PLOTGROUND: GROUNDING PLOT DIGITIZATION IN REAL SCIENTIFIC FIGURES AND THEIR SOURCE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific figures often encode quantitative results that are not readily available in machine-readable form, making accurate plot digitization important for verifying and reusing published findings. Yet it remains unclear how accurately current models recover plotted values from real scientific figures, as existing benchmarks rely largely on synthetic charts or cover only a limited range of chart types. We introduce **PlotGround**, an automated pipeline for building plot digitization benchmarks from real scientific figures and their author-released source data. PlotGround maps figures to source tables, identifies reconstructable panels, and generates quantitative questions with source-grounded reference values. We use PlotGround to construct **PlotGround-1k**, a human-verified benchmark of 1,119 questions from 1,066 bioRxiv preprints. Across sixteen multimodal models, the best reaches 87.5% accuracy at a $\pm 5\%$ relative-error tolerance. Tightening the tolerance to $\pm 2\%$ lowers every model's accuracy by 11–24 percentage points, revealing a gap between approximate visual reading and precise quantitative recovery. PlotGround's paired figure–source structure lets us compare how accurately the same values are recovered from figures and from source tables. Providing source tables instead of figures raises a coding agent's accuracy from 90.0% to 97.4% while cutting cost by 72%.

1 INTRODUCTION

Scientific figures typically sit at the end of rich computational pipelines: raw measurements may be filtered, normalized, and statistically summarized before the resulting quantities are rendered visually (Peng, 2011; Sandve et al., 2013; Wilson et al., 2017). Yet when an AI agent reads a paper, pixels are often all it gets. Recovering the underlying numerical values from those pixels, known as plot digitization, is therefore a basic prerequisite for quantitative analysis over scientific figures (Savva et al., 2011; Rohatgi, 2017; Liu et al., 2023a).

The reliability of this step becomes increasingly consequential as multimodal models and coding agents are deployed for scientific tasks. They are expected not only to ingest papers but also to verify reported findings and reproduce analyses (Skarlinski et al., 2024; Siegel et al., 2024; Starace et al., 2025). Figures often contain key evidence supporting a paper's claims (Roberts et al., 2024; Sun & Xiao, 2026), and even small digitization errors can silently propagate into subsequent analysis and conclusions.

Despite these stakes, we still know little about how accurately current models recover numerical values from real scientific figures. Existing digitization benchmarks offer limited guidance here because they have largely relied on synthetically rendered charts (Methani et al., 2020; Hassan et al., 2023; Xia et al., 2025; Berkane et al., 2026) or covered only a limited set of chart types (Luo et al., 2021; Touhidul Islam et al., 2026). