TxBench-PP: Analyzing AI Agent Performance on Small-Molecule Preclinical Pharmacology

Hannah Le[1,*], Ramesh Ramasamy[1,*], Alex Urrutia[1,*], Mahsa Yazdani[1,*], Tim Proctor[1], Kenny Workman[1]

[1]LatchBio, San Francisco, CA, USA
[*]These authors contributed equally.

Correspondence: kenny@latch.bio

ABSTRACT

Artificial intelligence (AI) agents promise to accelerate drug discovery by compressing interpretation and decision-making loops, but practical deployment requires trusted evaluation on realistic program decisions. We introduce TherapeuticsBench Preclinical Pharmacology (TxBench-PP), a verifiable benchmark for small-molecule preclinical pharmacology and the first focused slice of a broader TherapeuticsBench effort across drug-discovery stages and therapeutic modalities. TxBench-PP tests whether agents can recover accurate conclusions from real-world assay data rather than memorized facts from literature. The benchmark contains 100 evaluations indexed by program stage, assay type, and task structure, spanning mechanism-of-action (MoA) and pharmacodynamic (PD) reasoning, compound-target engagement, causal target validation, developability and safety, and translational efficacy. Agents receive realistic workflow snapshots, inspect files in a coding environment, and return structured answers graded deterministically. Across 16 model–harness configurations, comprising 11 models and 4,800 trajectories, no system reliably recovered preclinical pharmacology decisions. The strongest configuration, Claude Opus 4.8 / Pi, passed 59.3% of endpoint attempts (178/300; 95% CI, 51.1–67.6), followed by GPT-5.5 / Pi at 55.3% (166/300; 47.0–63.6).

TxBench-PP: Overview

TxBench-PP is a focused TherapeuticsBench benchmark for small-molecule preclinical pharmacology. Each evaluation supplies realistic workflow data from a local preclinical decision point and asks whether an agent can recover the result supported by the provided data.

Across 16 model–harness configurations, the strongest agent system passed 59.3% of endpoint attempts (178/300; 95% CI, 51.1–67.6).

TherapeuticsBench is the broader roadmap across drug-discovery stages and therapeutic modalities. The schematic below contextualizes TxBench-PP with future work.

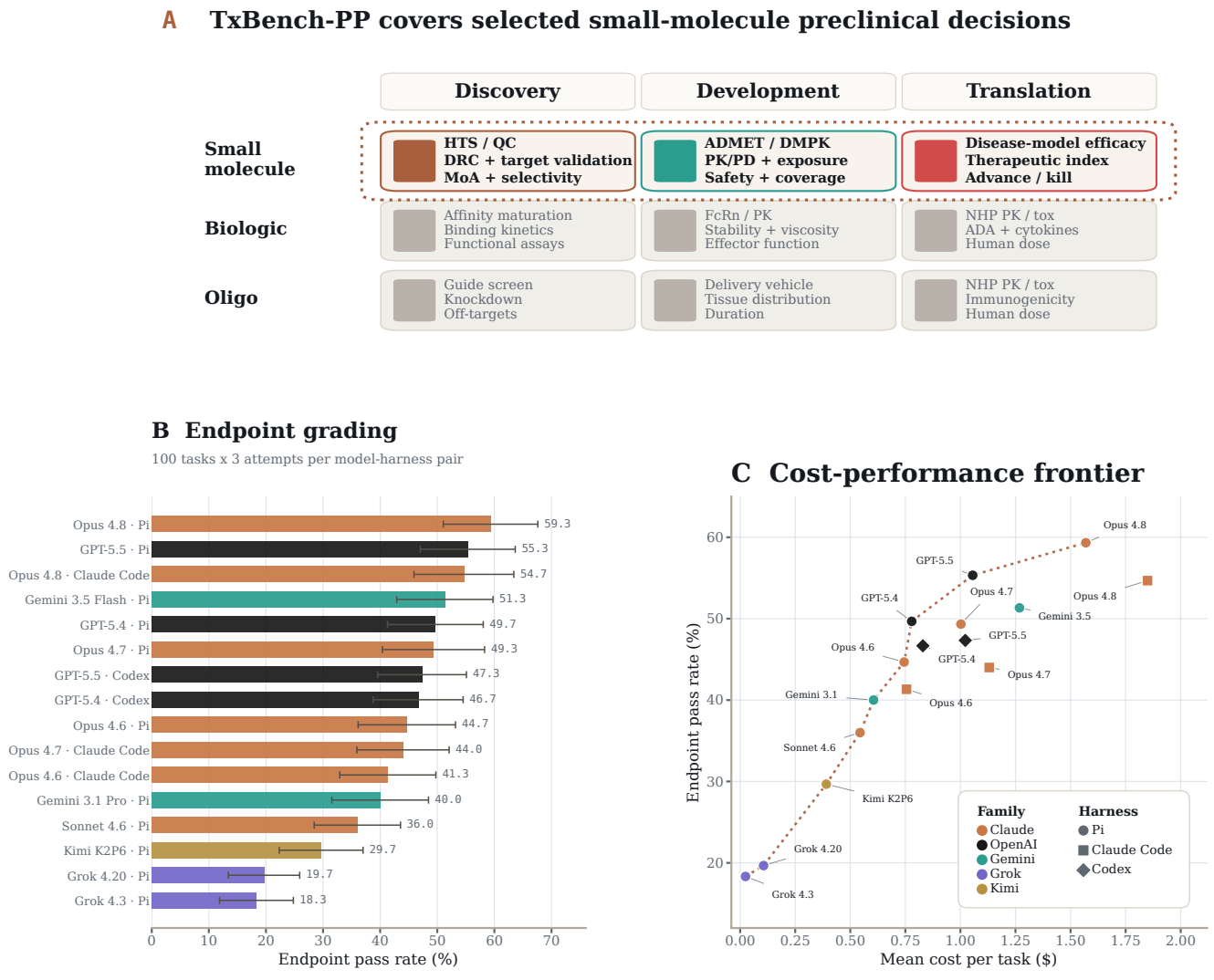

Figure 1: TxBench-PP scope and topline performance. **A**, TxBench-PP places local small-molecule preclinical pharmacology decisions within the broader TherapeuticsBench roadmap. The dotted boundary marks the task groups in this release; biologics, oligonucleotides, and full program histories are out of scope. **B**, Endpoint pass rates across 16 model–harness configurations; y-axis labels identify both model and harness. Confidence intervals are computed over task-level pass rates. **C**, Mean cost per task against endpoint pass rate; color denotes model family and marker shape denotes harness.

Introduction

While experiments are rate-limited by natural processes, human decisions and organizational consensus often make up significant components of program timelines in drug discovery [1–4]. Agents promise to accelerate discovery, development, and translation by compressing these interpretation and decision-making loops. Specific knowledge from constituent steps of the drug discovery pipeline can be encoded into frontier models through post-training, tool use, and domain-specific scaffolding, then deployed into practical data interpretation workflows [5–7].

However, the practical use of agentic systems in industrial workflows requires standard and trusted methods of evaluating performance. Recent biology-agent benchmarks have emphasized verifiable analysis tasks and realistic scientific data [8–14]. This is especially challenging in drug discovery because the ecosystem is a sprawling landscape of assay categories, development stages, therapeutic modalities, and decision types. Benchmarks must therefore measure realistic tasks while providing focused treatment of the many local scientific judgements that tile the biotech ecosystem.

