from which new combinations can be drawn.

The remainder of the paper is organised as follows. Section 2 describes the protein folding problem and AF2. Section 3 develops the theoretical model. Section 4 describes the data. Section 5 presents the empirical strategy. Section 6 reports results. Section 7 concludes.

2 The Protein Folding Problem and AlphaFold2

The protein folding problem has been a central question in molecular biology since the 1960s. A protein is a linear chain of amino acids—each called a *residue*—and a typical protein contains several hundred. The folding problem is to predict the three-dimensional structure that a given sequence of residues will adopt. Its importance was firmly established by the experiments of Christian Anfinsen, who demonstrated that a protein's native structure is fully encoded in its amino acid sequence. This insight, recognised with the 1972 Nobel Prize in Chemistry, implied that structure prediction from sequence alone was possible in principle. Because a protein's structure governs its biochemical function, progress on the folding problem has broad implications for drug discovery, disease biology, synthetic biology, and our basic understanding of cellular processes.

Despite its conceptual clarity, the problem resisted solution for decades. The core obstacle is combinatorial: a chain of just 100 amino acids can, in theory, adopt more conformations than there are atoms in the universe (10^{100}). This makes it computationally impossible to predict structure by brute force. Moreover, proteins do not always fold in a rigid, predictable manner. Many contain flexible or intrinsically disordered regions, adopt multiple biologically relevant conformations, or change structure depending on their molecular environment.

A major breakthrough came in 2020 when DeepMind's AF2 achieved near-experimental accuracy in blind structure prediction at the 14th Critical Assessment of protein Structure

Prediction, a biennial community benchmark. DeepMind—working with EMBL-EBI—subsequently released predicted structures for approximately 350,000 proteins in July 2021, including nearly the entire human proteome, and expanded this to over 200 million proteins in July 2022. The AF2 code was also made publicly available, enabling any research group to generate predictions for proteins of interest. In recognition of this advance, the 2024 Nobel Prize in Chemistry was awarded to Demis Hassabis and John Jumper, the lead developers of AF2.

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. 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 experimental structures in the PDB released before April 30, 2018. High-quality predictions can be generated from sequence and MSA alone—AF2 internalizes evolutionary signals and physical regularities rather than relying on solved templates or explicit folding simulations.

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 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 homologs—also tend to receive lower confidence scores. 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 of structural biology. Traditionally, determining a protein's three-dimensional structure required a lengthy process of experimental trial and error using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (Kuhlman and Bradley, 2019). Researchers typically cycled through multiple stages—hypothesizing about protein constructs, expressing the protein, crystallizing or imaging it, solving the structure, and iteratively refining the result—often repeating the process several times before obtaining a reliable model. Each stage involved substantial uncertainty and resource costs: even under

the Protein Structure Initiative (PSI)—a large-scale structural genomics program that ran between 2000 and 2017—fewer than 3% of selected targets survived the full experimental pipeline from selection to PDB deposition (Figure 6). Determining a single protein structure could require months to years of work and cost on the order of $100,000–$300,000 per structure, with even optimized high-throughput PSI efforts reducing costs only to approximately $66,000–$138,000 (Chandonia and Brenner, 2006; Terwilliger et al., 2009).

AF2 shifted this process from de novo discovery toward targeted validation. High-confidence predictions now serve as default starting points, guiding construct design, accelerating hypothesis generation, and allowing experimental effort to focus on unresolved regions, alternative conformations, or protein complexes (Campbell, 2024). For example, predicted structures help researchers identify domain boundaries before initiating experiments, enabling them to express only well-folded protein fragments rather than attempting to express full-length proteins that may prove unstable. AF2 models also serve as starting templates for molecular replacement in X-ray crystallography and as structural guides for fitting models into cryo-electron microscopy density maps, substantially reducing the time required to solve structures experimentally.

At the same time, AF2 did not eliminate the need for experiments. Instead, it enabled hybrid workflows in which computational predictions and experimental methods complement one another. Many laboratories now iterate between AF2 predictions and experimental data, alternating prediction, refinement, and validation. AF2 models are now routinely used as molecular-replacement templates in crystallography, allowing structures that previously resisted experimental phasing to be solved (Barbarin-Bocahu and Graille, 2022). Large-scale benchmarks show that AF2 predictions are sufficiently accurate to support molecular replacement in the vast majority of cases, after which experimental refinement yields high-quality structures (Terwilliger et al., 2023, 2024).

Beyond accelerating existing experiments, AF2 has expanded the feasible scope of structural biology. For some proteins—particularly those that are experimentally intractable—predicted structures can provide sufficient information to guide downstream functional studies even in the absence of a solved experimental structure.[1] By narrowing the experimental search space, lowering the cost of structural determination, and enabling work on previously inaccessible proteins, AF2 has sharply reduced the marginal cost of obtaining reliable structural information.

Taken together, AF2 dramatically lowered the cost of accessing structural knowledge,

[1]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).

while introducing systematic heterogeneity in where that knowledge is most reliable. These features—large productivity gains combined with uneven confidence across proteins and regions, and continued complementarity with experimental validation—are central to understanding how AF2 reshaped scientific production. They motivate our empirical analysis of how the availability of high-confidence predictions altered research directions, experimental effort, and the organization of scientific work.

3 A Model of Experimental Choice

This section develops a stylised model of experimental choice in structural biology. The purpose of the model is qualitative rather than fully structural: it is meant to discipline signs and interactions in the empirical analysis, not to provide a literal description of protein evolution or a tight mapping from primitives to regression coefficients. The framework builds on Carnehl and Schneider (2025), but it adapts that logic to several features of the protein-structure setting that matter here: knowledge spillovers live on an estimated sequence-similarity tree, experiments improve several distinct dimensions of structural knowledge, laboratories face protein-specific setup costs that create persistent local incumbency, and frontier difficulty evolves as the public stock of solved proteins expands.

The core insight of the Carnehl and Schneider (2025) (henceforth CS) framework is that when scientific problems are related to one another—so that solving one problem generates informational spillovers for nearby problems—the choice of *where* to acquire information becomes as important as the choice of *how much* to acquire. CS model an agent who sequentially acquires costly signals on a metric space, where the value of each signal depends on its location because observations propagate information to neighbouring points. This framework captures the essential economics of structural biology: resolving the structure of one protein generates a homology-modelling template for evolutionarily related proteins, so the value of an experiment depends on where the target protein sits relative to existing knowledge.

However, the CS framework requires several substantive modifications to map into the protein-folding scientific workflow. The protein knowledge space is not a generic metric space but a tree-like sequence-similarity geometry with heterogeneous density. Experiments are not homogeneous signals but differ qualitatively in what they reveal. Costs are shaped by laboratory-specific capital that creates an asymmetry between studying new proteins and conducting follow-up work on familiar ones. The value of structural knowledge flows through both scientific and commercial channels, each with distinct

incentive structures. And AlphaFold2 does not simply reduce the cost of information acquisition but differentially affects the value and cost of different types of experiments through method-specific channels.

The model is organised as follows. Section 3.1 defines the knowledge space and clarifies how the Brownian structure should be interpreted. Section 3.2 introduces experiments and quality dimensions. Section 3.3 incorporates laboratory capital and maps the theory to the empirical margins used later in the paper. Section 3.4 defines the value of structural knowledge. Section 3.5 introduces AF2 as a universal but heterogeneous information shock. Section 3.7 states the empirical predictions. For ease of reference, Table 1 summarises the main notation used throughout the model and empirical sections.

3.1 The Knowledge Space

Let $\mathcal{P}$ denote the set of proteins. Proteins are embedded in a tree $(\mathcal{T}, D_{\mathcal{T}})$ that should be interpreted empirically as an estimated sequence-similarity scaffold rather than as a literal gene or species phylogeny, and a σ-finite measure μ on $\mathcal{T}$ records the density of proteins. Dense regions correspond to large protein families; sparse regions correspond to lineage-specific or orphan proteins. As in Carnehl and Schneider (2025), the location of an observation matters because information spills over locally on the tree.

We model the *dominant structural component* of protein x by a scalar Gaussian process

$$Y(x) \sim \mathrm{BM}(\mathcal{T}, D_{\mathcal{T}}), \tag{1}$$

with a covariance kernel

$$k(x_i, x_j) = \tfrac{1}{2}\Big[D_{\mathcal{T}}(x_i, x_0) + D_{\mathcal{T}}(x_j, x_0) - D_{\mathcal{T}}(x_i, x_j)\Big]$$

for a fixed root x_0.

The Brownian specification is a tractable approximation, not a literal model of protein structure. Real proteins are high-dimensional, modular across domains, and often context-dependent: disorder, ligand-induced rearrangements, allostery, post-translational modification, and convergent motifs all violate literal tree-Brownian evolution. We, therefore, do *not* interpret $Y(x)$ as the full atomic structure of protein x. Instead, $Y(x)$ is a reduced-form latent index of how informative nearby solved proteins are for predicting a protein's dominant fold or template-relevant structural scaffold. Brownian motion is adopted because it imposes three qualitative features that are central for the empirical analysis: nearby homologues are more informative than distant ones, tree topology matters

for spillovers, and the value of a new experiment depends on where it sits relative to the stock of solved proteins. In the data, the underlying tree is itself an estimated similarity scaffold constructed from sequence comparisons (Appendix A.3.3), so the Brownian object should be read as a reduced-form homology kernel on that scaffold. The empirical work below uses only these qualitative implications; it does not require the Brownian model to be literally correct or structurally estimated.[2]

An experimental observation at protein x_i yields

$$s_i = Y(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \nu_i), \tag{2}$$

where the observation noise ν_i is determined by the quality of the experiment. Let $\mathcal{F}_t$ denote the public stock of experimental observations that are informative about the dominant static scaffold at time t. Standard Gaussian updating then defines the posterior variance

$$\sigma_0^2(x; \mathcal{F}_t) = \mathrm{Var}\big(Y(x) \mid \mathcal{F}_t\big).$$

The model uses only two qualitative properties of σ_0^2: it falls when nearby observations become more precise, and it rises as one moves away from solved relatives.

For local calculations, it is useful to work with the augmented tree $\bar{\mathcal{T}}$ obtained by inserting every observation location in $\mathcal{F}_t$ and every branch point as a vertex. Each open edge of $\bar{\mathcal{T}}$ is therefore branchless and observation-free. The next lemma gives the exact local variance formula on such an edge.[3]

Lemma 1 (Local posterior variance on a branchless tree edge)**.** *Let $\bar{\mathcal{T}}$ be the augmented tree just described. Suppose x lies in the interior of an edge $e = (v_l, v_r)$ of $\bar{\mathcal{T}}$, so that the open segment (v_l, v_r) contains no observation location and no branch point. Let $\mathcal{G}$ be the σ-field generated by all observations whose locations lie outside (v_l, v_r). Write $d_l = D_{\mathcal{T}}(x, v_l)$, $d_r = D_{\mathcal{T}}(x, v_r)$ and $X = d_l + d_r$, and define $P_l = \mathrm{Var}(Y(v_l) \mid \mathcal{G})$, $P_r = \mathrm{Var}(Y(v_r) \mid \mathcal{G})$, $C_{lr} = \mathrm{Cov}(Y(v_l), Y(v_r) \mid \mathcal{G})$. Then*

$$\mathbb{E}[Y(x) \mid \mathcal{G}] = \frac{d_r}{X}\mathbb{E}[Y(v_l) \mid \mathcal{G}] + \frac{d_l}{X}\mathbb{E}[Y(v_r) \mid \mathcal{G}],$$

[2] In Appendix A.2, we show that a stochastic structure that is closer to that in the science of protein exploration can be developed, leaving the qualitative predictions that follow from the model below essentially unchanged.

[3] It should be read as a *local decomposition* after all off-edge information has been summarised in endpoint posterior moments; it is not a claim that only the two nearest observations matter.

and

$$\mathrm{Var}(Y(x) \mid \mathcal{G}) = \frac{d_l d_r}{X} + \left(\frac{d_r}{X}\right)^2 P_l + \left(\frac{d_l}{X}\right)^2 P_r + 2\,\frac{d_l d_r}{X^2}\,C_{lr}. \tag{3}$$

The proofs of all results are in Appendix A.1. The branchless-edge restriction is essential. If side branches attach to the interior of the segment, then off-segment observations can affect $\mathrm{Var}(Y(x) \mid \mathcal{G})$ through those attachments and the simple two-endpoint formula need not hold.

To connect the theory more directly to the empirics, let $F_{xt} = F(x; \mathcal{F}_t)$ denote a latent *frontier index* that is increasing in distance to the nearest solved protein and local gap size and is typically decreasing in cluster saturation and neighbour density. The exact mapping from the vector of frontier measures used in the data to this scalar is not important for the theory. In the baseline regressions, we therefore use the corresponding pre-period frontier proxies directly rather than impose a single calibrated index; in the dynamic analysis, the relevant neighbour-share analogue is allowed to evolve with the public stock. The sharpest sign predictions concern remoteness variables such as nearest-solved distance and gap. Local density variables are best read as auxiliary proxies that can also capture residual opportunities inside otherwise well-mapped protein families.

3.2 Experiments and Quality Dimensions

In structural biology, the relevant choice object is an experiment rather than a generic signal. Experiments differ both in what they reveal and in how they map into the data.

Definition 1 (Experiment)**.** *An experiment is a tuple $\xi = (x, a, m)$ where $x \in \mathcal{P}$ is the target protein, $a \in \mathcal{A} = \{R, Q, S, L\}$ is the action type, and $m \in \{Xray, cryoEM, NMR\}$ is the method. R denotes residue-coverage work, Q resolution improvement, S a distinct conformational state, and L a ligand-bound structure.*

For each protein x let

$$\mathbf{q}_x = (e_x, \rho_x, n_x^S, n_x^L), \tag{4}$$

where $e_x \in [0, 1]$ is residue coverage, $\rho_x \geq 0$ is a resolution-quality index (for example $\rho_x = 1/r_x$ when resolution is measured in Ångströms), and $n_x^S, n_x^L \in \mathbb{Z}_{\geq 0}$ are the numbers of distinct conformational states and ligand-bound contexts deposited for x. For proteins with no public structure, $(e_x, \rho_x, n_x^S, n_x^L) = (0, 0, 0, 0)$.

Rather than forcing all forms of experimental progress through a single coverage

variable, we track three uncertainty states:

$$u_0(x; \mathcal{F}_t, \mathbf{q}_x) = \sigma_0^2(x; \mathcal{F}_t) + \kappa_E h_E(e_x) + \kappa_Q h_Q(\rho_x), \tag{5}$$

$$u_S(x; \mathbf{q}_x) = \kappa_S h_S(n_x^S), \tag{6}$$

$$u_L(x; \mathbf{q}_x) = \kappa_L h_L(n_x^L). \tag{7}$$

Here, u_0 is the total uncertainty about the dominant static scaffold, u_S is the uncertainty about conformational diversity, and u_L is the uncertainty about ligand-bound structure. For later reference, define the Bayesian scaffold component

$$u_f(x; \mathcal{F}_t) \equiv \sigma_0^2(x; \mathcal{F}_t),$$

so that $u_0(x; \mathcal{F}_t, \mathbf{q}_x) = u_f(x; \mathcal{F}_t) + \kappa_E h_E(e_x) + \kappa_Q h_Q(\rho_x)$. The functions h_E and h_Q are positive, decreasing, and convex. The functions h_S and h_L are positive, decreasing, and discrete-convex, so one-step gains shrink with the number already observed. The additive form in (5)–(7) is deliberate: it avoids the degenerate implication that once one quality dimension is nearly exhausted, all other dimensions become valueless. In particular, resolution improvement, conformational-state work, and ligand-binding work can remain valuable even when residue coverage is already high. Proposition 1 uses only these shape restrictions; whenever later comparative statics invoke derivatives or single-valued choices, the relevant smoothness or uniqueness assumptions are imposed explicitly in the statement.

The public and private state at date t is

$$\mathcal{S}_t = \left(\mathcal{F}_t, \{\mathbf{q}_x\}_{x \in \mathcal{P}}, \{J_t(x)\}_{x \in \mathcal{P}}, \{I_{\ell x t}\}_{\ell, x}\right).$$

The baseline transition convention is as follows. An R experiment on x adds a new static public signal at x to $\mathcal{F}_t$, raises e_x, sets the post-experiment public indicator for x to one, and adds x to the acting laboratory's pipeline. A Q experiment adds a more precise static signal at x, raises ρ_x, and likewise adds x to the acting laboratory's pipeline. An S experiment raises n_x^S by one and an L experiment raises n_x^L by one; in the baseline model these actions do *not* update $\mathcal{F}_t$ or u_f. This is a maintained abstraction used in Propositions 3 and 6: any static-coordinate information bundled with an S or L deposition is absorbed into the already-solved state $J_t(x) = 1$ rather than modeled as a separate scaffold update.

Action R increases e_x, action Q increases ρ_x, action S raises n_x^S by one, and action L raises n_x^L by one. Diminishing returns is characterised by the next proposition.

Proposition 1 (Diminishing returns by quality dimension)**.** *Fix a protein x and hold all other arguments fixed.*

(i) For any feasible $\Delta_E > 0$, the reduction in static uncertainty produced by a residue-coverage improvement,

$$G_R(e_x; \Delta_E) = \kappa_E \Big[h_E(e_x) - h_E(e_x + \Delta_E) \Big],$$

is weakly decreasing in e_x on $[0, 1 - \Delta_E]$.

(ii) For any feasible $\Delta_Q > 0$, the reduction in static uncertainty produced by a resolution improvement,

$$G_Q(\rho_x; \Delta_Q) = \kappa_Q \Big[h_Q(\rho_x) - h_Q(\rho_x + \Delta_Q) \Big],$$

is weakly decreasing in ρ_x. Equivalently, equal increments in the resolution-quality index yield smaller uncertainty reductions when resolution is already high (for $\rho_x = 1/r_x$, this corresponds to smaller gains at already low Ångström values).

(iii) The one-step reduction in conformational uncertainty,

$$G_S(n_x^S) = \kappa_S \Big[h_S(n_x^S) - h_S(n_x^S + 1) \Big],$$

is weakly decreasing in n_x^S. The analogous one-step reduction in ligand uncertainty,

$$G_L(n_x^L) = \kappa_L \Big[h_L(n_x^L) - h_L(n_x^L + 1) \Big],$$

is weakly decreasing in n_x^L.

Intuitively, each quality dimension has a low-hanging-fruit logic. Early progress—resolving the first chunks of residue coverage, improving very coarse resolution, or adding the first new conformational/ligand observations—removes broad uncertainty. Once a protein is already well characterised along a given margin, equally sized additional improvements tend to answer narrower questions, so the uncertainty reduction from another step is smaller. Convexity (and discrete convexity for counts) is exactly the shape restriction that formalises this diminishing-return pattern.

3.3 Costs: Laboratory Capital and the Mapping to Empirical Margins

These diminishing-return objects pin down the benefit side of experimentation, but they do not yet explain *who* works on which proteins or why empirical activity separates into entry versus follow-up margins. To connect the model to the data, we next introduce

laboratory-specific setup costs and protein-level incumbency capital, which create distinct wedges between public novelty and novelty to a given lab.

Let $J_t(x) \in \{0,1\}$ indicate whether protein x has any public experimental structure by time t. The feasible action set is

$$\mathcal{A}(x,t) = \begin{cases} \{R\}, & J_t(x) = 0, \\ \{R, Q, S, L\}, & J_t(x) = 1. \end{cases} \tag{8}$$

Let $I_{\ell xt} \in \{0,1\}$ indicate whether laboratory ℓ has already built protein-specific capital for x by time t, and define the laboratory's pipeline stock by

$$\mathcal{L}_{\ell t} = \{x \in \mathcal{P} : I_{\ell xt} = 1\}.$$

$J_t(x)$ is a public-state variable; $I_{\ell xt}$ is lab-specific incumbency. The distinction matters empirically because a publicly solved protein can still be *new* to a laboratory.

Let $B_{\ell t}(x,a,m)$ denote the gross value of experiment (x,a,m) to laboratory ℓ, before costs. The cost of that experiment is

$$c_{\ell t}(x,a,m) = \eta_m \Big[\phi_m \mathbf{1}\{I_{\ell xt} = 0\} + g_{a,m}(\mathbf{q}_x)\Big] \chi(F_{xt}), \tag{9}$$

where $\phi_m > 0$ is a method-specific setup cost, $g_{a,m}(\mathbf{q}_x) \geq 0$ is the action-specific marginal cost, $\eta_m > 0$ is a method scale, and $\chi(F_{xt})$ is a positive increasing frontier-difficulty function. The crucial asymmetry is the indicator $\mathbf{1}\{I_{\ell xt} = 0\}$: the setup cost is paid whenever the protein is new to the laboratory, regardless of whether it is new to the field.

Let

$$\Pi_{\ell t}(x,a,m) = B_{\ell t}(x,a,m) - c_{\ell t}(x,a,m)$$

denote the net payoff. The next proposition maps the model directly to the margins used in the data.

Proposition 2 (Laboratory capital and empirical margins)**.** *Under Assumption 1, for any laboratory ℓ and time t:*

(i) A public-extensive *choice solves*

$$\max_{\substack{x \in \mathcal{P}:\, J_t(x)=0 \\ m}} \Pi_{\ell t}(x, R, m).$$

(ii) A within-pipeline follow-up *choice solves*

$$\max_{\substack{x \in \mathcal{L}_{\ell t}:\, J_t(x)=1 \\ a \in \mathcal{A}(x,t),\, m}} \Pi_{\ell t}(x, a, m).$$

(iii) An outside-pipeline follow-up *on an already solved protein solves*

$$\max_{\substack{x \in \mathcal{P}:\, J_t(x)=1,\, x \notin \mathcal{L}_{\ell t} \\ a \in \mathcal{A}(x,t),\, m}} \Pi_{\ell t}(x, a, m).$$

For any publicly solved protein x, feasible action a, and method m,

$$\Pi_{\ell t}(x, a, m \mid I_{\ell x t} = 1) - \Pi_{\ell t}(x, a, m \mid I_{\ell x t} = 0) = \eta_m \phi_m \chi(F_{xt}) > 0. \tag{10}$$

In words, the three choice problems differ by whether a protein is new to the field, new to the laboratory, or both. The public-extensive problem chooses proteins with $J_t(x) = 0$, so the lab is creating first public structures. The within-pipeline follow-up problem chooses proteins with $J_t(x) = 1$ and $x \in \mathcal{L}_{\ell t}$, so the lab is refining proteins where it already has protein-specific capital. The outside-pipeline follow-up problem also has $J_t(x) = 1$ but imposes $x \notin \mathcal{L}_{\ell t}$, so the target is already solved publicly yet still new to that laboratory's pipeline.

Equation (10) is important for the empirical design. It implies that public-intensive work on proteins outside a laboratory's current pipeline is not equivalent to a within-pipeline follow-up: it faces the same setup wedge as a new target even when the protein is already solved in the public record. This theory object is narrower than the broader lab-experience measures used later in the paper, which are best interpreted as noisy proxies for being outside the relevant protein pipeline rather than as literal measures of $I_{\ell x t}$.

3.4 The Value of Structural Knowledge

This subsection closes the model by specifying the objective that laboratories optimize. After defining what actions are feasible and how costly they are, we now define how much laboratories value different kinds of knowledge gains—both on the focal protein and through spillovers in the similarity tree—so that payoffs and comparative statics are well defined.

Laboratory ℓ values reductions in the three uncertainty states according to

$$v_\ell(\mathcal{F}_t) = \int_{\mathcal{T}} w_\ell(x) \left[\alpha_0 \max\left\{1 - \frac{u_0(x)}{q_0}, 0\right\} + \alpha_S \max\left\{1 - \frac{u_S(x)}{q_S}, 0\right\} + \alpha_L \max\left\{1 - \frac{u_L(x)}{q_L}, 0\right\} \right] d\mu(x), \tag{11}$$

where $w_\ell(x) > 0$ is an institution-specific weight function and $\alpha_0, \alpha_S, \alpha_L, q_0, q_S, q_L$ are positive constants. The gross value of experiment $\xi = (x, a, m)$ is

$$B_{\ell t}(x, a, m) = v_\ell(\mathcal{F}_t \cup \{\xi\}) - v_\ell(\mathcal{F}_t). \tag{12}$$

Assumption 1 (Well-posedness and attainment)**.** *For every laboratory ℓ, the weight function w_ℓ is measurable and integrable, so $\int_{\mathcal{T}} w_\ell(x)\, d\mu(x) < \infty$. The feasible sets used below are nonempty. Whenever the text refers to a maximizing experiment, target, or coverage choice, the relevant maximum is attained.*

Because the bracketed term in (11) is bounded between 0 and $\alpha_0 + \alpha_S + \alpha_L$, Assumption 1 guarantees that $v_\ell(\mathcal{F}_t)$ and hence $B_{\ell t}(x, a, m)$ are well defined. Because v_ℓ integrates over the tree, experiments on proteins that illuminate dense or topologically central neighbourhoods generate larger spillovers. This is the economic content of family density and hub-like positions in the similarity tree. Different institutions can be represented by different weight functions $w_\ell(x)$ and, in the comparative statics below, by different distributions of pipeline capital $I_{\ell x t}$. The model does not need a formal theorem for academic versus industry behaviour; that heterogeneity enters as a conditional prediction in Section 3.7.

3.5 AlphaFold2 as an Information Shock

AF2 provides a predicted structure for essentially every protein. In the baseline one-component Gaussian special case, we model it as an independent signal on the dominant scaffold state $Y(x)$:

$$\hat{y}^{AF}(x) = Y(x) + \varepsilon^{AF}(x), \qquad \varepsilon^{AF}(x) \sim \mathcal{N}(0, \omega(x)). \tag{13}$$

Let $A_x \in [0, 1]$ denote the empirical AF informativeness measure used later in the paper, namely the share of residues whose AF prediction exceeds a usable confidence threshold. We assume A_x is decreasing in $\omega(x)$, so higher A_x corresponds to more informative AF predictions.

Because $u_f(x;\mathcal{F}_t) = \sigma_0^2(x;\mathcal{F}_t)$ is a literal posterior variance, AF2 updates the scaffold component by exact Gaussian precision addition:

$$u_f^{AF}(x) = \left(\frac{1}{u_f(x)} + \frac{1}{\omega(x)}\right)^{-1} = \frac{u_f(x)\omega(x)}{u_f(x)+\omega(x)}. \tag{14}$$

The total static uncertainty entering laboratory payoffs is therefore

$$u_0^{AF}(x) = u_f^{AF}(x) + \kappa_E h_E(e_x) + \kappa_Q h_Q(\rho_x).$$

This keeps the AF2 update literally Bayesian on the spillover-rich scaffold component while allowing coverage and resolution gaps to remain separate own-target state variables. Appendix A.2 develops a richer OU-plus-nugget version in which AF2 attenuates only the smooth scaffold component and leaves a target-specific residual untouched; the result below is the one-component benchmark nested in that extension.

For $a \in \{R, Q\}$, let $d\Delta_f(y;x,a) \geq 0$ denote the infinitesimal reduction in scaffold uncertainty $u_f(y)$ induced by the spillover-rich part of an infinitesimal experiment of type a at target x. Under the Brownian-tree structure, $d\Delta_f(y;x,a)$ is largest near x and declines with tree distance.

This setup motivates the next proposition: AF2 scales down the scaffold component of the marginal value of static-structure experiments, with stronger attenuation where AF is more informative over the experiment's spillover footprint.

Proposition 3 (Bayesian attenuation of the scaffold component and differential substitutability)**.** *Fix laboratory ℓ and target protein x. For $a \in \{R, Q\}$, let $d\Delta_f(y;x,a) \geq 0$ denote the infinitesimal reduction in scaffold uncertainty $u_f(y)$ induced by the spillover-rich part of an infinitesimal experiment of type a at x. Suppose all affected proteins remain in the interior of the static in-play region, so $u_0(y) < q_0$ whenever $d\Delta_f(y;x,a) > 0$. Let $dB_{\ell t}^{pre,f}(x,a)$ and $dB_{\ell t}^{AF,f}(x,a)$ denote the corresponding pre- and post-AF marginal values transmitted through u_f. Then*

$$dB_{\ell t}^{AF,f}(x,a) = \bar{\lambda}_{\ell t}(x,a)\, dB_{\ell t}^{pre,f}(x,a), \tag{15}$$

where

$$\bar{\lambda}_{\ell t}(x,a) = \frac{\int_{\mathcal{T}} w_\ell(y)\lambda(y)\, d\Delta_f(y;x,a)\, d\mu(y)}{\int_{\mathcal{T}} w_\ell(y)\, d\Delta_f(y;x,a)\, d\mu(y)} \in (0,1), \qquad \lambda(y) = \frac{\omega(y)^2}{\big(u_f(y)+\omega(y)\big)^2}.$$

If an R or Q experiment affects payoffs only through the scaffold component u_f, so there are no separate contemporaneous gains through e_x or ρ_x, then (15) is the full marginal-value effect.

Under the analogous interior condition for u_S and u_L, and under the maintained transition convention from Section 3.2, AF2 does not change the marginal value of S and L experiments:

$$dB^{AF}_{\ell t}(x, S) = dB^{pre}_{\ell t}(x, S), \qquad dB^{AF}_{\ell t}(x, L) = dB^{pre}_{\ell t}(x, L).$$

The proposition is intentionally local. For finite experiments, the relevant attenuation is an average of marginal factors along the path of uncertainty reduction, not a universal scalar multiplier. It deliberately isolates the Bayesian scaffold channel. When static experiments also improve own-target coverage or resolution, AF2 attenuates the scaffold part of their payoff but need not mechanically attenuate those additional own-target gains. Appendix A.2 shows how the same logic extends when AF2 loads on a smooth scaffold component but leaves a target-specific residual untouched.

AF2 can also change costs, but the relevant channel differs by method. Let $\delta^{\phi}_{m}(A_x) \geq 0$ denote setup-cost savings and $\delta^{g}_{a,m}(A_x) \geq 0$ variable-cost savings generated by AF2, with

$$0 \leq \delta^{\phi}_{m}(A_x) \leq \phi_m, \qquad 0 \leq \delta^{g}_{a,m}(A_x) \leq g_{a,m}(\mathbf{q}_x).$$

The post-AF cost is

$$c^{AF}_{\ell t}(x, a, m) = \eta_m \Big[\big(\phi_m - \delta^{\phi}_{m}(A_x)\big)\mathbf{1}\{I_{\ell xt} = 0\} + g_{a,m}(\mathbf{q}_x) - \delta^{g}_{a,m}(A_x) \Big] \chi(F_{xt}). \tag{16}$$

For X-ray crystallography, the natural case is large δ^{ϕ}_{m} and small $\delta^{g}_{a,m}$ because AF2 is useful for molecular replacement. For cryo-EM, the natural case is small δ^{ϕ}_{m} and larger $\delta^{g}_{a,m}$ because AF2 helps mainly with map interpretation. For NMR, both are typically small.

Write $\Pi^{E,s}_{\ell t}(x, R, m)$ for the net payoff from a public-extensive choice on target x under regime $s \in \{pre, AF\}$, and $\Pi^{I,s}_{\ell t}(y, a, m)$ for the corresponding within-pipeline intensive payoff. To connect the cost decomposition to observable margin reallocation, compare a laboratory's payoff from opening a new public target that is also new to the laboratory to its payoff from running an additional follow-up experiment inside an existing pipeline. The next proposition shows that setup-cost relief has a direct extensive-margin effect, whereas variable-cost relief has a sign that depends on relative frontier difficulty.

Proposition 4 (Setup-cost versus variable-cost channels)**.** *Fix method m, a candidate public-extensive target x with $J_t(x) = 0$ and $I_{\ell xt} = 0$, and a within-pipeline follow-up target y with $J_t(y) = 1$ and $I_{\ell yt} = 1$. Let*

$$D_m(x, y, a) = \Pi^{E,AF}_{\ell t}(x, R, m) - \Pi^{I,AF}_{\ell t}(y, a, m)$$

be the post-AF payoff difference between the public-extensive option and the within-pipeline intensive option.

(i) Holding all benefits and other cost terms fixed,

$$\frac{\partial D_m(x,y,a)}{\partial \delta_m^{\phi}(A_x)} = \eta_m \chi(F_{xt}) > 0.$$

A setup-cost saving unambiguously tilts the comparison toward the public-extensive option.

(ii) Suppose a variable-cost saving δ^g applies symmetrically to the relevant action on both margins. Then

$$\frac{\partial D_m(x,y,a)}{\partial \delta^g} = \eta_m \left[\chi(F_{xt}) - \chi(F_{yt})\right],$$

whose sign is in general ambiguous.

Proposition 4 gives the core margin intuition. A setup-cost saving acts like an entry subsidy for new targets when the candidate target is also new to the laboratory, because that cost is paid only when opening a pipeline ($I_{\ell xt} = 0$), so it mechanically tilts choices toward public-extensive work. If a laboratory has already sunk the setup cost on x despite $J_t(x) = 0$, the derivative in part (i) is zero because the setup term is absent. By contrast, a variable-cost saving applies to experiments on both margins, so much of it cancels in the comparison; what remains depends on relative frontier scaling through $\chi(F_{xt})$ versus $\chi(F_{yt})$, making the sign ambiguous in general.

To study where AF2-induced reallocation goes across candidate targets, we next summarise all AF benefit and cost changes on the public-extensive margin in a single object and ask how it varies with frontier position. Write

$$\Delta_{\ell t}^{E}(x,m) = \Pi_{\ell t}^{E,AF}(x,R,m) - \Pi_{\ell t}^{E,pre}(x,R,m)$$

for the net AF shock to the payoff from a public-extensive experiment at protein x.

Proposition 5 (Frontier comparative statics on the public-extensive margin)**.** *Under Assumption 1, for each method m, let $x^{\star}_{pre}(m)$ be the maximal-frontier element of*

$$\arg \max_{x:\, J_t(x)=0} \Pi_{\ell t}^{E,pre}(x,R,m),$$

and let $x^{\star}_{AF}(m)$ be the maximal-frontier element of

$$\arg \max_{x:\, J_t(x)=0} \Pi_{\ell t}^{E,AF}(x,R,m).$$

If the net AF shock $\Delta^E_{\ell t}(x,m)$ is weakly increasing in the frontier index F_{xt}, then

$$F_{x^\star_{AF}(m),t} \geq F_{x^\star_{pre}(m),t}.$$

A sufficient condition is that AF2 attenuates scaffold-related value less at higher frontier positions and that any setup-cost saving is not so strongly concentrated away from the frontier as to overturn that ordering.

Intuitively, this is a monotone-selection result: if AF2 raises the payoff to first-ever public work more strongly for harder, more frontier proteins, then the best post-AF target cannot move inward. At minimum it stays put, and in general it shifts toward the frontier.

Because F_{xt} is time-varying, first-ever public structures can also change future payoffs even conditional on the AF regime. This creates a dynamic complementarity: today's extensive entry can lower tomorrow's local frontier difficulty for nearby solved proteins, so follow-up work may become cheaper and more attractive even when its direct scaffold informational content is unchanged under the maintained transition convention. Let $B^s_{\ell t}(x,a,m)$, $c^s_{\ell t}(x,a,m)$, and $\Pi^s_{\ell t}(x,a,m)$ denote gross value, cost, and net payoff under regime $s \in \{pre, AF\}$.

Proposition 6 (Dynamic local spillovers of extensive entry)**.** *Maintain the transition convention from Section 3.2: S and L actions update only u_S and u_L and do not change $\mathcal{F}_t$ or u_f. Suppose protein z receives a first-ever public structure between t and $t+1$, so that $\mathcal{F}_{t+1} = \mathcal{F}_t \cup \{s_z\}$ for some new public signal at z and $J_{t+1}(z) = 1$. Fix another protein $y \neq z$ with $J_t(y) = 1$, laboratory ℓ, and method m. If the new entry weakly lowers local frontier difficulty at y, so $F_{y,t+1} \leq F_{yt}$, then for any $a \in \{S, L\}$,*

$$B^s_{\ell,t+1}(y,a,m) = B^s_{\ell t}(y,a,m), \qquad c^s_{\ell,t+1}(y,a,m) \leq c^s_{\ell t}(y,a,m), \qquad \Pi^s_{\ell,t+1}(y,a,m) \geq \Pi^s_{\ell t}(y,a,m).$$

Thus neighbouring first-ever structures weakly raise the payoff to later auxiliary follow-up work on already solved proteins whenever they reduce local frontier difficulty. For $a \in \{R, Q\}$, the same local entry lowers frontier costs but also changes the remaining scaffold uncertainty on the benefit side, so the sign of the net payoff effect is in general ambiguous.

The same first-ever structure also mechanically expands the feasible action set on z itself from $\{R\}$ to $\{R, Q, S, L\}$ by (8). Proposition 6 concerns the additional cross-protein effect on nearby already solved targets.

3.6 Coverage-Depth Comparative Statics

To connect the model to coverage-depth regressions, consider proteins that are comparable in all respects other than frontier position. Conditional on selecting target x and method m, let the laboratory choose a continuous coverage increment $\Delta e \in [0, 1 - e_x]$ by solving

$$\max_{\Delta e \in [0, 1-e_x]} \left\{ B(\Delta e) - \eta_m \left[\phi_m \mathbf{1}\{I_{\ell xt} = 0\} + g_{R,m}(\Delta e) \right] \chi(F_{xt}) \right\}, \tag{17}$$

where B is increasing and concave in Δe and $g_{R,m}$ is increasing and convex in Δe.

Proposition 7 (Breadth over depth at the frontier)**.** *Suppose the maximizer of* (17) *is unique for each* (x, m)*, and denote it by* $\Delta e^\star(x, m)$*. Then* $\Delta e^\star(x, m)$ *is weakly decreasing in* F_{xt}*.*

Intuitively, frontier targets are more expensive to deepen: the frontier multiplier $\chi(F_{xt})$ scales the marginal cost of additional coverage, while the marginal benefit of depth is diminishing because B is concave. As frontier position rises, the laboratory optimally chooses a smaller within-target increment Δe and shifts effort toward breadth across targets rather than depth within a target.

3.7 Theory-guided Empirical Predictions

The propositions above deliver a few sharp comparative statics and a broader set of mechanism-based interpretations. The items below should therefore be read as theory-guided empirical predictions derived from the results above. Prediction 2 and Prediction 4 map most directly into Propositions 5 and 7. The remaining predictions combine those formal results with auxiliary assumptions about budgets, empirical coding, method-specific cost channels, and institutional objective functions.

Prediction 1 (Reallocation rather than pure expansion)**.** *With approximately fixed research budgets, AF2 reweights experimental effort away from within-pipeline scaffold-oriented follow-up work and toward public-extensive activity, with the strongest additional response on margins that still pay the new-to-protein setup wedge. Total experimental volume may rise or fall, but the robust model implication is reallocation across margins rather than a uniform expansion. This is a budget-conditioned synthesis rather than a literal implication of the static choice problem alone.*

Prediction 2 (Frontier exploration on the margins that bear the setup wedge)**.** *When the net AF shock* $\Delta^E_{\ell t}(x, m)$ *is increasing in baseline frontier position, post-AF reallocation toward new public targets is stronger at the frontier. The same logic applies to public-intensive choices on proteins outside a laboratory's existing pipeline, because they pay the same setup wedge. This is*

the empirical analogue of Proposition 5: the theorem is about the best target, while the prediction extends that monotone-selection logic to extensive-type empirical margins more generally. The sharpest sign prediction concerns extensive-type outcomes on those margins; it does not *imply that every measure of total activity must rise more at the frontier, since follow-up work can also expand within already mapped regions. In specifications that use remoteness variables such as distance to the nearest solved protein or the local gap, this implies more positive post-AF interactions for more remote proteins. Density variables such as cluster and neighbour PDB shares proxy the same latent object less cleanly and can also reflect residual opportunities inside already active families, so their sign is less sharp.*

Prediction 3 (Weaker frontier response for within-pipeline follow-ups)**.** *Within-pipeline follow-up experiments need not be invariant to AF2, but their target choice is less sensitive to frontier conditions than public-extensive or outside-pipeline activity because they avoid the setup wedge only on proteins already in the laboratory's pipeline. This is a mechanism-based comparative static rather than a separate theorem.*

Prediction 4 (Breadth over depth)**.** *Conditional on a public-extensive deposition, frontier proteins should receive smaller coverage increments Δe on average than proteins closer to existing solved neighbours. This is the empirical counterpart of Proposition 7.*

Prediction 5 (Intensive-margin composition)**.** *Within public-intensive work, AF2 compresses the scaffold component of static-structure actions R and Q but need not eliminate own-target completeness or resolution gains. Under the maintained transition convention, it leaves the marginal value of S and L unchanged. At the same time, AF2 can reduce the cost of some static experiments. The net composition effect is therefore ambiguous ex ante and must be estimated. When the empirical coding aggregates conformational-state and ligand-binding depositions, the relevant theoretical prediction concerns their combined share relative to the more substitutable static-structure categories.*

Prediction 6 (Method heterogeneity)**.** *Because X-ray crystallography can benefit materially through setup-cost savings, it should exhibit a comparatively strong response on margins that require new-target setup or new-to-protein investment. Cryo-EM can also exhibit a strong extensive-margin response when AF2 mainly lowers variable costs on hard targets. NMR should respond least. This is a method-specific interpretation of Proposition 4, not a universal theorem ranking methods in all environments.*

Prediction 7 (Institutional heterogeneity)**.** *If academic laboratories have flatter weight functions $w_\ell(x)$ and smaller effective pipeline stocks than industry laboratories, then AF2 should induce a stronger shift toward public-extensive and outside-pipeline exploration in academia. This is*

a qualitative prediction rather than a formal theorem because it depends on institution-specific objective weights and stock distributions.

Prediction 8 (Within-pipeline versus new-to-protein channels)**.** *Among public-intensive depositions, the margins that behave most like public-extensive search should be stronger on proteins that are new to the laboratory than on within-pipeline follow-ups, because only the former pay the setup wedge in* (16)*. Empirical exercises based on broad pre-period lab experience are, therefore, best read as noisy proxies for this channel rather than literal tests of* $I_{\ell xt}$*. This is a mechanism-oriented interpretation of the model's cost wedge rather than a separate theorem.*

Proposition 6 yields a separate dynamic implication. As neighbouring first-ever structures accumulate, local frontier difficulty can fall even after the direct AF2 shock is conditioned out. In the model this has the sharpest sign for already solved neighbouring proteins: if a nearby first-ever structure lowers frontier difficulty, the payoff to later SL-type follow-up work weakly rises, because the benefit side of S/L is unchanged while the frontier-related cost term falls. For R/Q work and for further extensive entry, the same dynamic accumulation changes both costs and remaining scaffold uncertainty, so the sign is not pinned down. This motivates the later time-varying neighbour-share analysis as a supplementary empirical exercise rather than one of the headline eight predictions.

The empirical category of "corroborating" depositions does not map one-for-one into a primitive action in the model. The most natural interpretation is as low-insight public-intensive work that is closer to the static-structure component than to the S or L components. The model, therefore, predicts that corroborating activity should behave more like the substitutable part of intensive work than like conformational or ligand-specific experiments. In the empirical coding, any first-ever public structure is assigned to the extensive margin even if it simultaneously improves coordinate precision, while conformational-state and ligand-binding depositions are pooled into a combined SL category. Throughout the empirical sections, the refinement category Q is coded narrowly by improved reported resolution. A method change that does not improve resolution is not treated as a separate primitive action in the data and is absorbed into corroborating work unless it also generates an R or SL contribution.

Taken together, these predictions are best read as a structured mapping from the model's comparative statics into empirical margins and coding choices. Some are direct theorem-guided sign predictions; others are mechanism-based syntheses used to discipline the empirical interpretation.

3.8 Summary and Relationship to the Empirical Design

The model is intentionally stylised. Brownian motion on an estimated similarity tree is a tractable description of homology spillovers, not a literal model of protein structure.[4] The scaffold posterior u_f has a literal Bayesian interpretation in the one-component benchmark, while the broader states u_0, u_S, and u_L summarise how that scaffold uncertainty combines with own-target quality gaps, conformational heterogeneity, and ligand-specific uncertainty. What the model does provide is a clean bridge from theory to the empirical design. The main latent objects map naturally into observables: AF informativeness is proxied by AF coverage A_x, frontier position by the separate frontier proxies used in the empirical sections rather than by a calibrated scalar index, dynamic local frontier movement by the time-varying neighbour-share analogue, public-extensive versus public-intensive status by whether a protein has any prior PDB deposition, and laboratory incumbency by whether a target lies inside a laboratory's existing protein pipeline. The later empirical lab-experience measures are broader than that object and should be interpreted only as noisy proxies for $I_{\ell xt}$. The empirical specifications, therefore, test qualitative comparative statics and mechanism-based predictions suggested by the theory rather than a tight structural calibration of its primitives. Several of the predictions in Section 3.7 are best read as disciplined empirical interpretations of the preceding propositions rather than as one-to-one theorem tests.

4 Data

To study how AF2 reshaped scientific production in structural biology, we construct a novel dataset that links protein-level characteristics, structural knowledge, and experimental output over time. For each protein, we measure the extent of experimentally resolved structural information and the quality and coverage of AF2 predictions. The analysis operates at two levels: a protein–year panel tracking cumulative experimental activity, and individual PDB depositions recording the specific experimental choices made by research teams.

4.1 Proteins as Scientific Problems

Each protein represents an open scientific problem: determining how its amino acid chain folds into a three-dimensional structure that governs biological function. The starting point

[4] Appendix A.2 demonstrates robustness to more complex modelling choices that are closer to protein structure characteristics.

is the Universal Protein Resource Knowledgebase (UniProtKB), the central repository for protein sequence and functional information (The UniProt Consortium, 2025).[5] UniProtKB currently contains over 250 million protein entries, of which 573,661 have been manually curated through literature review and curator-verified computational analysis (Swiss-Prot). We restrict the analysis to proteins in Swiss-Prot created by 2016, ensuring that all proteins in the sample were known before the treatment period begins. This yields 552,566 proteins that constitute the sample for this study.

Each UniProtKB record is indexed by a stable accession number and provides rich 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 relevance.

4.2 Experimental Structures

The Protein Data Bank (PDB) is the worldwide archive of experimentally determined structures of biological macromolecules (Berman et al., 2000).[6] Each deposition is a revealed 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 depositions 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 deposition records the year it was made, the depositing team, the method used (X-ray crystallography, cryo-electron microscopy (cryo-EM), or nuclear magnetic resonance (NMR) spectroscopy), the resolution of the structure, and the associated publication.[7] We identify the scientists who determined each PDB structure from the deposition's author list and match them to OpenAlex authors. Using all matched structure authors, we classify affiliations as university, company, government, or other, and compute the share of each

[5] UniProt constructs protein sequences by translating gene sequences from major genome repositories, providing a standardized 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 from most non-germline immunoglobulins and T-cell receptors, synthetic sequences, patent application sequences, small fragments (< 8 amino acids), and pseudogenes.

[6] 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.

[7] 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.

type per deposition by counting each listed affiliation equally. The last author serves as a proxy for the laboratory, and laboratory experience is measured by the number of PDB depositions associated with the last author before 2017; a laboratory is classified as incumbent if its last author has at least one prior deposition.

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. For each deposition, SIFTS identifies which residues are experimentally observed. Multiple depositions may exist for a single protein, covering different sequence regions, conformational states, or ligand-bound forms, and a single deposition typically resolves only a fragment of the protein. We define cumulative *PDB coverage* as the share of protein p's residues that have been experimentally resolved in at least one deposition by year t:

$$e_{pt} = \frac{1}{L_p} \sum_{i=1}^{L_p} \mathbf{1}\{\text{residue } i \text{ observed in at least one deposition by year } t\}. \tag{18}$$

Coverage is inherently continuous—a protein may have select domains resolved across multiple depositions spanning many years before the full chain is characterised. Each deposition d can contribute to the coverage of the protein:

$$\Delta e_d = \frac{1}{L_p} \sum_{i=1}^{L_p} \mathbf{1}\{\text{residue } i \text{ observed in } d \text{ but not in any prior deposition}\}. \tag{19}$$

Together, e_{pt} and Δe_d measure whether, and to what degree, the structural problem has been solved at both the protein–year and the deposition level.

A deposition is *extensive* if it provides the first experimentally resolved structure for protein p; it is *intensive* if at least one prior deposition for p already exists. Within the intensive margin, follow-up depositions serve scientifically distinct purposes—extending residue coverage, improving resolution, capturing alternative conformational states, or characterizing ligand-bound forms (see Appendix A.3.2 for a detailed discussion). We classify every deposition d into one of five mutually exclusive categories according to its primary scientific contribution.

A *novelty* deposition (E) provides the first experimentally resolved structure for protein p. A *completeness* deposition (R) resolves residues not covered by any prior deposition, extending the fraction of the protein that has been experimentally observed. A *mechanistic*

insight deposition (SL) captures the protein in a previously unobserved conformational state (e.g., active/inactive, open/closed), as identified by Foldseek structural clustering, or includes a biologically relevant ligand (drug, cofactor, substrate) not previously co-crystallized with the protein. A *refinement* deposition (Q) achieves higher resolution than any prior deposition for the same protein, or uses a new experimental method. Finally, a *corroborating* deposition (Corr.) provides an independent determination of a known structure—a different crystal packing environment, a validation of a previous model, or a control in a study.[8]

These categories are constructed from observable features of depositions and are intended to capture economically meaningful distinctions in experimental activity. They should therefore be interpreted as approximations to underlying scientific objectives rather than direct observations of experimental intent.

4.3 The Structural Knowledge Frontier

Proteins are not isolated problems. They are related to one another through evolutionary descent, and this relatedness has a direct economic implication: resolving the structure of one protein generates a positive externality for its sequence neighbours, because it provides a homology modelling template for structurally similar proteins. Proteins sharing more than 30% sequence identity almost always share the same three-dimensional fold (Rost, 1999). 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.

We construct a tree that organises the ∼573K SwissProt proteins by evolutionary relatedness. The construction proceeds in three steps. First, we group proteins into *clusters* of closely related sequences using MMseqs2 (Steinegger and Söding, 2017): 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* (∼69,000 proteins, ∼12% of the sample). Third, we build a tree over the

[8] Because these flags are not mutually exclusive, we assign each deposition to a unique primary category using the priority ordering $E \succ R \succ SL \succ Q \succ$ Corr. This ordering reflects the theoretical hierarchy: a first structure generates information that propagates to sequence neighbours via homology modelling, residue extension and resolution improvement refine the baseline structural model, while conformational and ligand discoveries create target-specific value.

connected clusters using the neighbour-joining algorithm (Saitou and Nei, 1987), which constructs a tree by iteratively joining the pair of nodes that minimises total branch length, producing a topology that approximates the true evolutionary distances between clusters (see Appendix A.3.3 for construction details).

The tree topology is fixed—determined by evolutionary biology, 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 p and year t, we characterise p's position relative to this evolving frontier along two dimensions: distance and density. We measure each protein's distance to the frontier using two tree-path metrics: the *distance to the nearest solved protein*, which directly captures how far a protein is from a usable homology modelling template;[9] and the *gap*, defined as the sum of distances to the nearest solved protein on each side, which measures the width of the unsolved region surrounding the protein.

We complement these with two measures of local knowledge density: the *share of cluster members* with at least one deposition by year t, which captures within-cluster saturation—a cluster where most members are already solved offers diminishing returns to an additional deposition—and the *share of adjacent clusters* on the tree with PDB coverage, which captures the density of structural knowledge in the protein's neighbourhood. Together, these four measures characterise the possibility of homology modelling—and therefore the marginal value of a new deposition—for each protein at each point in time. All are recomputed each year as the set of solved proteins expands.[10]

[9] This measure is related to the pairwise "homology score" used by Hill and Stein (2026), defined as the raw sequence identity between a protein and its nearest experimentally solved neighbour. When a protein's cluster is directly connected to a solved cluster on the tree, the two measures are equivalent up to a scale transformation: a homology score of 30% corresponds to a tree distance of 0.70, and scores above 30% map monotonically to shorter distances. The measures diverge when the nearest solved protein is reached only through intermediate clusters on the tree: the tree-path distance sums edge weights along the path, which can differ from the direct pairwise identity.

[10] Distance measures are undefined for proteins in clusters that share less than 30% sequence identity with all other clusters, since these clusters have no edges in the similarity network and therefore no path to any solved protein. The neighbour density measure is undefined for clusters with no adjacent clusters on the tree. For these isolated proteins, we impute distance and gap variables at the observed maximum and share variables at zero, and flag them with an indicator (isolated). Regressions that use frontier measures control for this indicator.

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).[11]

To summarise the treatment intensity at the protein level, we define the fraction of residues in protein p whose predicted pLDDT exceeds threshold q:

$$\text{AF cov}_p^q = \frac{1}{L_p} \sum_{i=1}^{L_p} \mathbf{1}\{\text{pLDDT}_{pi} \geq q\}, \tag{20}$$

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).[12] 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.

The quality of each prediction—and thus AF coverage—depends on how informative the inputs are for a given protein. MSA depth is the primary driver: proteins from large, evolutionarily diverse families yield rich covariation patterns and high per-residue confidence, whereas lineage-specific proteins with few homologs yield shallow alignments and lower confidence. Template availability provides a secondary boost, though Jumper et al. (2021) show that AF2 performs nearly as well without it. Coverage is also lower in intrinsically disordered regions and in structurally complex proteins (e.g., multi-domain or membrane-spanning), reflecting genuine prediction difficulty. Crucially, these determinants—MSA depth, disorder, and structural complexity—are largely fixed by evolutionary history and therefore predetermined with respect to the release of AF2. At the same time, these biological characteristics may be correlated with broader patterns of

[11] 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.

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

scientific attention and experimental tractability. As a result, variation in AF2 coverage should be interpreted as capturing differential exposure to the AF2 shock across proteins with different underlying properties, rather than as a source of quasi-random treatment intensity.

4.5 Descriptive Statistics

Table 2 reports summary statistics for the four components of the dataset. The protein–year panel (Panel A) covers approximately 553,000 SwissProt proteins observed around the AF2 release between 2017 and 2025. Experimental activity is sparse: in the average protein–year, there are 0.033 depositions with the intensive margin dominating, and the cumulative PDB coverage ($e_{p,t}$) averages 4%.

At the deposition level (Panel B), the PDB is overwhelmingly intensive: only a small fraction of depositions provide the first-ever structure for a protein. Each deposition contributes modest new coverage on average (Δe_d), with extensive-margin depositions contributing substantially more than intensive-margin ones. Approximately two-thirds of depositions originate from academic institutions, and a substantial share of intensive-margin depositions come from laboratories that have previously deposited a structure for the same protein.

Figure 1 displays the time series of PDB depositions by category and method. There is no obvious discontinuity in total deposition volume at the AF2 release, consistent with reallocation rather than expansion. Figure 2 shows the lab composition over time: entrant labs account for a growing share of annual depositions, while incumbent labs maintain substantially larger protein portfolios.

Panel C characterises the structural knowledge frontier as of 2016, prior to the treatment period. Most proteins have zero PDB coverage, and frontier distances are large for the majority of the proteome. Approximately 16% of proteins are classified as isolated. The two distance measures (distance to nearest solved, gap) and two density measures (share PDB within cluster, share PDB in neighbouring cluster) are strongly but imperfectly correlated, loading on related but distinct dimensions of structural isolation and local knowledge density. Figure 5 shows that the frontier has been contracting steadily and the density of structural knowledge within clusters increasing since the 1990s.

Panel D summarises AF2 predictions. The treatment is strong and widespread: the mean share of residues predicted with pLDDT ≥ 70 exceeds 85%, with a pronounced mass near 100% (Figure 3, right panel). In contrast, experimental coverage was sparse and highly uneven prior to AF2—the left panel of Figure 3 shows that most proteins had

zero PDB coverage as of 2020, and the inset reveals that even among proteins with some experimental structure, coverage was frequently partial. The field goes from near-zero to near complete structural information thanks to AF2.

A critical feature for identification is the relationship between AF2 coverage and frontier position. If AF2 were just a proxy for well-studied proteins, high coverage would correlate with short frontier distance. Figure 4 shows that AF2 coverage is largely orthogonal to frontier distance, supporting the identifying assumption that treatment variation is not mechanically correlated with pre-existing frontier position. The exception is proteins with very low AF coverage ($< 20\%$), which tend to have higher frontier distances—these are lineage-specific sequences with few homologs that produce both shallow MSAs and few solved relatives. This group is small and the confound works against our results, attenuating estimated interactions between AF coverage and frontier distance.

These features—widespread but heterogeneous AF2 coverage, sparse and intensive-margin-dominated experimental activity, rich variation in frontier position, and approximate orthogonality between treatment intensity and frontier distance—motivate the empirical strategy.

5 Empirical Strategy

We exploit cross-protein variation in exposure to AF2 to estimate how a major AI advance reshapes scientific production. The identifying variation is heterogeneity in the quality of AF2's predictions: while AF2 generates predictions for essentially all known proteins, the share of each protein predicted at usable confidence varies substantially across proteins, providing a continuous measure of treatment intensity. We estimate the effect at two complementary levels. The protein–year panel identifies which proteins gain or lose experimental attention; the deposition-level analysis characterises the composition of experimental choices conditional on attention being directed to a protein.

Both levels draw from the same universe: SwissProt proteins created by 2016, prior to AF2. The protein–year panel tracks these proteins over 2017–2025; the deposition-level analysis studies all PDB depositions for these proteins over the same period. The baseline specification is

$$Y_i = \beta \, \text{AF cov}_p \times \text{Post}_t + \alpha_p + \lambda_t + \epsilon_i \tag{21}$$

where i indexes either protein–year pairs or individual depositions, and Post_t equals one for years after 2021.[13] In the baseline specification, AF cov_p is the share of residues predicted by

[13] AF2 predicted structures and source code were first released in July 2021. We define the post-period as

AF2 with pLDDT ≥ 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. Protein fixed effects α_p absorb all time-invariant protein characteristics—sequence length, taxonomic family, intrinsic disorder, baseline experimental tractability, and the availability of pre-existing structural templates—so identification comes entirely from the interaction with the post-period indicator. Year fixed effects λ_t capture global trends in experimental output. Standard errors are clustered at the protein level.

The coefficient β measures whether proteins that received more informative AF2 predictions experienced a differentially larger change in experimental activity after 2021 relative to proteins that received less informative predictions. This is a continuous difference-in-differences design: treatment intensity varies across proteins by a characteristic that is largely determined by amino-acid sequence, evolutionary constraints, and biophysical foldability—all fixed well before 2021.