Preclinical small-molecule pharmacology is one important piece of this landscape. Potency, target engagement, mechanism-of-action evidence, pharmacodynamic response, exposure, safety, and in vivo efficacy are converted into practical decisions about whether to advance, hold, repeat, or kill a molecular candidate. These decisions sit upstream of costly medicinal chemistry, animal studies, and human experiments. They are easy to get wrong and often depend on fragmented expert judgement.

We introduce TxBench-PP, a verifiable benchmark for evaluating AI agents on small-molecule preclinical pharmacology. TxBench-PP is best understood as a focused benchmark within the broader TherapeuticsBench roadmap rather than a holistic evaluation of drug discovery. Each evaluation supplies realistic workflow artifacts from a local preclinical decision point and asks the agent to return a structured, deterministically gradable answer. Importantly, the benchmark tests if agents can recover decisions from supplied data instead of leaning on well-understood mechanisms and literature knowledge, with many tasks intentionally designed to penalize memorized solutions.

TxBench-PP contains 100 evaluations indexed by program stage, assay type, and task structure. Tasks span mechanism-of-action and pharmacodynamic reasoning, compound-target engagement, causal target validation, developability and safety, and translational efficacy. Other parts of the therapeutic ecosystem, including clinical program stages and non-small-molecule modalities, are left to future TherapeuticsBench efforts. The benchmark offers a measuring stick by which scientists can evaluate agentic systems on local preclinical pharmacology decisions toward practical deployment in drug-discovery programs.

Benchmark Construction

Agents are evaluated on realistic preclinical pharmacology decisions over 100 evaluations spanning eight program stages: disease and model context, screening and hit confirmation, drug response/pharmacogenomics, causal target validation, MoA/PD, compound-target characterization, developability/safety/pharmacokinetics (PK), and translational efficacy. Human genetic target-support tasks are reserved for future benchmarking work. Across the benchmark, evaluations draw on assay types including drug-response screens, genetic perturbation, transcriptomic and proteomic profiling, chemoproteomic and live-cell target engagement, imaging, drug metabolism and pharmacokinetics (DMPK) and absorption, distribution, metabolism, excretion, and toxicity (ADMET) panels, toxicogenomics, and in vivo efficacy or safety studies. These assay families build on established measurement strategies for pooled drug response, growth-rate correction, genetic dependency, transcriptional profiling, proteomics, target engagement, toxicogenomics, and safety pharmacology [15–27,34,36,39,40].

The benchmark is built around recurring local decisions in small-molecule programs. Candidate tasks begin as specific analysis points where assay data might be turned into decisions. Examples include identifying a pathway mechanism from dose-resolved phosphoproteomics, separating true chemoproteomic target engagement from complex contaminants, and interpreting in vivo safety data.

Each evaluation includes the data artifacts needed for that decision, metadata needed to interpret the files, a task description, and a structured answer. Task context is calibrated to approximate what a scientist would know at the point of analysis. We avoid prescribing a method unless the method is part of the scientific decision. The intended answer is a result supported by the supplied data, not some fact or mechanism well understood in the literature. Many tasks are designed to intentionally trap systems that overfit on trained knowledge without empirical exploration.

Candidate evaluations are refined through manual review and model trajectory inspection. We remove tasks if the answer can be recovered by an obvious shortcut that does not require interaction with data, when the prompt over-specifies the analysis, or when plausible analysis choices produce answers that break grading assumptions. The retained tasks are tagged by program stage, assay type, and task structure.

Grading uses deterministic functions over structured final answers. Depending on the task, graders check label sets, numerical tolerances, rank-order relationships, categorical choices, or all required fields. Manual trajectory inspection is part of benchmark construction, following the same broad verifiable-benchmark principle used in related biology-agent evaluations [8–10]. We track decision-blocking gaps that should change the program decision if recognized: tox-species mismatch, insufficient free-drug coverage, hook effects, contaminated chemoproteomic signals, survivor bias in pathology sampling, and related failure points. These annotations are not exposed to agents; they are used to interpret failures after endpoint grading and to guide future benchmark updates.

Benchmark Anatomy

TxBench-PP is organized around the anatomy of a practical preclinical pharmacology decision. Each evaluation is tagged along three axes: where the decision sits in a small-molecule program, the assay type that generated the data, and the structure of the task.

The program-stage axis follows the labels shown in Figure 2: disease and model context, screening and hit confirmation, drug response/pharmacogenomics, human genetic target support, causal target validation, MoA/PD, compound-target characterization, developability/safety/PK, and translational efficacy. Human genetic target support is reserved for future benchmarking work because adequately measuring this category requires a separate focused benchmark. The order gives readers context for task progression; it does not imply that every small-molecule program proceeds through this strict sequence in practice [3,4,28,29].

The assay and task axes describe the actual work required. Assay labels collapse source kits into measurement families such as drug-response screens, protein or post-translational modification (PTM) measurements, genetic perturbation, target engagement, molecular state, safety profiling, and in vivo experiments. Task labels describe broad families of decisions: mechanism inference, quality control (QC), dependency calls, differential response, safety assessment, imaging, hit prioritization, and related judgements.