At the protein–year level, we estimate five outcomes that progressively characterise whether a protein receives experimental attention, how much, how deep, and where: (1) an indicator for any PDB deposition in year t; (2) the total number of depositions; (3) the number of intensive (follow-up) depositions, estimated on proteins with at least one PDB structure before 2017; (4) cumulative PDB coverage e_{pt}; and (5) frontier distance interacted with the extensive indicator, which captures where in sequence space new exploration occurs. Outcomes (1), (4), and (5) are estimated by OLS on the full balanced panel (all SwissProt proteins $\times$ 9 years, 2017–2025) with protein and year fixed effects. The count outcomes (2) and (3) are estimated by Poisson pseudo-maximum likelihood (PPML) with the same fixed effects.[14]

At the deposition level, we estimate outcomes that characterise the composition of experimental choices conditional on a deposition being made: (1) an indicator for whether the deposition is a first-ever structure (extensive), estimated with protein and year fixed effects; (2) the coverage increment Δe_d for extensive depositions, estimated with organism and year fixed effects and protein-level controls (log sequence length, number of Pfam domains, membrane protein indicator, disease association flag, and log pre-treatment

2022–2025, allowing the second half of 2021 to serve as a partial adoption period. Results are robust to using the expanded release date.

[14]The PPML estimator conditions on proteins with non-zero within-group variation in the outcome, which focuses identification on the relevant margin: proteins that receive at least some experimental attention over the sample period. The extensive margin—whether a protein receives its first-ever experimental structure—is not estimated in the protein–year panel because the survival structure of the outcome (each protein can enter the PDB at most once) precludes protein fixed effects. Instead, the extensive margin enters at the deposition level, where every depositing protein contributes at least one observation and protein fixed effects are estimable.

publication count);[15] (3) the coverage increment Δe_d for intensive depositions, estimated with protein and year fixed effects; and (4) mutually exclusive action-type indicators for intensive depositions. The action types—completeness (R: residue extension), mechanistic insight (SL: conformational states and ligand complexes), refinement (Q: improved resolution), and corroboration (Corr.: independent validation)—decompose follow-up work into economically distinct activities that map onto different motives for experimentation. Binary outcomes and the coverage measure are estimated by OLS. We interpret these results as characterising what is deposited conditional on a protein receiving attention, not as causal effects on a fixed population: selection into the deposition sample is precisely the protein-panel extensive outcome.

To test whether AF2 reshapes the *direction* of exploration, we interact AF coverage with measures of each protein's position on the structural knowledge frontier:

$$Y_i = \beta_1 \,\text{AF cov}_p \times \text{Post}_t + \beta_2 \,\text{AF cov}_p \times \text{Post}_t \times F_{pt} + \gamma_1 \, F_{pt} + \gamma_2 \, F_{pt} \times \text{Post}_t + \alpha_p + \lambda_t + \epsilon_i \tag{22}$$

where F_{pt} is one of the four frontier measures defined Section 4.3: distance to nearest solved protein, gap, share of cluster members with PDB coverage, and share of neighbouring clusters with PDB coverage. The coefficient β_2 tests whether AF2 has a larger effect on experimental activity for proteins closer to existing structural knowledge or embedded in denser regions of the similarity tree.

The design relies on the assumption that, absent AF2, proteins with different levels of AF2 coverage would have followed parallel trends in experimental activity. Three features of the setting support this assumption. First, AF2 coverage is primarily determined by evolutionary depth—MSA richness, sequence conservation, and structural complexity—not by prior experimental effort. Protein fixed effects absorb these characteristics in levels; the identifying assumption requires only that they do not generate differential trends. Second, Figure 4 shows that AF2 coverage is largely orthogonal to pre-existing frontier distance: two proteins at the same point on the knowledge frontier can receive very different AF2 coverage depending on the evolutionary depth of their protein family. Third, Table 15 confirms that observable protein characteristics do not exhibit differential pre-trends across AF2 coverage quartiles, and the event-study estimates in Figure 7 show no

[15] Log sequence length and the number of Pfam domains control for structural complexity and evolutionary family size—longer proteins and those with fewer annotated domains tend to receive lower-confidence AF2 predictions while also being harder to resolve experimentally. The membrane protein indicator accounts for the distinct experimental pipeline (cryo-EM rather than X-ray crystallography) and systematically lower AF2 confidence in transmembrane regions. The disease association flag and pre-treatment publication count absorb variation in scientific interest that predicts both the likelihood of experimental follow-up and, through historical research intensity on related families, the depth of evolutionary information available to AF2.

systematic pre-treatment divergence.

We nonetheless note two residual concerns. Proteins with high AF2 coverage come from large, well-studied evolutionary families whose research trajectories could differ from those of less well-characterised proteins for reasons unrelated to AF2. The post-2021 period also coincides with the continued expansion of cryo-EM, which has disproportionately benefited certain protein classes. Year fixed effects absorb aggregate method trends, and we estimate method-specific treatment effects (Table 10) to assess whether AF2 operates through a particular experimental technology. We further probe robustness by (i) adding lab fixed effects to control for compositional shifts in the depositor community (Tables 29 and 30); (ii) varying the AF coverage threshold across pLDDT $\geq$ 50, 70, and 90 (Tables 16–21); (iii) restricting to proteins with AF coverage above 0.2 to exclude those where the treatment measure may correlate with frontier position (Tables 26–27); and (iv) running placebo tests that randomly permute AF coverage across proteins (Tables 13 and 14).

6 Results

6.1 AlphaFold2 Redirects Experimental Effort

Table 3 reports the protein–year panel baseline. The model predicts that AF2 may reduce the value of experiments whose primary contribution is baseline coordinate information while lowering the marginal cost of experiments that build on that prior. These two channels work in opposite directions: substitution reduces experimental activity on well-predicted proteins, while cost reduction and frontier reweighting enable deeper and broader work on proteins where predictions are informative but incomplete.

Patterns in the data are consistent with both channels. Proteins with higher AF2 coverage experience a significantly larger decline in the probability of receiving any PDB deposition after 2021 (-0.003^{***}, a -44% decline relative to the pre-period mean; Col 1), consistent with substitution away from experimental work on well-predicted proteins. Yet the total number of depositions per protein increases by approximately 50% ($+0.404^{***}$; Col 2) and intensive depositions increase by approximately 67% ($+0.513^{***}$; Col 3). These patterns are consistent with a reallocation rather than a uniform expansion of effort. AF2 reduces activity on a broad set of proteins while concentrating and deepening work on a narrower set of active targets. One interpretation of these patterns is that, by providing a structural starting point, AF2 lowers the cost of designing follow-up experiments—molecular replacement in crystallography, model-guided construct design, targeted cryo-EM—making it worthwhile to pursue multiple depositions on proteins where predictions

are useful but incomplete.

Cumulative PDB coverage declines more for high-AF-coverage proteins (-0.004^{***}; Col 4). The concentration of depositions on fewer proteins does not compensate for the broad withdrawal of attention from the rest: average structural knowledge declines. The frontier-distance outcome (Col 5) reveals where new exploration occurs. First-ever depositions after AF2 are relatively more likely to target proteins closer to the existing knowledge frontier (-0.001^{***}), consistent with AF2 predictions serving as molecular replacement templates that lower the cost of resolving proteins whose close homologs have already been solved.

Table 4 reports the deposition-level baseline. Conditional on a deposition being made, the probability that it provides a first-ever structure increases with AF2 coverage after 2021 ($+0.070^{***}$; Col 1). This shift toward the extensive margin is consistent with the model's prediction that AF2, by reducing setup costs through structural priors, tilts the choice between exploring a new protein and revisiting a known one toward exploration, with the strongest effects for proteins closer to existing structural knowledge. Coverage increments Δe_d also increase: each first-ever deposition resolves a larger fraction of the protein ($+0.801^{***}$; Col 2), and follow-up depositions likewise cover more new residues ($+0.013^{***}$; Col 3). These patterns suggest increases in both the breadth and depth of individual experiments, consistent with predicted structures guiding experimentalists toward unresolved regions and enabling more efficient experimental design.

Results are robust to alternative treatment definitions using mean pLDDT and AF coverage at 50 and 90 confidence threshold (Tables 16–17, 20–21, and 18–19 in the Appendix), treatment samples (Tables 22–23 and 26–27), and post period (Tables 24–25).

6.2 Reallocation Along the Knowledge Frontier

The model highlights a tension. Because information spills over locally on the sequence-similarity tree, the marginal value of a new experiment is highest at the structural frontier, where no close homology template exists. AF2 introduces two opposing forces: it attenuates the value of experiments where predictions are highly informative, but lowers setup costs by providing structural priors. Whether exploration shifts toward the frontier therefore depends on which force dominates across proteins.

Table 5 examines how AF2's effect varies with four measures of frontier position Z. The triple interactions on total depositions show that AF2's effect on experimental activity grows with proximity to existing structural knowledge. The distance-based measures (nearest distance, gap) yield negative triple interactions: AF2 increases depositions more

for proteins closer to solved neighbours, consistent with AF2 predictions serving as molecular replacement templates that complement existing experimental infrastructure. The share-based measures (cluster PDB share, neighbour PDB share) yield positive triple interactions: AF2 also increases depositions in well-covered clusters, where it helps resolve the remaining unsolved members. On the intensive margin, the frontier gradient is weaker and less precisely estimated, indicating that frontier position shapes the overall level of experimental attention but not the depth of follow-up work conditional on a protein already being in the PDB.

Table 6 examines frontier interactions at the deposition level. The model predicts that frontier proteins should receive smaller coverage increments per deposition, because they are harder to resolve experimentally. The data confirm this: Δe_d for first-ever depositions declines with frontier distance (-0.129^{***} for the distance to the nearest solved structure, -0.146^{***} for gap). AF2 pushes exploration into new territory, but each deposition there covers less of the protein. The reallocation is about breadth of exploration, not depth at the frontier.

6.3 Composition of Intensive-Margin Depositions

The model distinguishes two channels through which AF2 affects intensive-margin composition. On one hand, AF2 reduces the marginal value of static-structure experiments—completeness (R) and refinement (Q)—by providing baseline coordinate information that these experiments would otherwise generate. On the other hand, AF2 reduces the cost of these same experiments by supplying a structural template that guides experimental design. The net effect on composition is theoretically ambiguous: if the cost channel dominates, AF2 steers follow-up work toward more residue extension, not less.

Table 7 tests this and shows that the cost channel dominates. AF2 steers follow-up work toward extending residue coverage into previously unresolved regions of known proteins: by providing a structural template, it makes it cheaper to design experiments that target specific uncovered residues. Higher AF2 coverage significantly increases the probability of a completeness deposition ($+0.138^{***}$, a 124% change relative to the pre-period rate). Depositions providing mechanistic insights or refinement are unaffected. The decline in corroborating depositions is consistent with AF2 reducing the need for independent structural validation—the activity whose informational value AF2 most directly substitutes.

6.4 Who Responds to AF2?

The preceding sections establish that AF2 reallocates experimental effort and reshapes the composition of follow-up work. The model generates predictions about which laboratories and experimental methods should respond most strongly. This section tests these predictions across three dimensions: institution type, laboratory experience, and experimental method.

6.4.1 Institutional heterogeneity

The model predicts that academic laboratories, with broader research objectives and smaller effective pipeline stocks, should shift more toward exploration than industry labs. Table 8 shows that company teams exhibit no significant differential response to AF2 on any outcome. The AF2 effect on structural biology operates primarily through academic laboratories, consistent with institutions with broader scientific objectives being more responsive to a general-purpose informational shock.

6.4.2 Incumbent-lab vs. entrant-lab depositions

The model predicts that AF2 should attract new laboratories into a protein's research space by lowering entry costs: entrant labs, which must pay the full setup cost, benefit most from the structural prior that AF2 provides. Table 9 shows that labs with more PDB deposition, a proxy for experience and a continuous measure of incumbency, do not represent a statistically significant share of depositions.[16] The concentration effect from Table 3 is driven primarily by incumbent labs increasing their output on proteins they already study.

6.4.3 Method heterogeneity

The model predicts that AF2's impact differs across experimental methods. X-ray crystallography benefits from AF-predicted structures as molecular replacement templates, lowering the barrier to phasing. Cryo-EM gains mainly through variable-cost reductions—better particle picking, map interpretation—on both margins. NMR receives minimal direct benefit from AF templates. allows the AF2 effect to differ by method, with X-ray as the omitted category.

These patterns partially confirm the model's predictions: in Table 10, AF2 benefits cryo-EM most strongly on the extensive margin (consistent with variable-cost reductions

[16] Table 12 shows that AF2 decreased the share of entrants among intensive-margin depositors.

enabling first-time exploration of new targets), while X-ray benefits are concentrated in coverage depth (consistent with AF templates improving molecular replacement for more complete structures).

6.5 Dynamic Spillovers

Extensive entry today widens the support of future intensive work: each new first-ever structure enters a protein into the pipeline stock, creating opportunities for auxiliary experiments—conformational-state and ligand-binding studies—on nearby proteins. We test this dynamic spillover in Table 11 using two measures of neighbourhood structural knowledge.

First, we measure the cross-sectional average AF2 coverage of neighboring clusters on the sequence-similarity tree. This variable is pre-determined by the biology of the neighborhood—evolutionary conservation, MSA depth—and is not affected by post-treatment experimental choices. It captures treatment-intensity amplification: proteins whose neighbors also received informative AF2 predictions may experience stronger indirect effects. The effect post AF2 exhibits the same substitution-versus-concentration pattern as the baseline and mirrors the own-protein AF2 effect. Higher neighborhood AF2 coverage reduces the probability of any intensive deposition (Col 1) but increases the count conditional on activity (Col 2). The informational shock attenuates the value of some follow-up experiments while enabling deeper work where it occurs.

Second, we use the time-varying share of neighboring clusters with any PDB structure. This variable evolves within protein as new extensive entries accumulate, capturing pipeline accumulation—the hypothesis that each new first-ever structure creates experimental opportunities (templates, demonstrated feasibility) for nearby proteins. However, because neighbour entries are themselves outcomes of AF2 treatment, this measure is endogenous. Its coefficient reflects a conditional association, not a causal effect. We find that a higher share of solved neighbours is associated with more intensive depositions (Cols 3 and 4).

When both measures enter jointly (Cols 5–6), each retains its sign and significance, indicating that they capture distinct channels. The static coefficient reflects indirect treatment intensity—the substitution and cost-reduction effects of AF2 predictions propagating through the neighbourhood. The dynamic coefficient reflects pipeline accumulation—new experimental entries creating opportunities for nearby proteins. We emphasise that the dynamic measure is endogenous; its coefficient cannot be given a causal interpretation.

Table 28 in the Appendix decomposes the dynamic association by deposition type.

The effect operates entirely through conformational-state and ligand-binding experiments (SL), with no significant effect on residue extension (R). This pattern is consistent with the mechanism in Proposition 6, whereby a neighbour's first structure does not help extend a protein's own residue coverage, but it reveals conformational states and binding sites that raise the value of auxiliary experiments on nearby proteins—the knowledge channel that AF2 does not substitute.

7 Conclusion

This paper studies how a major AI advance reshapes the direction of scientific effort. We model protein research as search on a knowledge landscape with local spillovers and setup frictions, and we treat AF2 as a universal but heterogeneous information shock. The model predicts reallocation rather than mechanical expansion: AF2 substitutes for experiments that primarily recover baseline static structure, while reducing the cost of complementary follow-up work that AI does not directly replace.

Using protein-level variation in AF2 informativeness, we find evidence consistent with this reallocation pattern in the data. Proteins with higher AF2 coverage experience a substantially larger decline in the probability of experimental attention, yet effort concentrates rather than disappears: among proteins that remain active, experimental output and follow-up work increase sharply. The composition of new experiments shifts toward first-ever structure determinations, and among follow-up work, toward resolving previously uncharacterised regions and away from corroboration. The response is concentrated where AF2 predictions complement existing experimental infrastructure and is driven primarily by academic laboratories. Dynamic results further show that new structures generate local spillovers that support subsequent conformational and ligand-binding experimentation. Overall, AI in this setting acts less as a pure labour-saving substitute and more as a technology that reorganises what science gets done, by whom, and where.

References

Acemoglu, Daron, "The Simple Macroeconomics of AI," Working Paper 32487, National Bureau of Economic Research May 2024.

_ **and Pascual Restrepo**, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Aghion, Philippe, Benjamin F Jones, and Charles I Jones, "Artificial Intelligence and Economic Growth," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 237–282.

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.

_ **,** _ **,** **and** _ , "AI in Science," Working Paper 34953, National Bureau of Economic Research March 2026.

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.

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.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?," *American Economic Review*, 2020, *110* (4), 1104–1144.

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.

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.

Cavalli, Gabriel, "Responding to Advances in AI: The Impact of AlphaFold on the Organization of Academic Labs," 2024. Working Paper, Rotman School of Management, University of Toronto.

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

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," 2025. NBER Working Paper.

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

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.

Jones, Benjamin F, "Artificial Intelligence in Research and Development," Working Paper 34312, National Bureau of Economic Research 2025.

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*.

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

Mullainathan, Sendhil and Ashesh Rambachan, "Science in the Age of Algorithms," in Ajay K Agrawal, Erik Brynjolfsson, and Anton Korinek, eds., *The Economics of Transformative Artificial Intelligence*, University of Chicago Press, 2025.

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.

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.

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.

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.

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

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

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.

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

Tables and Figures

Table 1: Notation guide

Symbol(s)	Meaning
$\mathcal{P}, \mathcal{T}, D_{\mathcal{T}}, \mu$	Protein universe, sequence-similarity tree, tree distance, and density measure on the tree.
$x, y, z, p, t, \ell, m, a$	Protein, time, laboratory, method, and action indices; $a \in \{R, Q, S, L\}$ denotes residue extension, refinement, conformational-state, and ligand-bound actions.
$\mathcal{F}_t, \mathcal{S}_t, \mathcal{L}_{\ell t}, \mathcal{A}(x,t), J_t(x), I_{\ell x t}$	Public static-information stock, full model state, lab pipeline stock, feasible actions, public-solved indicator, and lab-incumbency indicator.
$Y(x), s_i, \nu_i, \mathbf{q}_x$	Latent dominant static structure index, experimental signal, signal-noise variance, and quality state vector $\mathbf{q}_x = (e_x, \rho_x, n_x^S, n_x^L)$.
$e_x, \rho_x, n_x^S, n_x^L$	Residue coverage, resolution-quality index, number of conformational states, and number of ligand-bound contexts.
u_f, u_0, u_S, u_L	Scaffold posterior variance, total static uncertainty, and conformational / ligand uncertainty states.
$h_E, h_Q, h_S, h_L, \kappa_E, \kappa_Q, \kappa_S, \kappa_L$	Diminishing-return functions and their scale parameters.
G_R, G_Q, G_S, G_L	Marginal uncertainty reductions from additional progress in each action dimension.
$v_\ell(\mathcal{F}_t), w_\ell(x)$	Laboratory objective and protein-specific weight function.
$\alpha_0, \alpha_S, \alpha_L, q_0, q_S, q_L$	Value-function weights and in-play thresholds.
$\phi_m, g_{a,m}(\mathbf{q}_x), \eta_m$	Setup cost, variable cost component, and method scale.
$F_{xt}, \chi(F_{xt})$	Frontier position and frontier-difficulty multiplier in the cost function.
$c_{\ell t}(x,a,m), B_{\ell t}(x,a,m)$	Cost and gross value of an experiment.
$\Pi_{\ell t}(x,a,m)$	Net payoff ($B - c$).
$\hat{y}^{AF}(x), \omega(x), A_x, u_f^{AF}(x), u_0^{AF}(x)$	AF2 signal, AF2 noise variance, empirical AF informativeness, post-AF scaffold uncertainty, and post-AF total static uncertainty.
$d\Delta_f(y;x,a), \lambda(y), \bar{\lambda}_{\ell t}(x,a)$	Local scaffold spillover reduction and local / average attenuation factors under AF2.
$\delta_m^\phi(A_x), \delta_{a,m}^g(A_x), \Delta_{\ell t}^E(x,m)$	AF-induced setup / variable cost savings and net AF shock on the public-extensive margin.
$i, Y_i, \text{AF cov}_p, \text{Post}_t, \alpha_p, \lambda_t, \epsilon_i$	Observation index / outcome, treatment and post indicator, fixed effects, and error term in baseline regressions.
$e_{pt}, \Delta e_d, Z$	Cumulative coverage, deposition-level coverage increment, and frontier proxy in heterogeneity analyses.