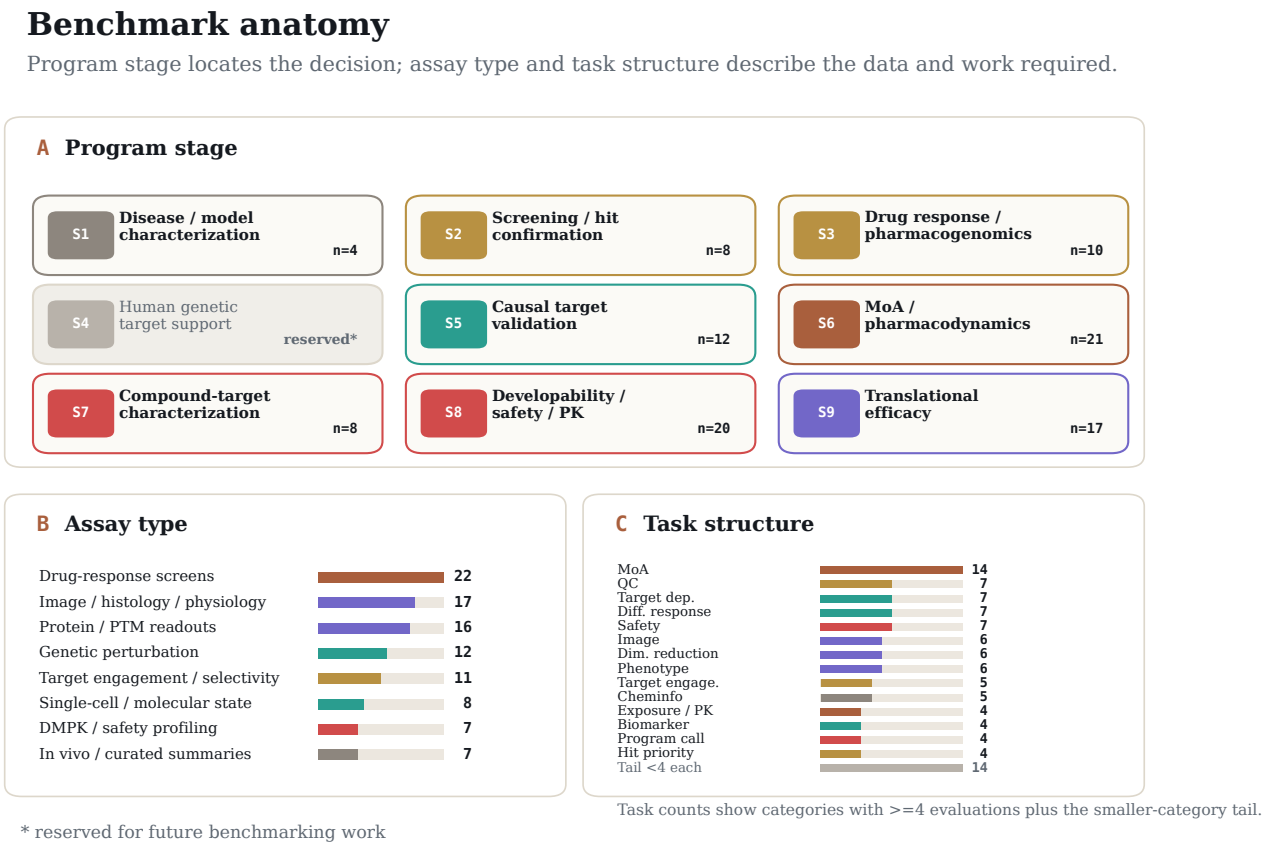

Figure 2: Benchmark anatomy. **A**, Program stage labels place evaluations along a roughly ordered small-molecule evidence spine; the asterisk marks a stage reserved for future benchmarking work. **B**, Assay-type labels summarize the measurement families represented in the source data. **C**, Task-structure labels describe the operation or judgement required for a passing answer.

Results

Top systems remain below 60%

We evaluated 16 model–harness configurations, comprising 11 models across three agent harnesses, on 100 preclinical pharmacology tasks. Each configuration was run three independent times per task, yielding 4,800 agent trajectories. A trajectory passed only when its submitted structured answer satisfied every task-specific grader. We report endpoint pass rate as the fraction of passing trajectories, with 95% confidence intervals computed across tasks.

The strongest configuration was Claude Opus 4.8 / Pi at 59.3% (178/300 trajectories; 95% CI, 51.1–67.6), followed by GPT-5.5 / Pi at 55.3% (166/300; 47.0–63.6), Claude Opus 4.8 / Claude Code at 54.7% (164/300; 45.9–63.4), and Gemini 3.5 Flash / Pi at 51.3% (154/300; 42.9–59.8). These confidence intervals overlap, so the benchmark identifies a leading group rather than a single clear winner. Even the top configuration failed 122 of 300 attempts and passed only 41 of 100 tasks in all three replicates. Current agents therefore remain unreliable on local preclinical pharmacology decisions, despite substantial differences in model family and harness implementation.

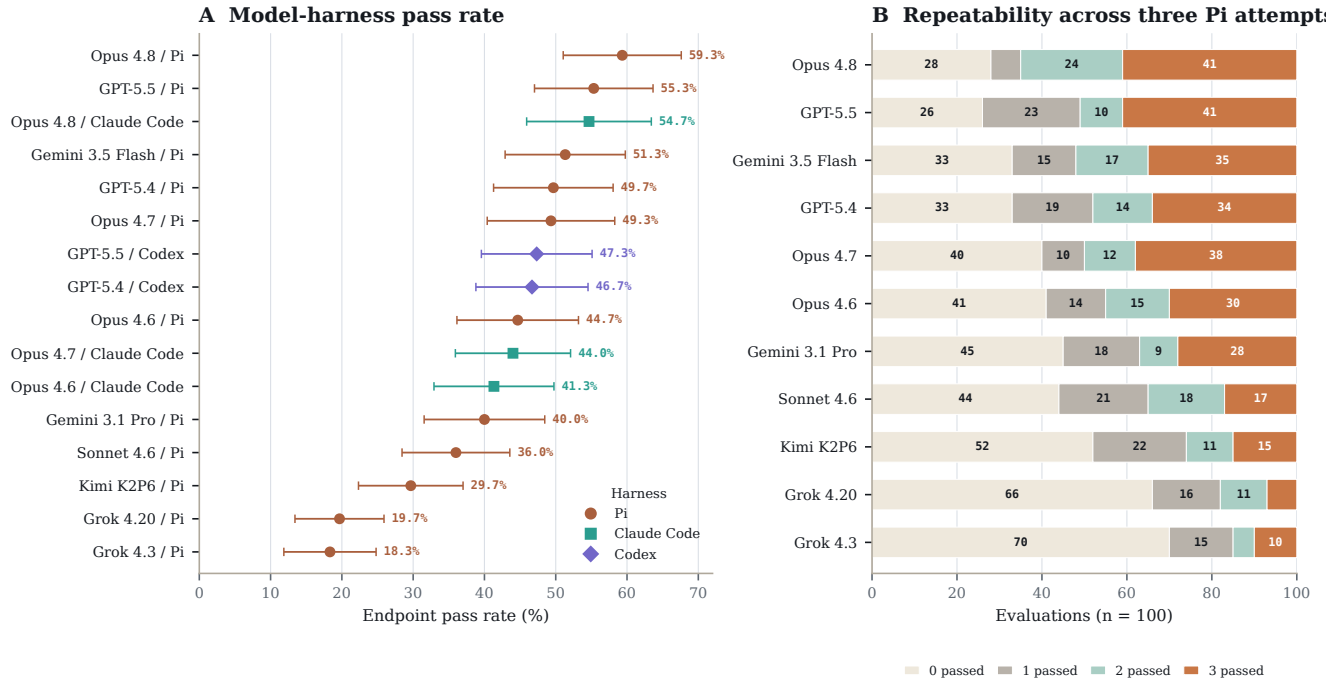

Figure 3: Top systems remain below 60%. **(A)** Endpoint pass rate for each of the 16 model–harness configurations across all 4,800 agent trajectories, with 95% confidence intervals computed over task-level pass rates. Rows are sorted by pass rate; color and marker shape identify the agent harness. **(B)** Task-level reliability for Pi-harness models, partitioning the 100 evaluations by how many of the three independent rounds passed.

Model / harness	Endpoint pass rate	$\geq 1/3$	$\geq 2/3$	$3/3$
Claude Opus 4.8 / Pi	59.3% (178/300; 51.1–67.6)	72/100	65/100	41/100
GPT-5.5 / Pi	55.3% (166/300; 47.0–63.6)	74/100	51/100	41/100
Claude Opus 4.8 / Claude Code	54.7% (164/300; 45.9–63.4)	65/100	59/100	40/100
Gemini 3.5 Flash / Pi	51.3% (154/300; 42.9–59.8)	67/100	52/100	35/100
GPT-5.4 / Pi	49.7% (149/300; 41.3–58.0)	67/100	48/100	34/100
Claude Opus 4.7 / Pi	49.3% (148/300; 40.4–58.3)	60/100	50/100	38/100
GPT-5.5 / OpenAI Codex	47.3% (142/300; 39.6–55.1)	70/100	46/100	26/100
GPT-5.4 / OpenAI Codex	46.7% (140/300; 38.8–54.5)	68/100	46/100	26/100
Claude Opus 4.6 / Pi	44.7% (134/300; 36.1–53.2)	59/100	45/100	30/100
Claude Opus 4.7 / Claude Code	44.0% (132/300; 35.9–52.1)	62/100	45/100	25/100
Claude Opus 4.6 / Claude Code	41.3% (124/300; 32.9–49.7)	55/100	43/100	26/100
Gemini 3.1 Pro / Pi	40.0% (120/300; 31.5–48.5)	55/100	37/100	28/100
Claude Sonnet 4.6 / Pi	36.0% (108/300; 28.4–43.6)	56/100	35/100	17/100
Kimi K2P6 / Pi	29.7% (89/300; 22.3–37.0)	48/100	26/100	15/100
Grok 4.20 reasoning / Pi	19.7% (59/300; 13.4–25.9)	34/100	18/100	7/100
Grok 4.3 / Pi	18.3% (55/300; 11.9–24.8)	30/100	15/100	10/100

Table 1: Pass rates by model–harness configuration. Pass rates are reported as percentages, with the numerator and denominator shown as passing trajectories out of 300 total trajectories per configuration, together with 95% confidence intervals. Confidence intervals were computed across task-level mean pass rates over the 100 preclinical pharmacology tasks. The final three columns report the number of tasks, out of 100, for which at least one, at least two, or all three independent attempts passed.

Trajectory analysis reveals gaps in scientific judgement

We manually reviewed 1,834 failing Pi-harness trajectories. Most failures revealed legitimate gaps in scientific judgement, where models inspected data and performed plausible analyses but ultimately reached incorrect conclusions.

Among failures with an assignable scientific error, method and calibration errors accounted for 71%. Method errors included skipped or inappropriately applied QC, incorrect statistical choices, and the use of an incorrect assay data layer. In one CRISPR-screen task, a model aggregated guides to gene-level scores without first removing guides that mapped to multiple loci, allowing paralog-family artifacts to flood the essential-gene list. In a proteasome-kinetics task, the supported answer lived in the ubiquitin-remnant layer of a dual-readout mass-spectrometry run, but the agent incorrectly used phosphoproteome data. Calibration errors came from cutoffs that were too permissive or too strict for the decision. In one competition-binding panel, models carried forward all 33 nonzero kinases when only the 8 strong binders above a clear affinity gap were supported.

Other failures reflected incorrect image interpretation, memorized literature, or claims not supported by the data. One model computed the strongest data-supported biomarker, a vanadium compound correlated with SLC26A2 at $r = -0.48$, then discarded it for the textbook elesclomol–FDX1 cuproptosis pair. Another reported tumor regression from a sub-baseline median even though the per-animal measurements showed growth arrest rather than regression.

Low step count was concentrated in the two weakest models, at 32% of Grok 4.3's failures and 17% of failures from Grok 4.20 reasoning, against 0–1% for every other model. Above that floor, more effort did not reliably improve accuracy. Gemini 3.5 Flash took the most steps without leading the benchmark, while GPT-5.5 improved on GPT-5.4 with fewer steps. Stronger models usually failed after meaningful tool use.

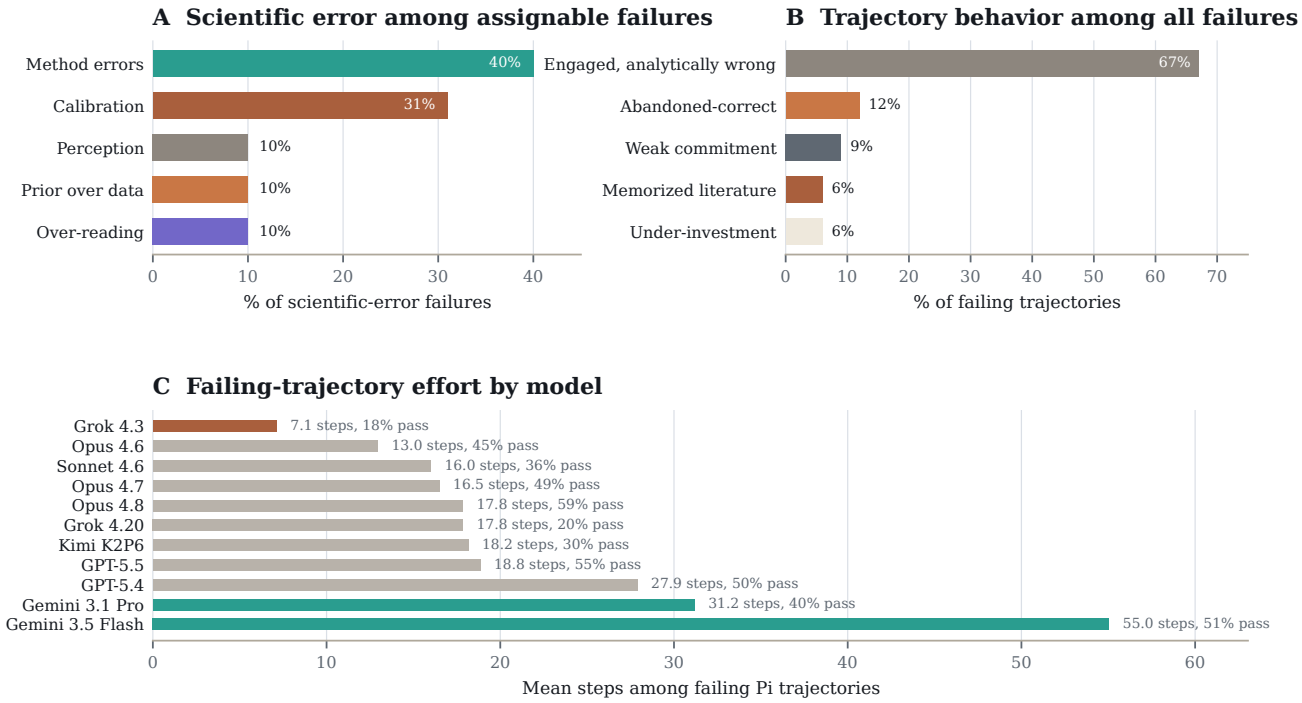

Figure 4: Reviewed failures mostly reflect scientific judgement errors. The 1,834 reviewed failing Pi-harness trajectories were labeled by scientific error and trajectory behavior. (**A**) Among failures with an assignable scientific error, method and calibration errors dominate. (**B**) Most failures followed data engagement and plausible analysis; smaller shares came from abandoned-correct answers, weak commitment, memorized literature, or under-investment. (**C**) Mean step count among failing Pi-harness trajectories by model, with each model's endpoint pass rate shown next to the bar.

Hard tasks require calibrated choices among candidates

Performance varied by program stage, from 27% in screening and hit prioritization to 55% in drug response (Figure 5). Difficult program stages involved decisions across QC, statistics, and chemical or biological judgement of molecular candidates.