Table 2: Descriptive Statistics

Variable	Mean	Std. Dev.	Min	Max
Protein panel ($N = 552,566$ proteins $\times$ 9 years)				
Nb. depositions	0.042	1.037	0.000	375.000
Nb. extensive	0.002	0.049	0.000	1.000
Nb. intensive	0.040	1.033	0.000	375.000
$e_{p,t}$	0.040	0.182	0.000	1.000
Dist. $\times$ Ext.	0.002	0.054	0.000	1.773
PDB depositions ($N = 348,143$ depositions)				
Share intensive	0.898	0.303	0.000	1.000
Share completeness (R)	0.109	0.312	0.000	1.000
Share mechanistic insight (SL)	0.524	0.499	0.000	1.000
Share refinement (Q)	0.012	0.111	0.000	1.000
Share corroborating (Corr.)	0.252	0.434	0.000	1.000
Δe_d	0.077	0.235	0.000	1.000
$\Delta e_d \mid$ extensive	0.680	0.333	0.001	1.000
$\Delta e_d \mid$ intensive	0.008	0.050	0.000	0.984
Share academic institution	0.615	0.375	0.000	1.000
Share company institution	0.044	0.184	0.000	1.000
Share government institution	0.027	0.116	0.000	1.000
Share of incumbent lab	0.689	0.463	0.000	1.000
Structural landscape ($N = 552,566$ proteins, as of 2016)				
$e_{p,2016}$	0.031	0.161	0.000	1.000
Share isolated	0.156	0.363	0.000	1.000
Dist. to nearest solved	0.342	0.366	0.000	1.346
Gap	0.634	0.165	0.000	1.199
Share PDB in cluster	0.043	0.125	0.000	1.000
Share PDB in neighbors	0.139	0.141	0.000	1.000
AF2 prediction ($N = 552,566$ proteins)				
AF2 coverage (pLDDT ≥ 90)	0.693	0.274	0.000	1.000
AF2 coverage (pLDDT ≥ 70)	0.868	0.190	0.000	1.000
AF2 coverage (pLDDT ≥ 50)	0.933	0.129	0.000	1.000
Mean pLDDT	87.673	10.511	26.181	98.754
Mean PAE	9.175	6.510	1.583	29.061

Notes. Panel A reports statistics for the protein $\times$ year panel covering 2017–2025. Panel B reports statistics at the level of individual PDB depositions. Panel C describes the structural knowledge frontier as a protein-level cross-section measured in 2016, before the treatment period. Panel D summarises AF2 predictions at the protein level.

Table 3: Protein-year panel baseline

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. × Ext.
Post × AF cov	-0.0035***	0.4043***	0.5133***	-0.0042***	-0.0013***
	(0.0004)	(0.1138)	(0.1487)	(0.0004)	(0.0003)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	171,252	72,342	4,774,455	4,774,455
R^2 / $\tilde{R}^2$	0.420	0.562	0.615	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.033	0.619	0.036	0.002

Notes. This table reports estimates of the effect of AF2 coverage on five protein-year outcomes. The dependent variable in column (1) is an indicator for any PDB deposition in year t; column (2) reports Poisson pseudo-ML estimates on the total number of depositions; column (3) reports Poisson estimates on the number of intensive (follow-up) depositions; column (4) reports cumulative PDB coverage e_{pt}; and column (5) reports frontier distance interacted with an indicator for first-ever PDB entry. The treatment variable is $\text{Post}_t \times \text{AFcov}_p$, where AFcov is the share of residues with pLDDT ≥ 70. Columns (1)–(2) and (4)–(5) use the full sample of SwissProt proteins created by 2016 (balanced panel, 2017–2025); column (3) restricts to proteins with at least one PDB structure before 2017. All columns include protein and year fixed effects. Columns (1), (4)–(5) are estimated by OLS; columns (2)–(3) by Poisson pseudo-ML, where coefficients are semi-elasticities. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 4: Deposition-level baseline

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post $\times$ AF cov	0.0703***	0.8009***	0.0135***
	(0.0114)	(0.0179)	(0.0020)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.272	0.396	0.257
$\bar{Y}_{p,t\leq 2021}$	0.068	0.674	0.007

Notes. This table reports OLS estimates at the deposition level for SwissProt proteins over 2017–2025. The dependent variable in column (1) is an indicator for whether the deposition provides the first-ever PDB structure for protein p; columns (2) and (3) report the coverage increment Δe_d conditional on extensive and intensive depositions, respectively. The treatment variable is $\text{Post}_t \times \text{AFcov}_p$. Columns (1) and (3) include protein and year fixed effects. Column (2) includes organism and year fixed effects with protein-level controls (log length, Pfam count, membrane indicator, disease association, and pharma flags). Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5: Protein-year panel: frontier interactions

$Z =$	(1) Dist.	(2) Gap	(3) Sh. PDB	(4) Sh. PDB nb.
Outcome: # PDB depositions				
Post × AF cov	0.628***	0.732**	-0.044	0.268*
	(0.188)	(0.287)	(0.118)	(0.141)
Post × Z	0.661***	0.308***	-1.729***	-1.242**
	(0.182)	(0.109)	(0.308)	(0.497)
Post × AF cov × Z	-0.531***	-0.185	1.287***	1.159**
	(0.205)	(0.129)	(0.353)	(0.555)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	171,252	171,252	171,252	171,252
$\tilde{R}^2$	0.563	0.563	0.564	0.562
Outcome: # Intensive depositions				
Post × AF cov	0.519**	0.736**	-0.069	0.360*
	(0.232)	(0.328)	(0.187)	(0.200)
Post × Z	0.075	0.210	-1.543***	-0.891
	(0.349)	(0.151)	(0.364)	(0.632)
Post × AF cov × Z	-0.061	-0.171	1.311***	0.846
	(0.392)	(0.179)	(0.408)	(0.708)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	72,342	72,342	72,342	72,342
$\tilde{R}^2$	0.615	0.616	0.616	0.616

Note: Poisson pseudo-ML estimates with standard errors clustered at the protein level (coefficients are semi-elasticities). Each column interacts the treatment (Post × AF cov) with a different frontier measure Z measured as of 2016: (1) tree distance to nearest solved; (2) gap; (3) share of cluster with PDB; (4) share of neighbouring clusters with PDB. All regressions control for Post × Isolated and protein and year fixed effects. The sample for total depositions covers all SwissProt proteins in a balanced panel; the intensive-margin sample restricts to proteins with at least one PDB structure before 2017. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 6: Deposition-level: frontier interactions and coverage depth

$Z =$	(1) Dist.	(2) Gap	(3) Sh. PDB	(4) Sh. PDB nb.
Outcome: Extensive				
Post × AF cov	0.011	0.012	0.111***	0.032**
	(0.009)	(0.018)	(0.019)	(0.013)
Post × Z	-0.149***	-0.056***	0.234***	0.019
	(0.032)	(0.012)	(0.031)	(0.038)
Post × AF cov × Z	0.114***	0.023	-0.177***	0.001
	(0.034)	(0.014)	(0.035)	(0.042)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	196,720	196,720	196,720	196,720
R^2_w	0.015	0.017	0.015	0.013
Outcome: Δe_d | Extensive				
Post × AF cov	0.946***	1.169***	0.784***	0.766***
	(0.034)	(0.065)	(0.019)	(0.023)
Post × Z	0.150***	0.150***	-0.555***	-0.443***
	(0.030)	(0.024)	(0.116)	(0.093)
Post × AF cov × Z	-0.129***	-0.146***	0.713***	0.405***
	(0.033)	(0.027)	(0.133)	(0.101)
Organism FE	X	X	X	X
Year FE	X	X	X	X
Protein features	X	X	X	X
N	10,575	10,575	10,575	10,575
R^2_w	0.227	0.229	0.228	0.228
Outcome: Δe_d | Intensive				
Post × AF cov	0.009***	0.007*	0.015***	0.012***
	(0.003)	(0.004)	(0.003)	(0.003)
Post × Z	-0.004	-0.004*	0.010	0.002
	(0.005)	(0.002)	(0.006)	(0.009)
Post × AF cov × Z	0.005	0.003	-0.013*	0.000
	(0.005)	(0.003)	(0.007)	(0.009)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	187,613	187,613	187,613	187,613
R^2_w	0.001	0.001	0.001	0.001

Notes. This table reports OLS estimates at the deposition level for SwissProt proteins over 2017-2025 in Panel A, restricted to extensive and intensive margin depositions in Panel B and C, respectively. Each column interacts the treatment (Post × AF cov) with a different frontier measure Z measured as of 2016: (1) tree distance to nearest solved protein; (2) gap; (3) share of cluster members with PDB; (4) share of neighboring clusters with PDB. All regressions control for Post × Isolated, along with protein and year fixed effects in Panel A and C, and organism and year fixed effects with protein-level controls (log length, Pfam count, membrane indicator, disease association, and pharma flag) in Panel B. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 7: Action-type decomposition on the intensive margin

	(1) Completeness (R)	(2) Mechanistic (SL)	(3) Refinement (Q)	(4) Corrob.
Post × AF cov	0.1376***	-0.0446	-0.0010	-0.0920***
	(0.0174)	(0.0283)	(0.0034)	(0.0270)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	187,613	187,613	187,613	187,613
R^2	0.404	0.356	0.249	0.333
$\bar{Y}_{t\leq 2021}$	0.111	0.566	0.009	0.314

Notes. This table reports OLS estimates of the effect of AF2 coverage on the composition of intensive-margin depositions. Each column regresses a mutually exclusive deposition-category indicator on Post × AFcov$_p$: completeness (R, new residues), mechanistic insight (SL, new conformation or ligand), refinement (Q, improved resolution), and corroborating (Corr., independent validation). Categories are assigned with priority $R \succ SL \succ Q \succ$ Corr. (see Appendix A.3.2). The sample consists of intensive-margin PDB depositions over 2017–2025. All columns include protein and year fixed effects. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8: Institutional heterogeneity: university vs. company teams

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov	0.0790***	0.8108***	0.0149***
	(0.0121)	(0.0203)	(0.0021)
Post × AF cov × Any company	-0.0059	0.0364	-0.0013
	(0.0123)	(0.0681)	(0.0025)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	167,664	9,298	159,542
R^2	0.272	0.390	0.264
$\bar{S}_{t\leq 2021}$	0.078	0.078	0.078
$\bar{Y}_{t\leq 2021}$	0.068	0.670	0.007

Notes. This table reports OLS estimates of institutional heterogeneity in the AF2 effect. The dependent variable in column (1) is an indicator for first-ever PDB entry; columns (2) and (3) report the coverage increment Δe_d conditional on extensive and intensive depositions, respectively. The interaction term $S_d = \mathbf{1}$[any company on the author team] is based on OpenAlex-matched institutional affiliations. The sample consists of PDB depositions with non-missing institution metadata over 2017–2025. Column (1) includes protein and year fixed effects; column (2) includes organism and year fixed effects with protein-level controls (log length, Pfam count, membrane indicator, disease association, and pharma flags); column (3) includes protein and year fixed effects. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 9: Deposition-level baseline with lab experience interaction

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.	(4) Dist. × Ext.
Post × AF cov	0.0677***	0.7791***	0.0130***	0.0804***
	(0.0124)	(0.0218)	(0.0021)	(0.0129)
Post × AF cov × Lab exp.	0.0008	0.0196**	0.0003	0.0009
	(0.0017)	(0.0085)	(0.0004)	(0.0019)
Organism FE		X		
Protein FE	X		X	X
Year FE	X	X	X	X
Protein features		X		
N	196,720	10,575	187,613	196,720
R^2	0.272	0.396	0.257	0.275

Notes. This table reports OLS estimates of how laboratory experience moderates the AF2 effect, adding a triple interaction with log lab experience (log of one plus the number of PDB depositions by the last author before 2017). The dependent variable in column (1) is an indicator for first-ever PDB entry; columns (2) and (3) report the coverage increment Δe_d conditional on extensive and intensive depositions; column (4) reports frontier distance interacted with first-ever PDB entry. The sample consists of PDB depositions with non-missing last author over 2017–2025. Columns (1) and (3)–(4) include protein and year fixed effects; column (2) includes organism and year fixed effects with protein-level controls. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 10: Method heterogeneity: triple interactions

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov (X-ray)	0.0357***	0.8813***	0.0088***
	(0.0108)	(0.0232)	(0.0019)
Post × AF cov × Cryo-EM	0.0749***	-0.1405***	0.0105***
	(0.0169)	(0.0312)	(0.0031)
Post × AF cov × NMR	-0.0360	-0.5718***	-0.0421**
	(0.0695)	(0.1142)	(0.0212)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.273	0.416	0.257
$\bar{Y}_{t\leq 2021}$	0.068	0.674	0.007

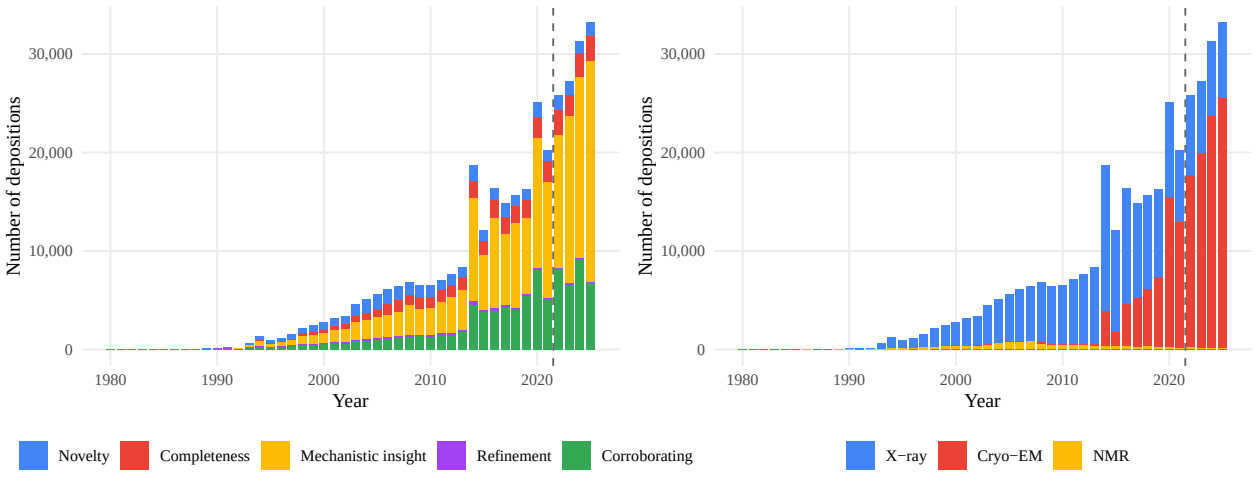
Notes. This table reports OLS estimates of method-specific AF2 effects, with X-ray crystallography as the omitted category. The coefficient on Post × AF cov captures the X-ray treatment effect; the interactions with cryo-EM and NMR measure differential effects relative to X-ray. The dependent variable in column (1) is an indicator for first-ever PDB entry; columns (2) and (3) report the coverage increment Δe_d conditional on extensive and intensive depositions, respectively. The sample consists of all PDB depositions for SwissProt proteins over 2017–2025. Column (1) includes protein and year fixed effects; column (2) includes organism and year fixed effects with protein-level controls; column (3) includes protein and year fixed effects. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 11: Dynamic network spillover on the intensive margin

	(1) Pr(Int.)	(2) # Int.	(3) Pr(Int.)	(4) # Int.	(5) Pr(Int.)	(6) # Int.
Post × AF cov	-0.0163***	0.3623***	-0.0207***	0.4653***	-0.0160**	0.3824***
	(0.0063)	(0.0870)	(0.0061)	(0.1438)	(0.0063)	(0.0873)
Post × Neighbor AF cov	-0.0074**	0.1975***			-0.0081***	0.1545***
	(0.0030)	(0.0458)			(0.0030)	(0.0449)
Share neighbors PDB			0.0478**	1.088***	0.0510**	2.975***
			(0.0200)	(0.2014)	(0.0201)	(0.3642)
Protein FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
N	203,202	203,202	203,202	72,342	203,202	203,202
R^2 / $\tilde{R}^2$	0.484	0.562	0.484	0.616	0.484	0.563
$\bar{Y}_{t\leq 2021}$	0.115	0.619	0.115	0.619	0.115	0.619

Notes. This table compares two measures of network spillovers on the intensive margin. Columns (1)–(2) include only the static measure (Post × neighbor AF coverage), which is pre-determined and exogenous to post-treatment experimental choices. Columns (3)–(4) include only the dynamic measure (time-varying share of neighboring clusters with any PDB structure), which is endogenous because neighbor entries are themselves AF2 outcomes. Columns (5)–(6) include both measures jointly. The dependent variable in odd columns is an indicator for any intensive deposition; even columns report the count of intensive depositions. The sample consists of proteins with at least one PDB structure before 2017 in a balanced panel over 2017–2025. All columns include protein and year fixed effects. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

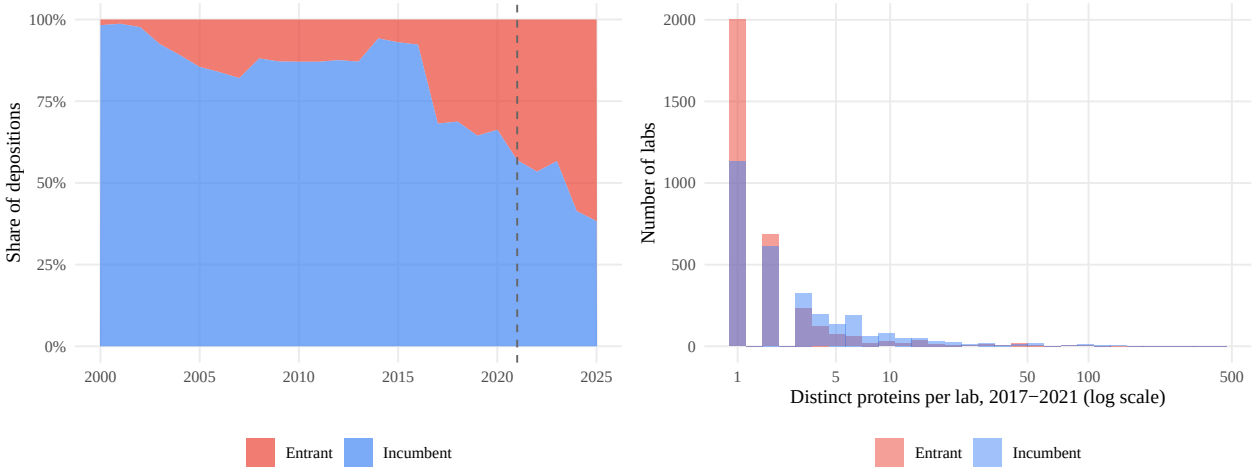
Figure 1: Annual depositions of experimental structure by category and method



Notes. The left panel plots the annual number of PDB depositions by category: novelty depositions provide the first-ever structure for a protein, completeness depositions resolve previously unobserved residues,

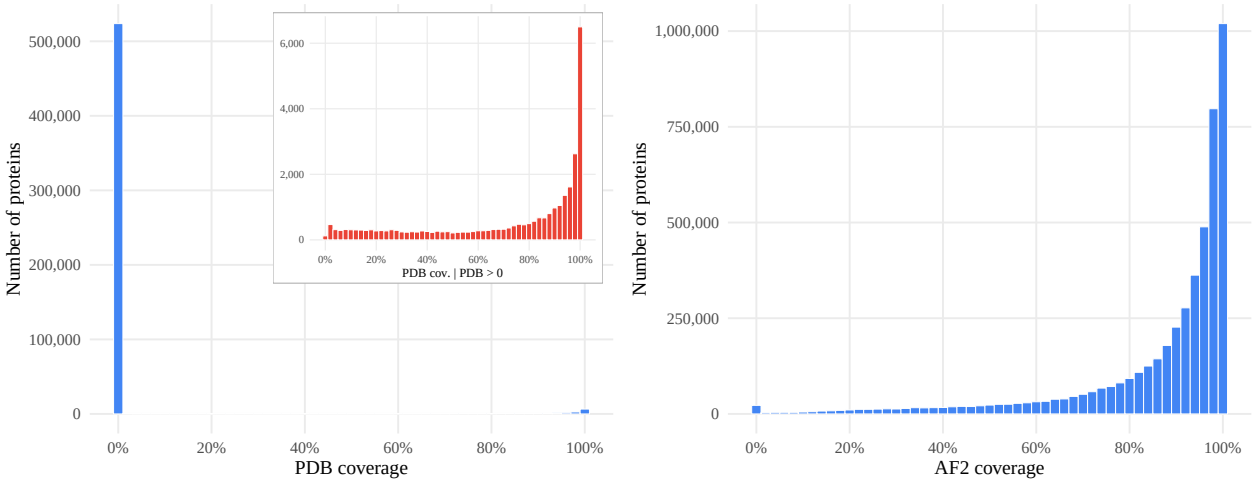
mechanistic insight depositions capture a new conformation or ligand, refinement depositions improve resolution, and corroborating depositions provide independent validation of a known structure. The right panel plots the annual number of depositions by experimental method. The dashed line marks the AlphaFold release in July 2021.