Analysis of trajectories helped diagnose model behavior. In screening, models handled mechanical steps such as collapsing noisy replicates (`PRISM05`, 58%), but were unable to apply biological judgement to separate technical artifacts from plausible leads. On outlier-pool QC (`PRISM01`, 18%), models excluded many cytotoxic positive-control wells as artifacts, and on oncology drug repurposing (`PRISM_S2_11`, 0%) they ranked by breadth of killing rather than selectivity. In causal target validation, models recovered clean dependencies but skipped guide-level QC: on multi-target guide contamination (`GG_01`, 16%) they aggregated guides without removing multi-locus single-guide RNAs (sgRNAs). In target engage-

ment, models detected binding signals but lacked a stable cutoff, accepting every weak binder in one kinase-binding panel (`TXBX_KB_v11_DA1`, 0%) and discarding genuinely engaged targets behind a fabricated stringency cutoff in another (`CTRL01`, 21%). In translational efficacy, models consistently struggled to reason accurately about microscopy or organoid images: on a two-marker immunostain (`HAI2027_MM_FIG2E`, 42%) models credited faint, near-vehicle single-agent staining as activity, and in organoid viability tasks apoptotic debris was sometimes counted as living organoids.

Task-type aggregation shows the same structure. Across the 14 task categories with at least four evaluations, the mean Pi-harness pass rate across the 11 models ranged from 70% on exposure/PK tasks to 19% on hit prioritization and 17% on cheminformatics (Figure 6). Exposure/PK, biomarker discovery, safety assessment, and target engagement averaged 57%; these tasks usually ask for a computed value, a binary call, or one supported entity. Target dependency, differential response, hit prioritization, and cheminformatics averaged 24%; these require selecting a defensible subset from many plausible candidates. No model exceeded 45% on hit prioritization or cheminformatics, indicating a shared limit across current systems.

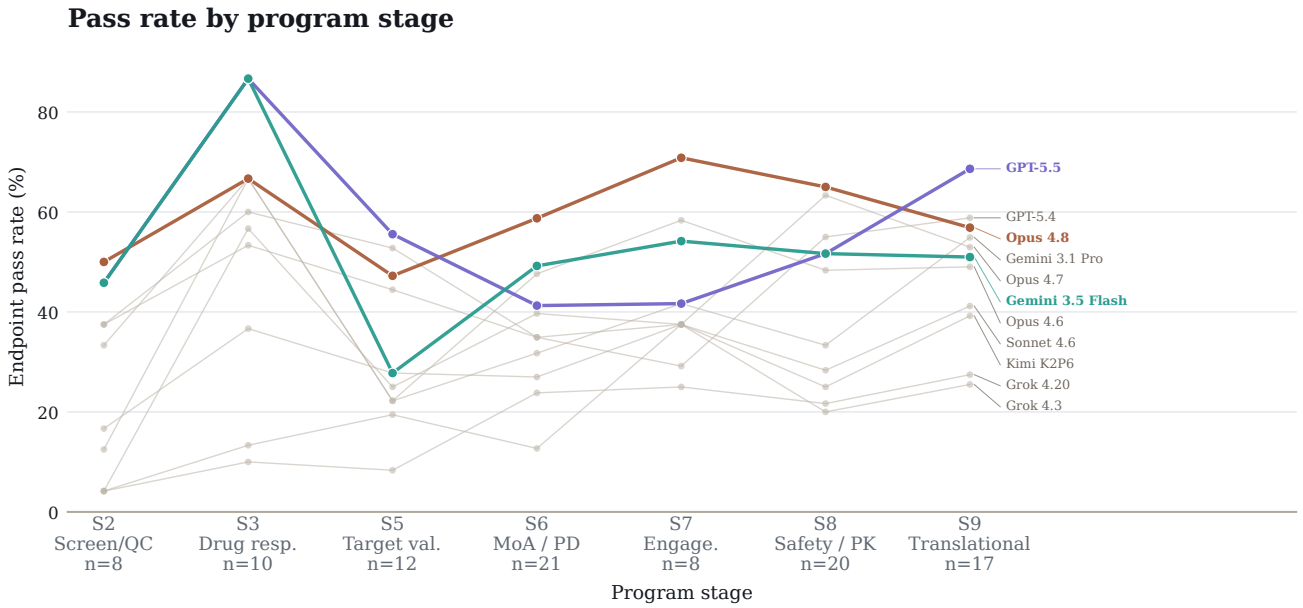

Figure 5: Pass rate by program stage. Endpoint pass rate by program stage across Pi-harness models. S1 disease/model characterization is omitted because it is backed by too few evaluations; S4 human genetic target support is reserved for future benchmarking work.

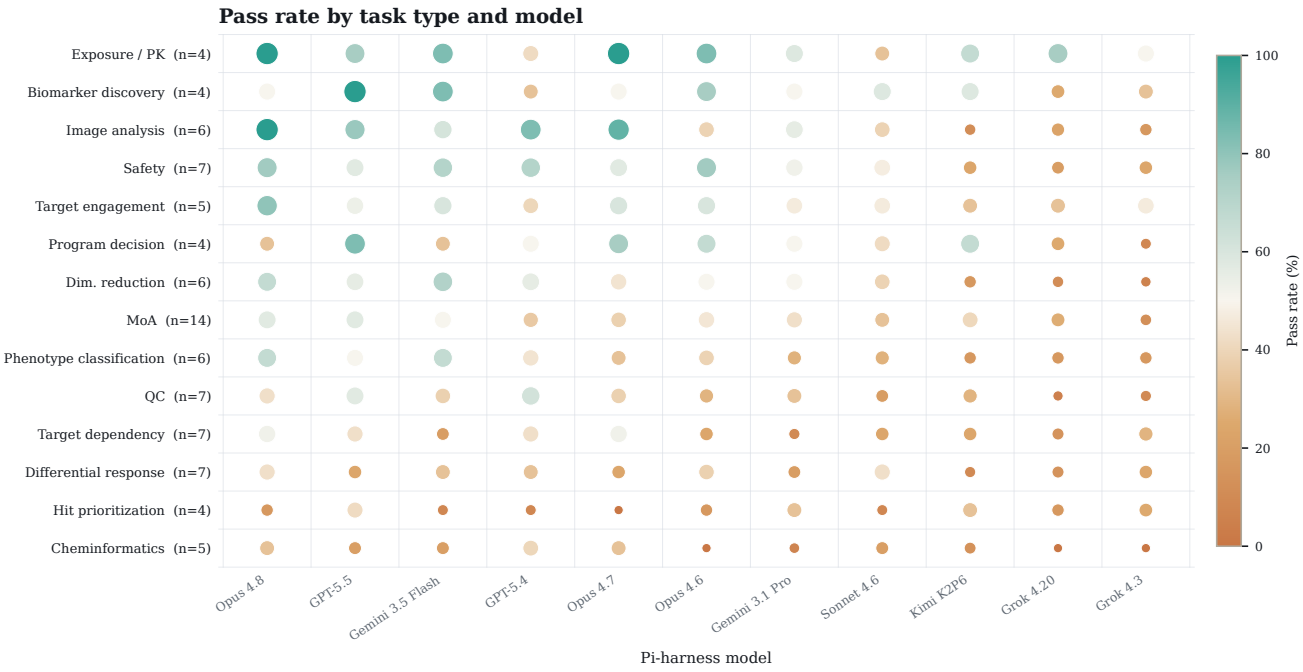

Figure 6: Task-type performance separates determinate calls from candidate selection. Pi-harness pass rate by task type and model for categories with at least four evaluations. Rows are ordered by the mean across the 11 models.

Program decisions expose practical failures

Some evaluations closely resemble real program decisions, giving the agent several assay readouts at once and asking for a single advance, hold, or prioritization call. The current suite includes seven such advancement decisions; across 230 runs, models pass only 35% of them. Models perform well when the data are clean and consistent, but struggle on more challenging cases, either advancing weak candidates or discarding strong ones.

Clean correct. An example organoid program go/no-go task is the clean case. Given a primary screen, a mechanism image, a toxicology counterscreen, and DMPK exposure for eight blinded compounds, the correct action is to advance the one compound that shows the desired cytostatic mechanism together with an adequate safety margin and exposure coverage, and most models do (82%).

Wrong synthesis. On some decision tasks, the agent must assemble a set of experimental models with convergent support across several endpoints, but models incorrectly overfit on a single endpoint (27%), so the program's next experiments are built on wrong assumptions.

False-go. An example safety advancement task requires that no candidate move forward, because none pairs activity with healthy-tissue sparing, yet models advance one on activity alone or on a relative tumor-versus-healthy index (36%), putting an unsafe compound into the pipeline.

False no-go. The reverse error also appears. On an example long-term combination task, models read a strong colony-stain reduction as a reason to drop a candidate, incorrectly culling a program with strong data (`HAI2027_FIG1_DUR01`, 48%).

Strong overall benchmark scores did not reliably predict these program decisions. The six advancement decisions with complete Pi-harness model comparisons yielded an order almost unrelated to overall accuracy (Spearman $\rho = 0.08$; Figure 7A). Claude Opus 4.8, first overall at 59.3%, ranked last on these decisions at 22.2% (4/18).

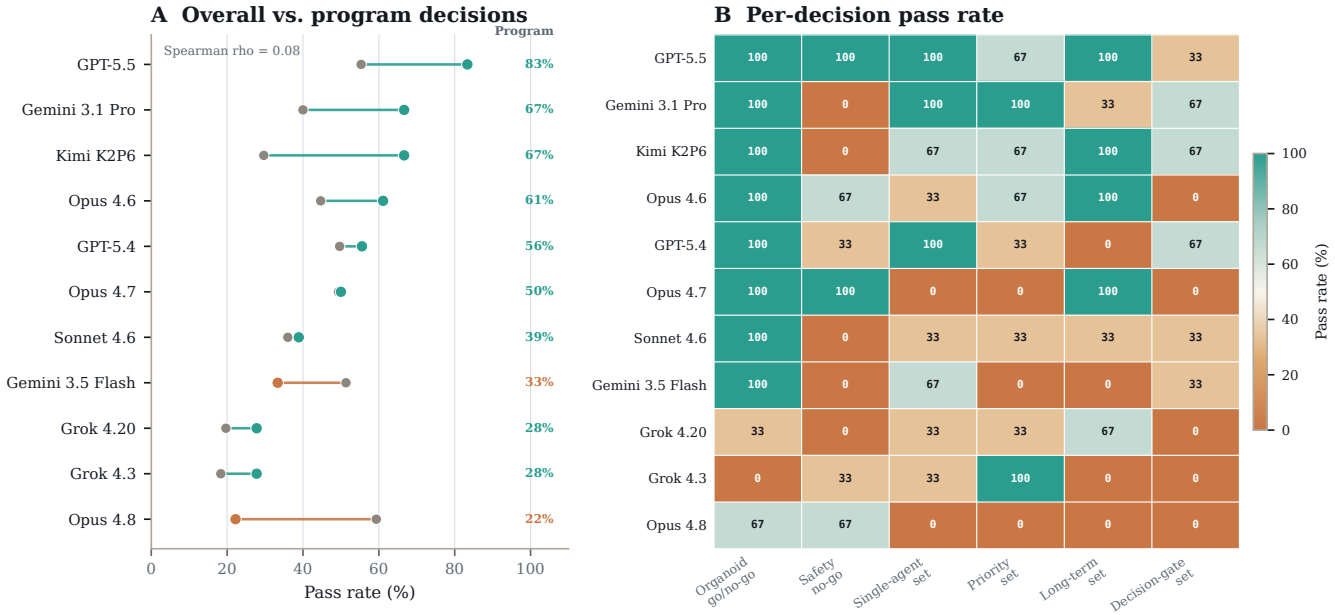

Figure 7: Overall accuracy does not predict advancement decisions. Pass rates on the six advancement decisions with complete Pi-harness model comparisons (up to 18 runs per model). (**A**) Overall endpoint pass rate versus pass rate on the six program decisions, with rows sorted by program-decision performance (Spearman $\rho = 0.08$). (**B**) Per-decision pass rate by model. The two single-compound calls are passed broadly, but tasks that require selecting the supported model set separate the field.

Scores depend on the agent harness

Each harness used the same model checkpoint, task text, and answer-submission protocol, but pass rate still varied by harness. In matched comparisons within the same model family, Pi outperformed Claude Code by 4.4 percentage points (95% CI: 1.2–7.8) on the shared Claude Opus models and outperformed OpenAI Codex by 5.5 percentage points (95% CI: 1.2–9.8) on the shared GPT models. Pi scored higher on every shared model, not only on average.

The size of this difference is comparable to improvements in model generations within families. Moving from Claude Opus 4.6 to Claude Opus 4.7 improved Pi performance by 4.6 points, similar to the harness effect. The Pi advantage was also similar on image-based and non-image tasks, so it is not explained only by differences in image reasoning. Claude Code rendered figures to the model in 100% of image-bearing tasks, compared

with 84% for Pi; OpenAI Codex, which lacks vision capabilities, never rendered figures.

The remaining differences point to implementation details. Pi's lean, three-tool loop failed on 0.6% of runs, compared with 1.8% for Claude Code, and reached answers in fewer steps.

OpenAI Codex was also penalized on figure-dependent tasks because it did not render figures. These effects are small in absolute terms and have wide confidence intervals, but they show that reported scores are properties of the model–harness configuration rather than the model alone.