Figure 2: Lab composition of PDB depositions



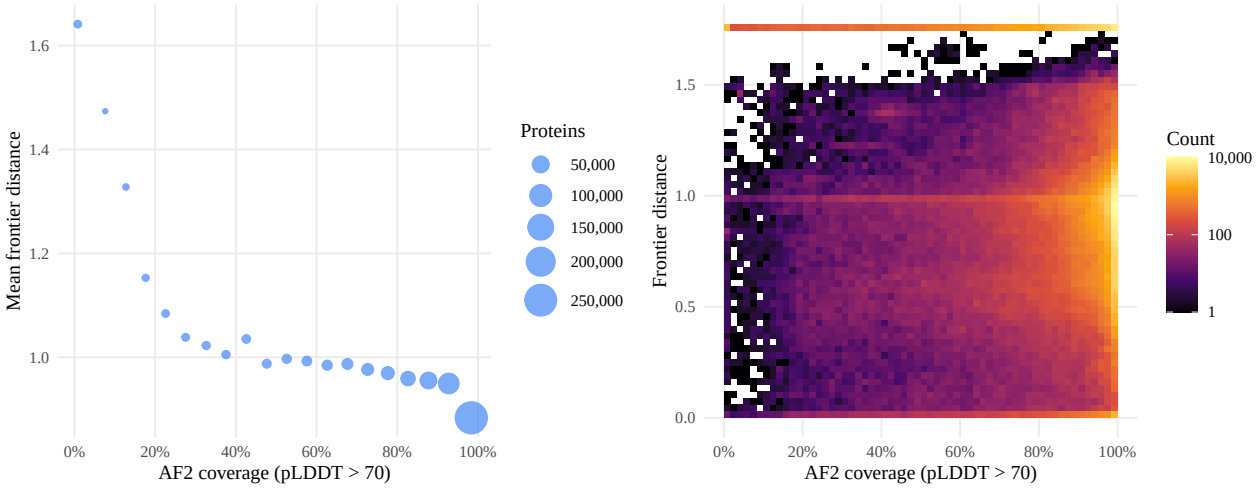
Notes. An incumbent lab is defined as one whose last (senior) author has at least one PDB deposition before 2017; an entrant lab has no depositions before 2017. The left panel shows the share of annual PDB depositions by lab type. The right panel shows the distribution of distinct proteins per lab during 2017–2021, the pre-AF2 window, computed over the same period for both types. The dashed line marks the AlphaFold release in July 2021.

Figure 3: Distribution of experimental and AF2 coverage



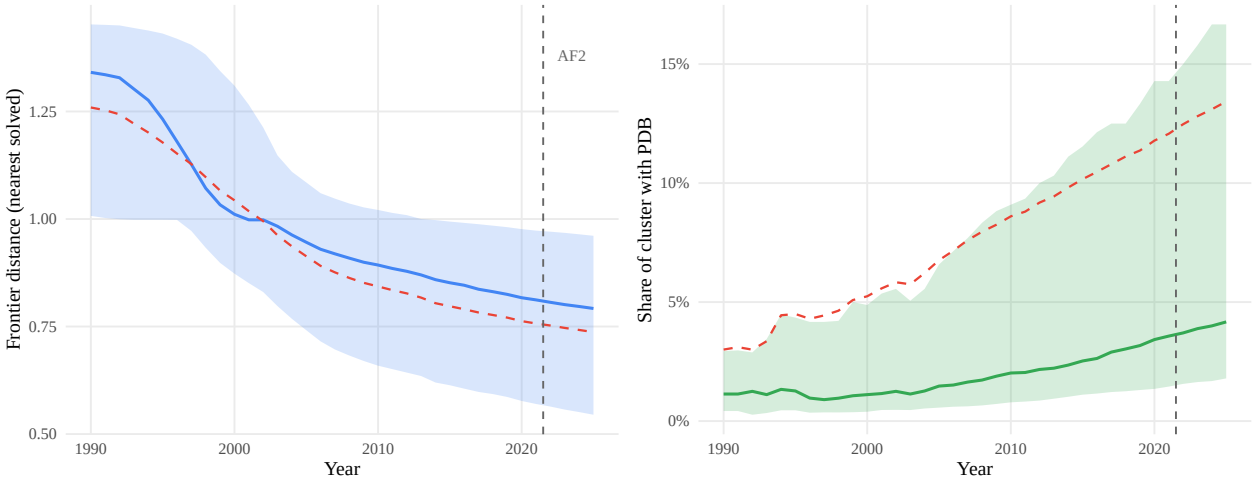
Notes. The left panel shows the distribution of experimental structural coverage ($e_{p,2020}$) as of 2020, and the inset shows the same distribution conditional on the protein having at least some experimental coverage ($e_{p,2020} > 0$). The right panel shows the distribution of AlphaFold2 coverage, defined as the share of residues with pLDDT ≥ 70.

Figure 4: AF2 coverage and frontier distance



Notes. The left panel is a binscatter of AF2 coverage (the share of residues with pLDDT ≥ 70) against mean frontier distance, measured as the tree-path distance to the nearest solved protein as of 2016, with point size proportional to the number of proteins in each bin. The right panel shows a two-dimensional histogram of the joint distribution on a log scale.

Figure 5: Frontier contraction over time



Notes. The left panel shows the distribution of frontier distance (tree-path distance to the nearest solved protein) over time, where the solid line is the median, the dashed red line is the mean, and the shaded band spans the interquartile range. The right panel shows the share of cluster members with PDB coverage over the same period. The frontier contracts steadily as depositions accumulate. The dashed vertical line marks the AlphaFold2 release in July 2021.

Appendix

A.1 Proofs for Section 3

Proof of Lemma 1. Let

$$\alpha = \frac{d_r}{X}, \qquad \beta = \frac{d_l}{X} = 1 - \alpha.$$

Because x lies in the interior of a branchless edge $e = (v_l, v_r)$ of the augmented tree $\bar{\mathcal{T}}$, removing the open segment (v_l, v_r) disconnects the tree into two components that meet the rest of the process only through the boundary values $Y(v_l)$ and $Y(v_r)$. The Gaussian Markov property on trees therefore, implies

$$Y(x) \perp\!\!\!\perp \mathcal{G} \mid Y(v_l), Y(v_r).$$

Conditional on $Y(v_l)$ and $Y(v_r)$, the restriction of Brownian motion to the edge e is an ordinary Brownian bridge of length X. Hence

$$Y(x) = \alpha Y(v_l) + \beta Y(v_r) + \zeta,$$

where

$$\zeta \mid Y(v_l), Y(v_r), \mathcal{G} \sim \mathcal{N}\left(0, \frac{d_l d_r}{X}\right)$$

and ζ is independent of $(Y(v_l), Y(v_r), \mathcal{G})$.

Taking conditional expectations yields

$$\mathbb{E}[Y(x) \mid \mathcal{G}] = \alpha \mathbb{E}[Y(v_l) \mid \mathcal{G}] + \beta \mathbb{E}[Y(v_r) \mid \mathcal{G}].$$

Applying the law of total variance conditional on $\mathcal{G}$ gives

$$\begin{aligned}
\mathrm{Var}(Y(x) \mid \mathcal{G}) &= \mathbb{E}[\mathrm{Var}(Y(x) \mid Y(v_l), Y(v_r), \mathcal{G}) \mid \mathcal{G}] \\
&\quad + \mathrm{Var}(\mathbb{E}[Y(x) \mid Y(v_l), Y(v_r), \mathcal{G}] \mid \mathcal{G}) \\
&= \frac{d_l d_r}{X} + \mathrm{Var}(\alpha Y(v_l) + \beta Y(v_r) \mid \mathcal{G}) \\
&= \frac{d_l d_r}{X} + \alpha^2 P_l + \beta^2 P_r + 2\alpha\beta C_{lr}.
\end{aligned}$$

Substituting $\alpha = d_r / X$ and $\beta = d_l / X$ yields (3). □

Proof of Proposition 1. For part (i), define

$$G_R(e) = \kappa_E \big[h_E(e) - h_E(e + \Delta_E)\big].$$

Take $e_1 < e_2$ in $[0, 1 - \Delta_E]$. Convexity of h_E implies that forward secant slopes are weakly increasing, so

$$\frac{h_E(e_1 + \Delta_E) - h_E(e_1)}{\Delta_E} \leq \frac{h_E(e_2 + \Delta_E) - h_E(e_2)}{\Delta_E}.$$

Multiplying by $\Delta_E > 0$ and rearranging gives

$$h_E(e_1) - h_E(e_1 + \Delta_E) \geq h_E(e_2) - h_E(e_2 + \Delta_E),$$

hence $G_R(e)$ is weakly decreasing in e.

For part (ii), define

$$G_Q(\rho) = \kappa_Q \big[h_Q(\rho) - h_Q(\rho + \Delta_Q)\big].$$

Applying exactly the same secant-slope argument to the convex function h_Q yields

$$h_Q(\rho_1) - h_Q(\rho_1 + \Delta_Q) \geq h_Q(\rho_2) - h_Q(\rho_2 + \Delta_Q) \qquad \text{whenever } \rho_1 < \rho_2,$$

so $G_Q(\rho)$ is weakly decreasing in ρ.

For part (iii), define

$$G_S(n) = \kappa_S \big[h_S(n) - h_S(n+1)\big].$$

Discrete convexity of h_S means precisely that the one-step decrements weakly shrink:

$$h_S(n) - h_S(n+1) \geq h_S(n+1) - h_S(n+2),$$

so $G_S(n)$ is weakly decreasing in n. The same argument applies to

$$G_L(n) = \kappa_L \big[h_L(n) - h_L(n+1)\big]$$

because h_L is also discrete-convex. □

Proof of Proposition 2. Parts (i), (ii), and (iii) follow directly from the feasible action set (8), the definition of the pipeline stock $\mathcal{L}_{\ell t}$, and the payoff formula

$$\Pi_{\ell t}(x, a, m) = B_{\ell t}(x, a, m) - c_{\ell t}(x, a, m).$$

For (10), hold fixed a publicly solved protein x, a feasible action a, and a method m. The gross value term $B_{\ell t}(x, a, m)$ is the same in the two counterfactual comparisons; the only difference is whether the laboratory has already built protein-specific capital for x. Hence

$$\begin{aligned}
&\Pi_{\ell t}(x, a, m \mid I_{\ell x t} = 1) - \Pi_{\ell t}(x, a, m \mid I_{\ell x t} = 0) \\
&\quad = \big[B_{\ell t}(x, a, m) - c_{\ell t}(x, a, m \mid I_{\ell x t} = 1)\big] - \big[B_{\ell t}(x, a, m) - c_{\ell t}(x, a, m \mid I_{\ell x t} = 0)\big] \\
&\quad = c_{\ell t}(x, a, m \mid I_{\ell x t} = 0) - c_{\ell t}(x, a, m \mid I_{\ell x t} = 1) \\
&\quad = \eta_m \phi_m \chi(F_{x t}) > 0.
\end{aligned}$$

This proves (10). □

Proof of Proposition 3. Fix $a \in \{R, Q\}$. By construction,

$$u_0(y) = u_f(y) + \kappa_E h_E(e_y) + \kappa_Q h_Q(\rho_y).$$

Because all affected proteins remain in the interior of the static in-play region, the derivative

of the static component of (11) with respect to the scaffold component $u_f(y)$ is

$$\frac{\partial}{\partial u_f(y)}\left[w_\ell(y)\alpha_0 \max\left\{1-\frac{u_0(y)}{q_0},0\right\}\right] = -\frac{\alpha_0}{q_0}w_\ell(y),$$

since $\partial u_0(y)/\partial u_f(y) = 1$. Therefore an infinitesimal reduction $d\Delta_f(y;x,a)$ in $u_f(y)$ raises value by

$$\frac{\alpha_0}{q_0}w_\ell(y)\,d\Delta_f(y;x,a),$$

and the pre-AF marginal value transmitted through the scaffold component is

$$dB_{\ell t}^{pre,f}(x,a) = \frac{\alpha_0}{q_0}\int_{\mathcal{T}} w_\ell(y)\,d\Delta_f(y;x,a)\,d\mu(y). \tag{23}$$

After AF arrives, the scaffold uncertainty at y becomes

$$u_f^{AF}(y) = \frac{u_f(y)\omega(y)}{u_f(y)+\omega(y)}.$$

Differentiating with respect to $u_f(y)$ yields

$$\frac{\partial u_f^{AF}(y)}{\partial u_f(y)} = \frac{\omega(y)^2}{\big(u_f(y)+\omega(y)\big)^2} = \lambda(y) \in (0,1).$$

Hence the same underlying infinitesimal reduction $d\Delta_f(y;x,a)$ in pre-AF scaffold uncertainty translates into an infinitesimal reduction

$$d\Delta_f^{AF}(y;x,a) = \lambda(y)\,d\Delta_f(y;x,a)$$

in post-AF scaffold uncertainty. Therefore the post-AF marginal value transmitted through u_f is

$$dB_{\ell t}^{AF,f}(x,a) = \frac{\alpha_0}{q_0}\int_{\mathcal{T}} w_\ell(y)\lambda(y)\,d\Delta_f(y;x,a)\,d\mu(y). \tag{24}$$

If the denominator in (15) is nonzero, dividing (24) by (23) gives the formula for $\bar{\lambda}_{\ell t}(x,a)$. Because $0<\lambda(y)<1$ pointwise, the weighted average $\bar{\lambda}_{\ell t}(x,a)$ also lies in $(0,1)$. If the denominator is zero, then both (23) and (24) are zero, so the claim is trivial.

If an R or Q experiment affects payoffs only through the scaffold component u_f, then $dB_{\ell t}^{pre,f}(x,a) = dB_{\ell t}^{pre}(x,a)$ and $dB_{\ell t}^{AF,f}(x,a) = dB_{\ell t}^{AF}(x,a)$, yielding the one-component special case.

For $a = S$, the maintained transition convention in Section 3.2 implies that the experiment changes only u_S and does not update $\mathcal{F}_t$ or u_f. Under the interior condition $u_S(x) < q_S$, the marginal value of an infinitesimal reduction in u_S is

$$\frac{\alpha_S}{q_S}w_\ell(x),$$

which is the same before and after AF. Hence

$$dB_{\ell t}^{AF}(x, S) = dB_{\ell t}^{pre}(x, S).$$

Exactly the same argument applies to $a = L$ because AF does not enter u_L and the baseline transition convention leaves u_f unchanged. □

Proof of Proposition 4. By definition,

$$D_m(x, y, a) = \Pi_{\ell t}^{E,AF}(x, R, m) - \Pi_{\ell t}^{I,AF}(y, a, m).$$

Holding benefit terms and other cost terms fixed, the public-extensive cost component is

$$\eta_m \Big[\phi_m - \delta_m^{\phi}(A_x) + \cdots\Big] \chi(F_{xt}),$$

because $I_{\ell xt} = 0$ by assumption, while the same-lab intensive option does not contain the setup-cost term because $I_{\ell yt} = 1$.

For part (i), differentiating $D_m(x, y, a)$ with respect to $\delta_m^{\phi}(A_x)$ yields

$$\frac{\partial D_m(x, y, a)}{\partial \delta_m^{\phi}(A_x)} = \eta_m \chi(F_{xt}) > 0.$$

Hence a larger setup-cost saving unambiguously raises the relative attractiveness of the public-extensive option when the target is new to the laboratory. If instead $I_{\ell xt} = 1$, the derivative is zero because the setup term is absent.

For part (ii), suppose a common variable-cost saving δ^g applies symmetrically to the relevant action on both margins. Then the public-extensive option receives cost reduction $\eta_m \delta^g \chi(F_{xt})$ and the same-lab intensive option receives cost reduction $\eta_m \delta^g \chi(F_{yt})$. Therefore

$$\frac{\partial D_m(x, y, a)}{\partial \delta^g} = \eta_m \chi(F_{xt}) - \eta_m \chi(F_{yt}) = \eta_m \big[\chi(F_{xt}) - \chi(F_{yt})\big],$$

whose sign is generally ambiguous. □

Proof of Proposition 5. Let

$$x_{pre}^{\star}(m) \in \arg \max_{x:\, J_t(x)=0} \Pi_{\ell t}^{E,pre}(x, R, m)$$

be the maximiser with the largest frontier index, and define $x_{AF}^{\star}(m)$ analogously after AF arrives. Suppose, for contradiction, that

$$F_{x_{AF}^{\star}(m)} < F_{x_{pre}^{\star}(m)}.$$

Because $x_{pre}^{\star}(m)$ is a pre-AF maximiser,

$$\Pi_{\ell t}^{E,pre}\big(x_{pre}^{\star}(m), R, m\big) \geq \Pi_{\ell t}^{E,pre}\big(x_{AF}^{\star}(m), R, m\big).$$

Because $\Delta^E_{\ell t}(x, m)$ is weakly increasing in F_{xt} and

$$F_{x^\star_{pre}(m)} > F_{x^\star_{AF}(m)},$$

we also have

$$\Delta^E_{\ell t}\big(x^\star_{pre}(m), m\big) \geq \Delta^E_{\ell t}\big(x^\star_{AF}(m), m\big).$$

Adding the two inequalities gives

$$\Pi^{E,AF}_{\ell t}\big(x^\star_{pre}(m), R, m\big) \geq \Pi^{E,AF}_{\ell t}\big(x^\star_{AF}(m), R, m\big).$$

If the inequality is strict, this contradicts the definition of $x^\star_{AF}(m)$ as a post-AF maximiser. If equality holds, then $x^\star_{pre}(m)$ is also a post-AF maximiser and has a strictly larger frontier index than $x^\star_{AF}(m)$, contradicting the tie-breaking rule that selects the maximal-frontier post-AF maximiser. Therefore

$$F_{x^\star_{AF}(m)} \geq F_{x^\star_{pre}(m)}.$$

□

Proof of Proposition 6. Fix a regime $s \in \{pre, AF\}$ and a protein $y \neq z$ with $J_t(y) = 1$. Maintain the transition convention from Section 3.2. Consider first an auxiliary action $a \in \{S, L\}$.