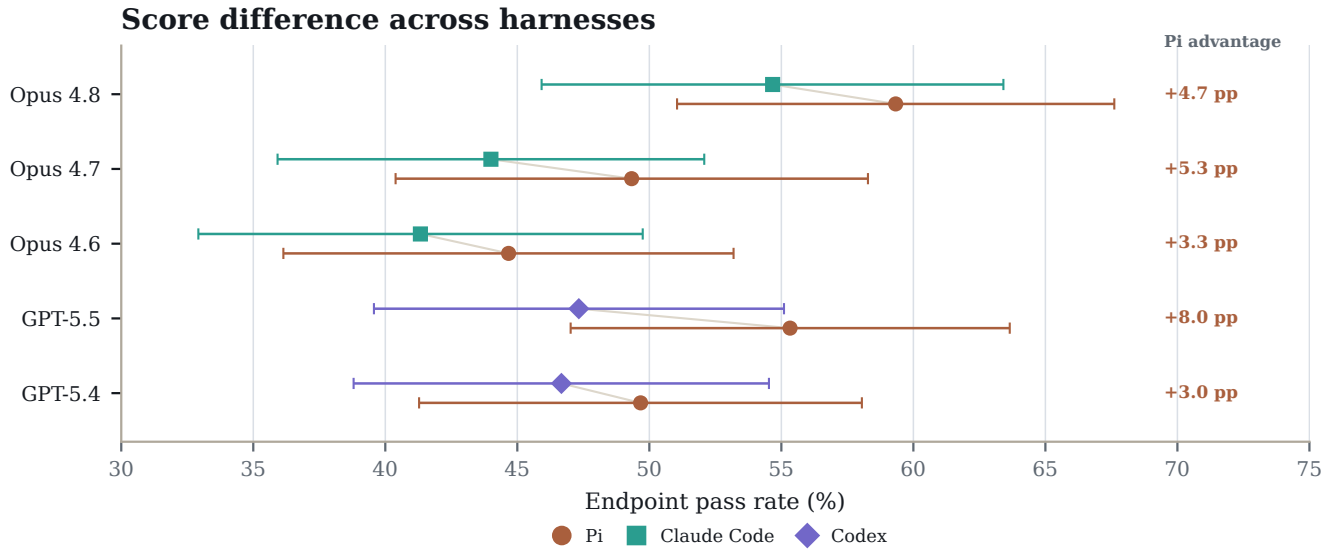

Figure 8: Harness choice changes scores within the same model. Endpoint pass rates for models run under Pi and an alternate harness, with 95% confidence intervals computed over task-level pass rates. Each row compares the same model checkpoint on the same task set; labels at right show the within-model Pi advantage in percentage points. These matched comparisons support interpreting reported scores as properties of the model–harness configuration, not the model alone.

Discussion

TxBench-PP measures a small but relevant category of drug discovery work: whether an agent can make accurate scientific decisions about preclinical pharmacology data with a similar amount of context to what a scientist would have in practice. It is a small-molecule-focused benchmark within a broader TherapeuticsBench roadmap, and should not be used to make more general claims about other areas of the ecosystem, such as human genetic target support, full disease-model selection, broad clinical translation, or non-small-molecule modalities.

The results show agents can complete some tasks correctly but cannot yet be trusted as reliable scientific assistants. Claude Opus 4.8 / Pi was the leading configuration, passing 59.3% of endpoint attempts (178/300; 95% CI, 51.1–67.6), but still failed more than two of every five attempts. Manual trajectory review revealed specific classes of scientific failure. Agents engaged in productive, exploratory behavior but made errors reasoning about QC, statistics, biological context, or molecular properties, or overfit on memorized facts. Local analysis errors can advance weak or unsafe candidates and discard supported ones. The data suggest work remains before these systems can be integrated reliably and reproducibly into preclinical biology workflows.

Future work will extend the benchmark downstream to clinical tasks, vertically to broader pharmacogenomics, and across modalities to antibodies, antibody-drug conjugates (ADCs), degraders, oligonucleotides, cell therapies, and gene therapies. Complementary long-horizon TherapeuticsBench work will follow entire programs across discovery, development, and translation, analogous to long-horizon benchmark designs in other biological domains [8].

Methods

Benchmark assembly

Benchmark construction is described in the Benchmark Construction and Benchmark Anatomy sections. Briefly, we assembled 100 evaluations from practical small-molecule preclinical pharmacology workflow states, using construction principles adapted from prior verifiable biology-agent benchmarks [8–10]. Candidate evaluations were retained when the target decision could be recovered from the supplied data artifacts, expressed through a constrained final answer, and graded deterministically. Tasks were revised or removed when the prompt over-specified the method, when the answer could be recovered by an obvious shortcut, when plausible analysis choices produced incompatible decisions, or when the grader could not distinguish the supported pharmacology conclusion from a plausible but unsupported answer.

Each retained evaluation includes a task prompt, workflow data artifacts, metadata needed to interpret the files, a structured answer schema, a hidden deterministic grader, and tags for program stage, assay type, and task structure. Evaluation notes record the intended empirical decision, known analysis traps, and decision-blocking gaps used for later trajectory interpretation. These notes and failure annotations are not shown to agents. Safety, exposure, toxicogenomics, and cheminformatics evaluations used standard concepts from cardiac safety, free-drug and tissue-exposure reasoning, toxicogenomics resources, molecular fingerprints, and scaffold analysis [34–42].

Agent runs

We evaluated 16 model–harness configurations across Pi, Claude Code, and OpenAI Codex. Pi was run with Claude Opus 4.6, Claude Opus 4.7, Claude Opus 4.8, Claude Sonnet 4.6, GPT-5.4, GPT-5.5, Gemini 3.1 Pro, Gemini 3.5 Flash, Kimi K2P6, Grok 4.20 reasoning, and Grok 4.3. Claude Code was run with Claude Opus 4.6, Claude Opus 4.7, and Claude Opus 4.8. OpenAI Codex was run with GPT-5.4 and GPT-5.5.

Each configuration was run on the same 100 evaluations with three independent attempts per evaluation, yielding 300 trajectories per configuration and 4,800 trajectories overall. Runs used the packaged task context, the supplied data artifacts, a coding environment for file inspection and analysis, and the same structured answer-submission protocol. For each run we recorded the final answer, grader output, trajectory, result log, duration, step count, and available cost metadata.

Failed, timed-out, incomplete, malformed, or unparsable runs were retained in the denominator and counted as nonpassing unless the hidden grader emitted a literal boolean passing result. The result parser treated missing pass fields or non-boolean pass fields as failures.

Endpoint grading

The primary benchmark score is deterministic endpoint grading of the final structured answer. Graders parse the submitted answer and compare the requested fields against task-specific criteria. Depending on the evaluation, these criteria include exact or normalized label matches, numerical tolerances, rank-order checks, categorical choices, and all-of field comparisons. A run passes only when all required task-specific checks pass. Missing fields, invalid JSON, answers outside the expected schema, and answers that cannot be parsed by the grader are counted as failures.

We also report task-level robustness counts for each model–harness configuration: the number of evaluations for which at least one, at least two, or all three replicate attempts passed. These counts are diagnostic and do not replace the endpoint pass rate.

Result aggregation and statistical analysis

For each model–harness configuration and evaluation, the three replicate passes were averaged to produce an evaluation-level pass score in $\{0, 1/3, 2/3, 1\}$. The primary endpoint pass rate is the mean of these 100 evaluation-level scores. We report raw pass counts for interpretability, but compute confidence intervals over evaluations rather than over individual trajectories, because replicate attempts on the same evaluation are not independent samples of the benchmark. Unless otherwise stated, 95% confidence intervals are Student t intervals over evaluation-level mean pass scores.

Stage, assay, and task-type analyses use the same evaluation-level aggregation within each metadata-defined subset. Task-type plots are restricted to categories with at least four evaluations. Cost, duration, and step summaries are averaged first across the three replicate attempts for each evaluation and then across evaluations. Cost analyses use the recorded total-cost metadata when available; for Anthropic model runs without a total-cost field, cost was computed from recorded token usage and the same model-specific price table used for aggregation.

Harness comparisons are matched by model and evaluation. The Claude-family comparison uses Claude Opus 4.6, Claude Opus 4.7, and Claude Opus 4.8 on Pi and Claude Code. The GPT-family comparison uses GPT-5.4 and GPT-5.5 on Pi and OpenAI Codex. Program-decision analyses use the curated subset of evaluations that ask for an advance, hold, prioritize, or related preclinical program decision from multiple evidence streams; the model-comparison panel uses the six such decisions with complete Pi-harness comparisons across the 11 Pi models.

Manual trajectory review

Manual trajectory inspection was used during benchmark construction and for post hoc failure analysis. During construction, trajectories were inspected to identify prompts that

over-specified the method, shortcut-solvable tasks, unstable answer surfaces, and graders that failed to separate supported answers from plausible but unsupported ones.