Under the additive uncertainty specification (5)–(7), an S experiment at y changes only $u_S(y)$ and leaves $u_f(\cdot)$, $u_0(\cdot)$, and $u_L(\cdot)$ unchanged except through terms that do not enter that experiment's payoff. Because the new first-ever structure at z affects future public information only through $\mathcal{F}_{t+1}$ and frontier position, it does not alter the benefit side of a future S experiment under the maintained transition convention. Hence

$$B^s_{\ell,t+1}(y, S, m) = B^s_{\ell t}(y, S, m).$$

Exactly the same argument applies to $a = L$, so

$$B^s_{\ell,t+1}(y, L, m) = B^s_{\ell t}(y, L, m).$$

Now turn to costs. In the pre-AF regime, costs are given by (9); in the AF regime, costs are given by (16). In either case, for a fixed target y, action $a \in \{S, L\}$, method m, and laboratory status $I_{\ell y t}$, all terms in the square brackets are unchanged between t and $t+1$. The only term affected by the new neighbouring entry is the frontier multiplier $\chi(F_{yt})$. Because χ is increasing and $F_{y,t+1} \leq F_{yt}$ by assumption,

$$c^s_{\ell,t+1}(y, a, m) \leq c^s_{\ell t}(y, a, m).$$

Combining the unchanged gross value with the weakly lower cost gives

$$\Pi^s_{\ell,t+1}(y, a, m) = B^s_{\ell,t+1}(y, a, m) - c^s_{\ell,t+1}(y, a, m) \geq B^s_{\ell t}(y, a, m) - c^s_{\ell t}(y, a, m) = \Pi^s_{\ell t}(y, a, m)$$

for $a \in \{S, L\}$.

For $a \in \{R, Q\}$, the same cost inequality continues to hold, but the addition of the new

public signal s_z changes the scaffold uncertainty term $u_f(\cdot)$ and therefore the benefit side of future static experiments. Without additional restrictions on how that benefit effect compares to the cost reduction, the sign of the net payoff change is not identified. This proves the proposition. □

Proof of Proposition 7. Define the objective in (17) by

$$\pi(\Delta e, F) = B(\Delta e) - \eta_m \big[\phi_m \mathbf{1}\{I_{\ell xt} = 0\} + g_{R,m}(\Delta e)\big]\chi(F).$$

Take $\Delta e_2 > \Delta e_1$ and $F_2 > F_1$. Because $g_{R,m}$ and χ are increasing,

$$\begin{aligned}\pi(\Delta e_2, F_2) - \pi(\Delta e_1, F_2) &= B(\Delta e_2) - B(\Delta e_1) - \eta_m \chi(F_2)\big[g_{R,m}(\Delta e_2) - g_{R,m}(\Delta e_1)\big] \\ &\le B(\Delta e_2) - B(\Delta e_1) - \eta_m \chi(F_1)\big[g_{R,m}(\Delta e_2) - g_{R,m}(\Delta e_1)\big] \\ &= \pi(\Delta e_2, F_1) - \pi(\Delta e_1, F_1).\end{aligned}$$

Hence $\pi(\Delta e, F)$ has decreasing differences in $(\Delta e, F)$. Standard monotone comparative statics therefore implies that the set of maximizers of (17) is weakly decreasing in F. Under the maintained uniqueness assumption, the optimal coverage increment $\Delta e^\star(x, m)$ is therefore weakly decreasing in F_x. □

A.2 Alternative knowledge-space foundation

This appendix shows that the main qualitative conclusions of Section 3 do not depend on the Brownian prior in (1). A slightly richer Gaussian specification is both tractable and closer to the protein research setting. It separates (i) a smooth homologous-scaffold component, which can be inferred from nearby solved relatives and which AF2 predicts relatively well, from (ii) a protein-specific residual component, which captures target-specific features such as flexible loops, disorder, alternative local packing, construct effects, and other idiosyncratic details that do not spill over strongly across the tree.

The economic point of the extension is simple. Brownian motion forces all static uncertainty to be smooth on the tree. In the protein setting, that is too strong: nearby homologues are informative about a dominant scaffold, but even proteins with close solved relatives retain target-specific uncertainty. An Ornstein–Uhlenbeck (OU) component plus a nugget is therefore a more faithful reduced form. The OU piece implies bounded, exponentially decaying dependence across the tree, while the nugget prevents the model from over-smoothing away protein-specific uncertainty.

Alternative prior. Replace (1) by

$$Y(x) = f(x) + \xi_x, \tag{25}$$

where f is a centered Gaussian process on $(\mathcal{T}, D_{\mathcal{T}})$ with covariance kernel

$$\mathrm{Cov}\big(f(x_i), f(x_j)\big) = \sigma_f^2 \exp\{-\alpha D_{\mathcal{T}}(x_i, x_j)\}, \qquad \alpha > 0,\ \sigma_f^2 > 0, \tag{26}$$

and

$$\xi_x \overset{ind}{\sim} \mathcal{N}(0, \tau_x^2), \qquad \tau_x^2 \geq 0. \tag{27}$$

The latent field $f(x)$ is the template-relevant scaffold component and ξ_x is a protein-specific nugget.

Experimental observations remain

$$s_i = Y(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \nu_i). \tag{28}$$

For a set of observed proteins $\mathcal{O}_t = \{x_1, \ldots, x_n\}$, define

$$K_f = \left[\sigma_f^2 e^{-\alpha D_{\mathcal{T}}(x_i, x_j)}\right]_{i,j=1}^{n}, \qquad T = \operatorname{diag}(\tau_{x_1}^2, \ldots, \tau_{x_n}^2), \qquad \Lambda = \operatorname{diag}(\nu_1, \ldots, \nu_n),$$

and

$$k_f(x, \mathcal{O}_t) = \left(\sigma_f^2 e^{-\alpha D_{\mathcal{T}}(x, x_i)}\right)_{i=1}^{n}.$$

Standard Gaussian conditioning gives the posterior variance of the smooth component:

$$\operatorname{Var}\big(f(x) \mid \mathcal{F}_t\big) = \sigma_f^2 - k_f(x, \mathcal{O}_t)\big(K_f + T + \Lambda\big)^{-1} k_f(\mathcal{O}_t, x). \tag{29}$$

Equivalently, the total latent field Y is itself a Gaussian process with kernel

$$k_Y(x_i, x_j) = \sigma_f^2 e^{-\alpha D_{\mathcal{T}}(x_i, x_j)} + \mathbf{1}\{x_i = x_j\}\tau_{x_i}^2,$$

so exact posterior moments of Y are available in closed form:

$$\operatorname{Var}\big(Y(x) \mid \mathcal{F}_t\big) = k_Y(x, x) - k_Y(x, \mathcal{O}_t)\big(K_Y + \Lambda\big)^{-1} k_Y(\mathcal{O}_t, x), \qquad K_Y = K_f + T. \tag{30}$$

This specification is a step more faithful to the protein setting for two reasons. First, homology information decays with evolutionary distance and then effectively saturates: once a protein is already far from any solved relative, additional distance matters less than it does locally. The OU kernel captures exactly that bounded, exponentially decaying dependence. Second, even proteins with close solved homologues retain target-specific uncertainty. The nugget captures this residual component directly.

Local edge formula. The Brownian bridge formula in Lemma 1 is replaced by an OU bridge plus a nugget.

Lemma 2 (Local posterior variance under OU plus nugget)**.** *Suppose x lies on a branchless edge between adjacent nodes x_l and x_r. Let*

$$d_l = D_{\mathcal{T}}(x, x_l), \qquad d_r = D_{\mathcal{T}}(x, x_r), \qquad X = d_l + d_r,$$

and let $\mathcal{G}$ be the σ-field generated by all observations outside that open segment. Write

$$P_l^f = \operatorname{Var}\big(f(x_l) \mid \mathcal{G}\big), \qquad P_r^f = \operatorname{Var}\big(f(x_r) \mid \mathcal{G}\big), \qquad C_{lr}^f = \operatorname{Cov}\big(f(x_l), f(x_r) \mid \mathcal{G}\big).$$

Then

$$\mathrm{Var}\big(Y(x) \mid \mathcal{G}\big) = \tau_x^2 + V^{OU}(d_l, d_r) + a_l(d_l, d_r)^2 P_l^f + a_r(d_l, d_r)^2 P_r^f + 2a_l(d_l, d_r) a_r(d_l, d_r) C_{lr}^f, \tag{31}$$

where

$$a_l(d_l, d_r) = \frac{\sinh(\alpha d_r)}{\sinh(\alpha X)}, \qquad a_r(d_l, d_r) = \frac{\sinh(\alpha d_l)}{\sinh(\alpha X)},$$
$$V^{OU}(d_l, d_r) = \sigma_f^2 \, \frac{(1 - e^{-2\alpha d_l})(1 - e^{-2\alpha d_r})}{1 - e^{-2\alpha X}}.$$

Proof. Because the OU field f is Gaussian and Markov on the tree, conditional on $f(x_l)$, $f(x_r)$, and $\mathcal{G}$, the process on the open edge is an OU bridge independent of $\mathcal{G}$. Standard Gaussian conditioning yields

$$f(x) = a_l(d_l, d_r) f(x_l) + a_r(d_l, d_r) f(x_r) + \zeta_x,$$

where

$$\zeta_x \mid f(x_l), f(x_r), \mathcal{G} \sim \mathcal{N}\big(0, V^{OU}(d_l, d_r)\big)$$

and ζ_x is conditionally independent of $(f(x_l), f(x_r), \mathcal{G})$. Since $Y(x) = f(x) + \xi_x$ and ξ_x is independent of $\mathcal{G}$ with variance τ_x^2, (31) follows from the law of total variance. □

The same frontier variables used in the main text remain natural reduced-form proxies under (25). A larger nearest-solved distance lowers the effective endpoint weights in (31); a larger gap raises the bridge term $V^{OU}(d_l, d_r)$; and larger cluster or neighbourhood PDB shares increase the number and precision of nearby conditioning signals entering (29)–(30). Relative to Brownian motion, the difference is that distance effects are saturating rather than approximately linear.

How AF2 enters. To reflect the fact that AF2 is much closer to a scaffold predictor than a full target-specific uncertainty eliminator, let AF2 load on the smooth component only:

$$\hat{y}^{AF}(x) = f(x) + \varepsilon^{AF}(x), \qquad \varepsilon^{AF}(x) \sim \mathcal{N}(0, \omega(x)). \tag{32}$$

Let $u_f(x)$ denote the unresolved scaffold-related uncertainty before AF2. Then the post-AF smooth uncertainty is

$$u_f^{AF}(x) = \left(\frac{1}{u_f(x)} + \frac{1}{\omega(x)} \right)^{-1}. \tag{33}$$

For the value-theoretic part of Section 3, it is convenient to summarise static uncertainty by the reduced-form decomposition

$$u_0(x; \mathcal{F}_t, \mathbf{q}_x) = u_f(x) + u_\xi(x) + \kappa_E h_E(e_x) + \kappa_Q h_Q(\rho_x), \tag{34}$$

where $u_\xi(x)$ is a target-specific residual component. Before any direct target-specific observation on x, one can take $u_\xi(x) = \tau_x^2$. The exact Gaussian posterior variance in

(30) need not decompose additively after direct observations on x; equation (34) is the reduced-form summary that preserves the distinction between spillover-rich scaffold uncertainty and target-specific residual uncertainty. Nearby solved relatives reduce $u_f(x)$ but, absent a direct target-specific observation, do not reduce $u_\xi(x)$. This is precisely why the specification is better aligned with the protein setting.

If one instead wished the full marginal value of static experiments to satisfy the scalar attenuation formula in (15), one could let AF2 provide a noisy signal on $Y(x)$ rather than on $f(x)$ and suppress the target-specific residual. The one-component benchmark in Section 3.5 is the nested special case. The stricter formulation in (32) is adopted here because it is the more biologically faithful restriction.

Robustness of the Section 3 results. The effect of the alternative prior on the formal results is summarised next.

Proposition 8 (Section 3 under OU plus nugget)**.** *Maintain the cost, action, and value definitions in Section 3, but replace* (1) *by* (25) *and let AF2 enter through* (32)*. Then:*

(i) Proposition 1 and Proposition 2 are unchanged. They do not depend on the specific Gaussian prior over the knowledge space.

(ii) Proposition 3 continues to hold for the smooth scaffold component, but the full marginal value of a static experiment becomes a partial attenuation *result once one allows a direct reduction in the target-specific nugget. For $a \in \{R, Q\}$, let $d\Delta_f(y; x, a) \geq 0$ denote the smooth spillover reduction in static uncertainty at y induced by an infinitesimal experiment at x, and let $d\Delta_\xi(x; x, a) \geq 0$ denote the direct reduction in the target-specific nugget at x. Under the same interior condition used in Proposition 3,*

$$dB_{\ell t}^{AF}(x,a) = \bar{\lambda}_{\ell t}^{f}(x,a)\, dB_{\ell t}^{pre,f}(x,a) + dB_{\ell t}^{pre,\xi}(x,a), \tag{35}$$

where

$$dB_{\ell t}^{pre,f}(x,a) = \frac{\alpha_0}{q_0} \int_{\mathcal{T}} w_\ell(y)\, d\Delta_f(y; x, a)\, d\mu(y),$$
$$dB_{\ell t}^{pre,\xi}(x,a) = \frac{\alpha_0}{q_0} w_\ell(x)\, d\Delta_\xi(x; x, a),$$

and

$$\bar{\lambda}_{\ell t}^{f}(x,a) = \frac{\int_{\mathcal{T}} w_\ell(y)\lambda_f(y)\, d\Delta_f(y; x, a)\, d\mu(y)}{\int_{\mathcal{T}} w_\ell(y)\, d\Delta_f(y; x, a)\, d\mu(y)} \in (0,1).$$

and

$$\lambda_f(y) = \frac{\omega(y)^2}{\big(u_f(y) + \omega(y)\big)^2}.$$

Hence AF2 still lowers the marginal value of static-scaffold experiments, but by less than in a one-component model whenever $dB_{\ell t}^{pre,\xi}(x,a) > 0$. For $a \in \{S, L\}$, the second half of Proposition 3 is unchanged because AF2 still does not enter u_S or u_L.

(iii) Proposition 4 is unchanged. It is a cost-side result and therefore does not depend on the covariance kernel.

(iv) Proposition 5 is unchanged as an ordinal comparative-statics statement: its proof only requires that the net AF shock $\Delta^{E}_{\ell t}(x, m)$ be weakly increasing in the frontier index F_{xt}. Under the present specification, that monotonicity is more naturally interpreted as saying that AF2 raises payoffs more for proteins with lower smooth information stock $u_f(x)$. A large nugget τ_x^2 can dampen the magnitude of the frontier shift, but it does not alter the proposition itself.

(v) Proposition 7 is unchanged for the comparison made in the main text, which holds fixed all protein characteristics other than frontier position. Under that comparison the proof is identical. If, in addition, frontier proteins systematically have larger nugget variances τ_x^2, then the own-target return to deeper experiments is larger at the frontier and the negative depth gradient can be weakened, though not mechanically reversed.

Proof. Parts (i), (iii), and the formal statement in part (iv) follow immediately because the corresponding proofs in Appendix A.1 do not use the Brownian bridge formula. For part (ii), decompose the infinitesimal value of a static experiment into a smooth spillover component and a direct target-specific component. AF2 attenuates the former through (33) and leaves the latter unchanged, which yields (35). Part (v) follows because Proposition 7 is a monotone-comparative-statics result for the depth problem (17); the additional nugget matters only if one allows frontier position to shift the benefit function B directly rather than holding all non-frontier target characteristics fixed, which the main-text proposition does not do. □

Implications for the empirical predictions. The theory-guided predictions in Section 3.7 survive this change with only modest reinterpretation.

(i) Prediction 1, Prediction 3, Prediction 6, Prediction 7, and Prediction 8 are unchanged.

(ii) Prediction 2 is unchanged in sign but may be weaker in magnitude. AF2 still especially helps proteins that are poor scaffold-inference candidates ex ante, but the response is smaller when frontier proteins also have large target-specific nuggets that AF2 does not resolve.

(iii) Prediction 4 is unchanged conditional on comparing proteins with similar nugget size. If high-frontier proteins also have larger τ_x^2, the negative frontier-depth gradient is attenuated because own-target depth remains valuable even after AF2.

(iv) Prediction 5 becomes more naturally ambiguous. In the Brownian one-component model AF2 is a broad substitute for static-structure work. Under OU plus nugget, AF2 is mainly a substitute for the smooth scaffold component, while experiments can still be valuable for target-specific residue completion and local refinement. This makes it easier to rationalize empirical cases in which AF2 reduces some follow-up work but increases target-specific completeness experiments.

In short, the OU+nugget specification preserves the paper's core comparative statics while weakening one implication that is too strong in a literal protein setting: AF2 should not be thought of as erasing all static uncertainty. It is better understood as sharply compressing the *shared* scaffold component of uncertainty while leaving a protein-specific residual that still gives experiments value.

A.3 Data Construction

A.3.1 Data Sources

Universal Protein Resource Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (e.g. the amino acid sequence, protein name or description, and taxonomic data) along with citation information. It consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis (UniProtKB/Swiss-Prot), and a section with 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 repository for structural coordinates of biological macromolecules, primarily proteins. Since its inception in 1971, it is freely available at https://cdn.rcsb.org/rcsb-pdb/pdb50mesh/index.html. It contains more than 170,000 structures; 150,000 of those had corresponding primary citation describing these entries in a peer-reviewed journal. 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 table are useful for this project, available here (downloaded in 2025-09-01) or the FTP. The first is the *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 the experimental coverage. The second is the *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 is the *pdb_pubmed* is a summary of the Pubmed id(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. I 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 catalog 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.3.2 Deposition Categories

Over 90% of PDB depositions 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. Each PDB deposition is assigned to exactly one of five mutually exclusive categories based on its primary scientific contribution. The classification proceeds in two stages: first, five binary flags are computed for each deposition; second, a priority ordering assigns each deposition to a unique category.

Binary flags. 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 is the most consequential type of deposition: it generates positive externalities for all sequence neighbours via homology modelling and opens the protein to structure-based drug design, mechanistic annotation, and comparative analysis across species.

- **Completeness *R* (or Improved Coverage).** The deposition resolves residues not covered by any prior deposition for the same protein. Formally, $R_d = 1$ if d is an intensive-margin deposition and the coverage increment $\Delta e_d > 0$, where Δe_d is defined in Equation 19. 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). 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 cryo-cooling 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.

Priority assignment. Because the five flags are not mutually exclusive—a deposition may simultaneously extend coverage, improve resolution, and contain a new ligand—each deposition is assigned to a unique primary category using the priority ordering $E \succ R \succ SL \succ Q \succ \text{Corr.}$, where SL combines the S and L flags (a deposition is classified as SL if $S_d = 1$ or $L_d = 1$, or both). The priority ordering reflects the theoretical hierarchy of informational value: a first-ever structure generates positive externalities for all sequence neighbours via homology modelling; residue extension expands the experimentally validated portion of the protein; conformational and ligand discoveries create target-specific mechanistic value; resolution improvement refines an existing structural model; and a corroborating structure provides independent validation without new structural information.

Formally:

$$\text{Category}(d) = \begin{cases} E & \text{if } E_d = 1 \\ R & \text{if } E_d = 0 \text{ and } R_d = 1 \\ SL & \text{if } E_d = 0,\ R_d = 0, \text{ and } (S_d = 1 \text{ or } L_d = 1) \\ Q & \text{if } E_d = 0,\ R_d = 0,\ S_d = 0,\ L_d = 0, \text{ and } Q_d = 1 \\ \text{Corr.} & \text{otherwise} \end{cases}$$

A.3.3 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 3.3 rests on an estimated tree that organises all Swiss-Prot 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 Swiss-Prot 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 percent 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 (**?**), 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 $D = 1.0$. 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. When p's own cluster contains solved members, the within-cluster path may be shorter, and we take the minimum:

$$d_{\text{tree}}(p, t) = \min\Big(\underbrace{d_{\text{intra}}(p) + d_{\text{within}}(p, t)}_{\text{within-cluster path}}, \underbrace{d_{\text{intra}}(p) + d^{\text{cluster}}_{\text{tree}}(t)}_{\text{inter-cluster path}} \Big),$$

where $d_{\text{within}}(p, t)$ is the spoke distance from p to the nearest solved member within the cluster (if any), and $d^{\text{cluster}}_{\text{tree}}(t)$ is the tree distance from p's cluster representative to the nearest solved cluster in year t.