For the failure taxonomy, reviewers labeled 1,834 failing Pi-harness trajectories against the evaluation's documented scientific intent. Each reviewed failure received a behavioral-cause label describing how the model failed. Failures with an assignable scientific issue also received a scientific-error label describing what was wrong with the submitted answer. These labels were used only for diagnostic analysis after endpoint grading.

Data availability

The public TxBench-PP repository is available at https://github.com/latchbio/txbench-pp. It contains a subset of evaluations and trajectories for manual inspection, along with task definitions, graders, and available data artifacts for those evaluations. Source assay data are distributed or linked according to the original data licenses and access terms.

References

[1] Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* 11, 191–200 (2012). doi: 10.1038/nrd3681.

[2] Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* 3, 711–716 (2004). doi: 10.1038/nrd1470.

[3] Cook, D. et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery* 13, 419–431 (2014). doi: 10.1038/nrd4309.

[4] Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *British Journal of Pharmacology* 162, 1239–1249 (2011). doi: 10.1111/j.1476-5381.2010.01127.x.

[5] Tuntland, T. et al. Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at Novartis Institute of Biomedical Research. *Frontiers in Pharmacology* 5, 174 (2014). doi: 10.3389/fphar.2014.00174.

[6] Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18, 463–477 (2019). doi: 10.1038/s41573-019-0024-5.

[7] Hasselgren, C. & Oprea, T. I. Artificial Intelligence for Drug Discovery: Are We There Yet? *Annual Review of Pharmacology and Toxicology* 64, 527–550 (2024). doi: 10.1146/annurev-pharmtox-040323-040828.

[8] Diks, I., Muralidharan, H., Proctor, T. & Workman, K. Verifiable Benchmarking of Long-Horizon Spatial Biology. *arXiv* 2605.28065 (2026). doi: 10.48550/arXiv.2605.28065.

[9] Workman, K., Yang, Z., Muralidharan, H. & Le, H. SpatialBench: Can Agents Analyze Real-World Spatial Biology Data? *arXiv* 2512.21907 (2025). doi: 10.48550/arXiv.2512.21907.

[10] Workman, K., Yang, Z., Muralidharan, H., Abdulali, A. & Le, H. scBench: Evaluating AI Agents on Single-Cell RNA-seq Analysis. *arXiv* 2602.09063 (2026). doi: 10.48550/arXiv.2602.09063.

[11] Laurent, J. M. et al. LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv* 2407.10362 (2024). doi: 10.48550/arXiv.2407.10362.

[12] Mitchener, L. et al. BixBench: a Comprehensive Benchmark for LLM-based Agents in Computational Biology. *arXiv* 2503.00096 (2025). doi: 10.48550/arXiv.2503.00096.

[13] Qu, Y. et al. BiomniBench: Process-level Evaluation of LLM Agents for Real-world Biomedical Research. *bioRxiv* (2026). doi: 10.64898/2026.05.12.724604.

[14] Li, J. & Ho, A. GeneBench: Assessing AI Agents for Multi-Stage Inference Problems in Genomics and Quantitative Biology. *bioRxiv* (2026). doi: 10.64898/2026.04.22.720113.

[15] Yu, C. et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nature Biotechnology* 34, 419–423 (2016). doi: 10.1038/nbt.3460.

[16] Corsello, S. M. et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer* 1, 235–248 (2020). doi: 10.1038/s43018-019-0018-6.

[17] Hafner, M., Niepel, M., Chung, M. & Sorger, P. K. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods* 13, 521–527 (2016). doi: 10.1038/nmeth.3853.

[18] Meyers, R. M. et al. Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics* 49, 1779–1784 (2017). doi: 10.1038/ng.3984.

[19] Pacini, C. et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nature Communications* 12, 1661 (2021). doi: 10.1038/s41467-021-21898-7.

[20] Subramanian, A. et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171, 1437–1452.e17 (2017). doi: 10.1016/j.cell.2017.10.049.

[21] Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* 75, 1895–1904 (2003). doi: 10.1021/ac0262560.

[22] Gerritsen, J. S. & White, F. M. Phosphoproteomics: a valuable tool for uncovering molecular signaling in cancer cells. *Expert Review of Proteomics* 18, 661–674 (2021). doi: 10.1080/14789450.2021.1976152.

[23] Bantscheff, M. et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnology* 25, 1035–1044 (2007). doi: 10.1038/nbt1328.

[24] Klaeger, S. et al. The target landscape of clinical kinase drugs. *Science* 358, eaan4368 (2017). doi: 10.1126/science.aan4368.

[25] Martinez Molina, D. et al. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* 341, 84–87 (2013). doi: 10.1126/science.1233606.

[26] Savitski, M. M. et al. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* 346, 1255784 (2014). doi: 10.1126/science.1255784.

[27] Vasta, J. D. et al. Quantitative, wide-spectrum kinase profiling in live cells for assessing the effect of cellular ATP on target engagement. *Cell Chemical Biology* 25, 206–214.e11 (2018). doi: 10.1016/j.chembiol.2017.10.010.

[28] Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nature Genetics* 47, 856–860 (2015). doi: 10.1038/ng.3314.

[29] Minikel, E. V. et al. Refining the impact of genetic evidence on clinical success. *Nature* 629, 624–629 (2024). doi: 10.1038/s41586-024-07316-0.

[30] Ioannidis, J. P. A. Why most published research findings are false. *PLOS Medicine* 2, e124 (2005). doi: 10.1371/journal.pmed.0020124.

[31] Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10, 712 (2011). doi: 10.1038/nrd3439-c1.

[32] Begley, C. G. & Ellis, L. M. Raise standards for preclinical cancer research. *Nature* 483, 531–533 (2012). doi: 10.1038/483531a.

[33] Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology. *eLife* 10, e67995 (2021). doi: 10.7554/eLife.67995.

[34] Redfern, W. S. et al. Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development. *Cardiovascular Research* 58, 32–45 (2003). doi: 10.1016/S0008-6363(02)00846-5.

[35] International Council for Harmonisation. ICH S7B: The nonclinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation) by human pharmaceuticals. ICH Guideline (2005).

[36] Gintant, G., Sager, P. T. & Stockbridge, N. Evolution of strategies to improve preclinical cardiac safety testing. *Nature Reviews Drug Discovery* 15, 457–471 (2016). doi: 10.1038/nrd.2015.34.

[37] Webborn, P. J. H., Beaumont, K., Martin, I. J. & Smith, D. A. Free Drug Concepts: A Lingering Problem in Drug Discovery. *Journal of Medicinal Chemistry* 68, 6850–6856 (2025). doi: 10.1021/acs.jmedchem.5c00725.

[38] Zhang, D. et al. Drug concentration asymmetry in tissues and plasma for small molecule-related therapeutic modalities. *Drug Metabolism and Disposition* 47, 1122–1135 (2019). doi: 10.1124/dmd.119.086744.

[39] Huang, R. et al. Modelling the Tox21 10K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nature Communications* 7, 10425 (2016). doi: 10.1038/ncomms10425.

[40] Igarashi, Y. et al. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research* 43, D921–D927 (2015). doi: 10.1093/nar/gku955.

[41] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50, 742–754 (2010). doi: 10.1021/ci100050t.

[42] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* 39, 2887–2893 (1996). doi: 10.1021/jm9602928.