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

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

where d_{left} and d_{right} are the tree distances from p's cluster to the nearest and second-nearest solved clusters. A wide gap means a large span of sequence space around the protein lacks a usable structural template.

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 assign these the maximum distance $d_{\text{tree}}(p, t) = 1.0$. 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 two local knowledge density measures. The *cluster PDB share* is the fraction of cluster members with any PDB structure by year t. Even when a cluster's 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 *neighbourhood PDB share* looks outward: we contract the tree by collapsing Steiner nodes so that each cluster is directly adjacent to its nearest cluster neighbours, then compute the fraction of those neighbours that have any PDB structure by year t. A cluster whose tree neighbours are all solved sits in a well-characterised region of sequence space; a cluster surrounded by unsolved neighbours lies at the frontier. Both measures are recomputed for each year as the solved set expands.

A.4 Additional Outcomes

Figure 6: PSI experimental funnel: number of targets reaching each pipeline stage. Percentages show cumulative survival relative to selected targets.

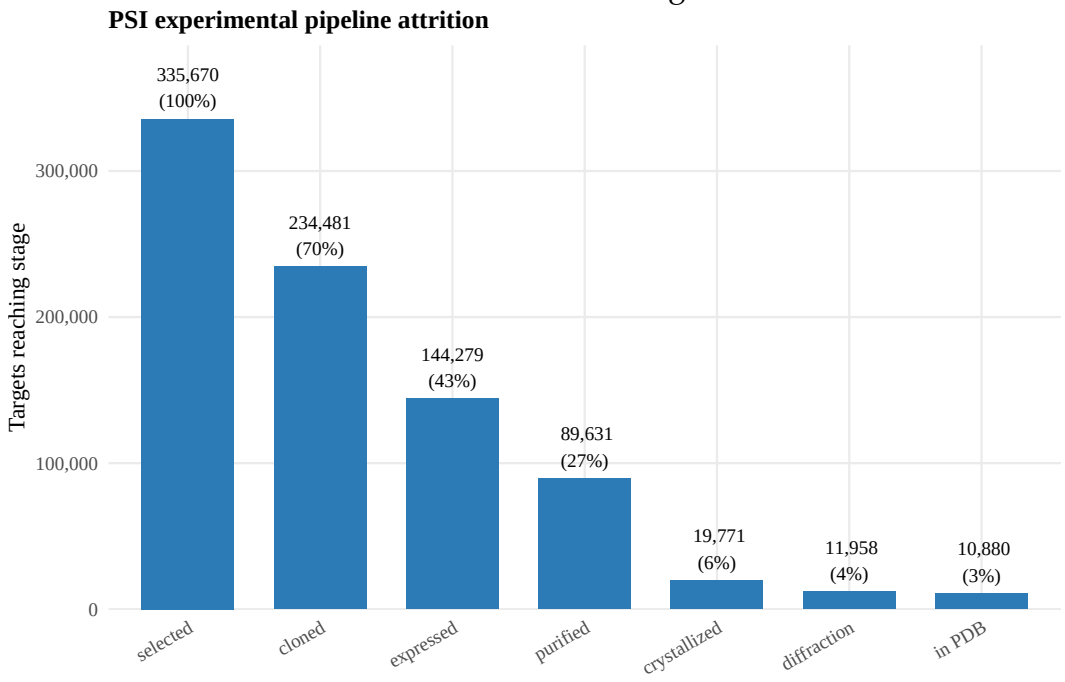


Table 12: Across-lab channel: probability of entrant-lab deposition

	(1) Pr(Entrant)	(2) Pr(Entrant)
Post × AF cov	-0.1126***	-0.1461***
	(0.0362)	(0.0537)
Post × AF cov × Dist.		0.1210
		(0.0811)
Protein FE	X	X
Year FE	X	X
N	166,147	166,147
R^2	0.368	0.368
$\bar{Y}_{t\leq 2021}$	0.299	0.299
$\hat{\beta}/\bar{Y}$	-37.7%	-48.9%

Notes. This table reports OLS estimates of the probability that an intensive-margin deposition is made by an entrant lab, defined as one whose last (senior) author has zero PDB depositions before 2017. Column (1) reports the baseline specification with protein and year fixed effects. Column (2) adds the frontier distance interaction (tree distance to nearest solved protein, as of 2016); the Post × Isolated control is included but not shown. The sample consists of intensive-margin PDB depositions with non-missing last author over 2017–2025. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5 Robustness Checks

A.5.1 Placebo, Pre-trend and Balance test, Event Study

Table 13: Placebo protein-year level

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. × Ext.
Post × AF cov (placebo)	-0.0002	0.0069*	0.0029	-0.0002	-0.0003
	(0.0004)	(0.0038)	(0.0062)	(0.0005)	(0.0003)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	4,774,455	206,424	4,774,455	4,774,455
R^2	0.414	0.529	0.478	0.931	0.110

Notes. Placebo test for Table 3. AFcov_p is randomly permuted across proteins (seed = 42), breaking the true treatment assignment while preserving the marginal distribution and time structure. All specifications identical to Table 3. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 14: Placebo deposition level

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov (placebo)	-0.0007	-0.0221	-0.0023*
	(0.0080)	(0.0189)	(0.0014)
Protein FE	X		X
Organism FE		X	
Year FE	X	X	X
Protein features		X	
N	194,810	10,702	185,418
R^2	0.272	0.299	0.259

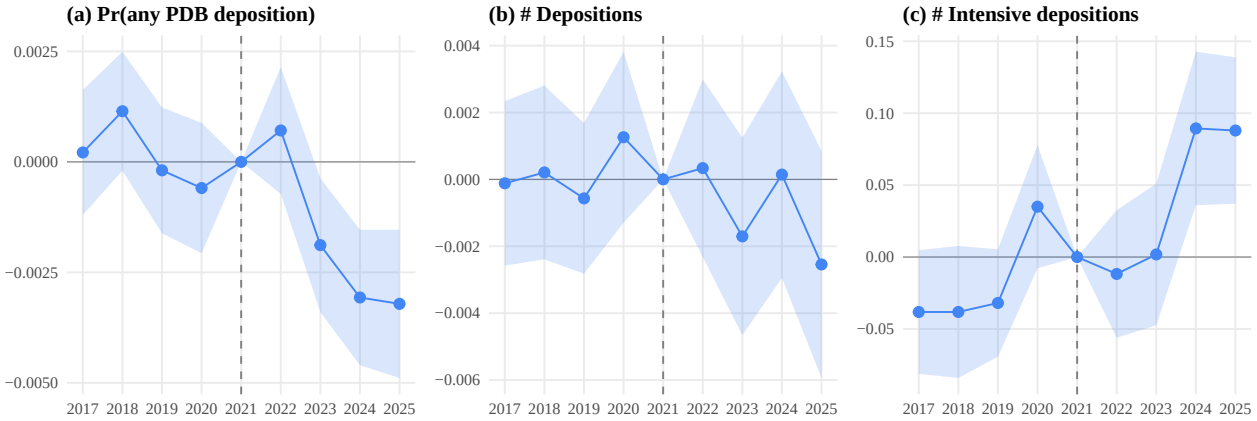
Notes. Placebo test for Table 4. Same permutation as Table 13. All specifications identical to Table 4. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 15: Covariate balance: differential pre-trends by AF coverage quartile

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

Notes. Each row regresses the covariate on AF coverage quartile × linear trend (2017–2021), with Q1 (lowest coverage) as reference, conditional on organism FE and year FE. Standard errors clustered at the protein level. Norm. diff. = coefficient / SD of the covariate. Values close to zero indicate no differential pre-trend between high- and low-coverage proteins.

Figure 7: Event study



Coefficients on year × high-group indicator, with 2021 as reference and low group as baseline. Groups defined by 2021 cross-section: high (AF cov > 0.9), low (AF cov < 0.1).

A.5.2 Alternative AF2 treatment and post period

Table 16: Protein-year panel baseline (treatment: mean pLDDT / 100)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. × Ext.
Post × pLDDT	-0.0085***	0.5381**	0.8418***	-0.0125***	-0.0030***
	(0.0008)	(0.2125)	(0.2711)	(0.0007)	(0.0006)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	171,252	72,342	4,774,455	4,774,455
R^2 / $\tilde{R}^2$	0.420	0.562	0.615	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.033	0.619	0.036	0.002
% effect	-107.4%	71.3%	132.0%	-34.2%	-140.1%

Notes. Same specification as Table 3 but replacing AF coverage with mean pLDDT (rescaled to [0,1]). Col (1), (4)–(5) are estimated with an OLS, and Col (2)–(3) with a Poisson pseudo-ML. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 17: Deposition-level baseline (treatment: mean pLDDT / 100)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × pLDDT	0.1402***	1.425***	0.0294***
	(0.0199)	(0.0323)	(0.0036)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.272	0.388	0.257
$\bar{Y}_{t\leq 2021}$	0.068	0.674	0.007
$\hat{\beta}/\bar{Y}$	206.3%	211.4%	407.0%

Notes. Same specification as Table 4 but replacing AF coverage with mean pLDDT (rescaled to [0,1]). Standard errors clustered at the protein level. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 18: Protein-year panel baseline (treatment: AF cov at pLDDT $\geq$ 90)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. × Ext.
Post × AF cov_{90}	-0.0034***	0.0522	0.1913**	-0.0053***	-0.0012***
	(0.0003)	(0.0681)	(0.0888)	(0.0003)	(0.0002)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	171,252	72,342	4,774,455	4,774,455
R^2 / $\tilde{R}^2$	0.420	0.562	0.615	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.033	0.619	0.036	0.002
% effect	-43.6%	5.4%	21.1%	-14.6%	-55.7%

Notes. Same specification as Table 3 but replacing AF cov_{70} (pLDDT $\geq$ 70) with AF cov_{90} (pLDDT $\geq$ 90, stricter threshold). Col (1), (4)–(5) are estimated with an OLS, and Col (2)–(3) with a Poisson pseudo-ML. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 19: Deposition-level baseline (treatment: AF cov at pLDDT $\geq$ 90)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post $\times$ AF cov_{90}	0.0591***	0.4907***	0.0135***
	(0.0063)	(0.0135)	(0.0013)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.272	0.362	0.258
$\bar{Y}_{t\leq 2021}$	0.068	0.674	0.007
$\hat{\beta}/\bar{Y}$	86.9%	72.8%	187.3%

Notes. Same specification as Table 4 but replacing AF cov_{70} with AF cov_{90} (pLDDT $\geq$ 90). Standard errors clustered at the protein level. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 20: Protein-year panel baseline (treatment: AF cov at pLDDT $\geq$ 50)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. $\times$ Ext.
Post $\times$ AF cov_{50}	-0.0040***	0.7948***	0.9005***	-0.0064***	-0.0016***
	(0.0007)	(0.1642)	(0.2084)	(0.0005)	(0.0005)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	171,252	72,342	4,774,455	4,774,455
R^2 / $\tilde{R}^2$	0.420	0.562	0.616	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.033	0.619	0.036	0.002
% effect	-50.1%	121.4%	146.1%	-17.4%	-72.5%

Notes. Same specification as Table 3 but replacing AF cov_{70} (pLDDT $\geq$ 70) with AF cov_{50} (pLDDT $\geq$ 50, stricter threshold). Col (1), (4)–(5) are estimated with an OLS, and Col (2)–(3) with a Poisson pseudo-ML. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 21: Deposition-level baseline (treatment: AF cov at pLDDT ≥ 50)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post $\times$ AF cov_{50}	0.0688***	0.9345***	0.0139***
	(0.0155)	(0.0263)	(0.0026)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.272	0.366	0.257
$\bar{Y}_{t\leq 2021}$	0.068	0.674	0.007
$\hat{\beta}/\bar{Y}$	101.2%	138.6%	193.1%

Notes. Same specification as Table 4 but replacing AF cov_{70} with AF cov_{50} (pLDDT ≥ 50). Standard errors clustered at the protein level. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 22: Protein-year panel baseline (initial AFDB release organisms)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. $\times$ Ext.
Post $\times$ AF cov	0.0067***	0.5091***	0.6842***	0.0434***	-0.0017**
	(0.0011)	(0.1266)	(0.1611)	(0.0012)	(0.0008)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	893,160	114,516	52,569	893,160	893,160
R^2 / $\tilde{R}^2$	0.434	0.587	0.634	0.917	0.106
$\bar{Y}_{p,t\leq 2021}$	0.030	0.137	0.891	0.097	0.008
% effect	22.1%	66.4%	98.2%	44.9%	-23.2%

Notes. Same specification as Table 3 but restricted to proteins from the 20 model organisms included in the initial AlphaFold database release (July 2021): *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *S. pombe*, *E. coli* K-12, *A. thaliana*, *O. sativa*, *Z. mays*, *G. max*, *T. cruzi*, *L. infantum*, *P. falciparum*, *M. tuberculosis*, *S. aureus*, *D. discoideum*, and *M. jannaschii*. These proteins had AF2 predictions available from mid-2021; the remaining SwissProt proteins received predictions in the July 2022 database expansion. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 23: Deposition-level baseline (initial AFDB release organisms)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov	0.0797***	0.8570***	0.0129***
	(0.0121)	(0.0201)	(0.0020)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	152,991	6,883	147,008
R^2	0.267	0.321	0.250
$\bar{Y}_{t\leq 2021}$	0.055	0.602	0.007
$\hat{\beta}/\bar{Y}$	143.7%	142.5%	173.2%

Note: Same specification as Table 4 but restricted to proteins from the 20 initial AFDB release organisms. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 24: Protein-year panel baseline (Post $\geq$ 2022)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. × Ext.
Post_{2022} × AF cov	-0.0035***	0.4043***	0.5133***	-0.0042***	-0.0013***
	(0.0004)	(0.1138)	(0.1487)	(0.0004)	(0.0003)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,774,455	171,252	72,342	4,774,455	4,774,455
R^2 / $\tilde{R}^2$	0.420	0.562	0.615	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.033	0.619	0.036	0.002
% effect	-44.1%	49.8%	67.1%	-11.6%	-57.9%

Note: Same specification as Table 3 but defining the post-period as 2022–2025 instead of 2021–2025, so that 2021 is coded as pre-treatment. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 25: Deposition-level baseline (Post $\geq$ 2022)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
$\text{Post}_{2022} \times$ AF cov	0.0703***	0.8009***	0.0135***
	(0.0114)	(0.0179)	(0.0020)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	196,720	10,575	187,613
R^2	0.272	0.396	0.257
$\bar{Y}_{t\leq 2021}$	0.068	0.674	0.007
$\hat{\beta}/\bar{Y}$	103.4%	118.8%	186.5%

Note: Same specification as Table 4 but defining the post-period as 2022–2025. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 26: Protein-year panel baseline (AF cov $\geq$ 0.2)

	(1) Pr(PDB dep.)	(2) # Dep.	(3) # Intensive	(4) PDB cov.	(5) Dist. $\times$ Ext.
Post $\times$ AF cov	-0.0052***	0.4737***	0.5849***	-0.0090***	-0.0011***
	(0.0006)	(0.1215)	(0.1555)	(0.0005)	(0.0004)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
N	4,695,660	168,921	71,442	4,695,660	4,695,660
R^2 / $\tilde{R}^2$	0.421	0.562	0.615	0.932	0.110
$\bar{Y}_{p,t\leq 2021}$	0.008	0.034	0.641	0.037	0.002
% effect	-66.0%	60.6%	79.5%	-24.4%	-52.5%

Notes. Same specification as Table 3 but restricted to proteins with AF $\text{cov}_{70} \geq 0.2$, dropping proteins with negligible AlphaFold predictions for which AF2 might be correlated with with pre-existing frontier position. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 27: Deposition-level baseline (AF cov ≥ 0.2)

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov	0.0796***	0.8683***	0.0152***
	(0.0113)	(0.0174)	(0.0020)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Protein features		X	
N	195,437	10,417	186,460
R^2	0.271	0.390	0.257
$\bar{Y}_{t\leq 2021}$	0.066	0.691	0.007
$\hat{\beta}/\bar{Y}$	121.3%	125.7%	222.0%

Note: Same specification as Table 4 but restricted to proteins with AF $\text{cov}_{70} \geq 0.2$. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5.3 Additional Controls

Table 28: Dynamic network spillover on the intensive margin

	(1) Pr(R)	(2) Pr(SL)	(3) # R	(4) # SL
Share neighbors PDB	-0.0132	0.1289***	-0.0635	1.560***
	(0.0131)	(0.0177)	(0.3067)	(0.2769)
Post × AF cov	0.0070	-0.0163***	0.0361	0.3424*
	(0.0045)	(0.0047)	(0.1499)	(0.1778)
Post × Neighbor AF cov	-0.0016	-0.0060**	-0.0968	-0.0381
	(0.0019)	(0.0025)	(0.0836)	(0.0876)
Protein FE	X	X	X	X
Year FE	X	X	X	X
N	203,202	203,202	39,825	49,383
R^2 / $\tilde{R}^2$	0.190	0.509	0.087	0.561
$\bar{Y}_{t\leq 2021}$	0.035	0.073	0.048	0.364

Notes. This table decomposes the network spillover by deposition type, including both the static measure (Post × Neighbor AF cov, pre-determined) and the dynamic measure (Share neighbors PDB, time-varying). The dependent variable in column (1) is an indicator for residue-extension depositions (R); column (2) is an indicator for conformational-state and ligand-binding depositions (SL); columns (3) and (4) report Poisson pseudo-ML estimates on the counts of R and SL depositions, respectively. The sample consists of proteins with at least one PDB structure before 2017 in a balanced panel over 2017–2025. All columns include protein

and year fixed effects. Columns (1)–(2) are estimated by OLS; columns (3)–(4) by Poisson pseudo-ML, where coefficients are semi-elasticities. Robust standard errors in parentheses are clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 29: Protein-year panel baseline with lab fixed effects

	(1) Pr(PDB dep.)	(2) # Dep.	(3) Pr(Int.)	(4) PDB cov.	(5) Dist. × Ext.
Post × AF cov	-0.0044	0.3114***	-0.0106**	0.0595***	-0.0088***
	(0.0046)	(0.0603)	(0.0047)	(0.0046)	(0.0033)
Protein FE	X	X	X	X	X
Year FE	X	X	X	X	X
Lab FE	X	X	X	X	X
N	287,460	287,460	191,376	287,460	287,460
R^2	0.331	0.523	0.470	0.794	0.085
$\bar{Y}_{p,t\leq 2021}$	0.133	0.572	0.129	0.566	0.035

Notes. OLS estimates with standard errors clustered at the protein level. Same outcomes as Table 3 (Col 2 is OLS on deposition counts rather than Poisson; Col 3 is Pr(any intensive) rather than Poisson count) with lab (last-author) fixed effects added. The lab is the last author on the protein's first-ever PDB deposition. Proteins without a matched last author are excluded. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 30: Deposition-level baseline with lab fixed effects

	(1) Extensive	(2) Δe_d | Ext.	(3) Δe_d | Int.
Post × AF cov	0.0833***	0.5933***	0.0140***
	(0.0098)	(0.0274)	(0.0021)
Organism FE		X	
Protein FE	X		X
Year FE	X	X	X
Lab FE	X	X	X
Protein features		X	
N	177,769	8,161	168,932
R^2	0.343	0.616	0.323
$\bar{Y}_{t\leq 2021}$	0.067	0.670	0.007
$\hat{\beta}/\bar{Y}$	124.9%	88.5%	190.6%

Notes. Same specification as Table 4 with lab (last-author) fixed effects added. Standard errors clustered at the protein level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.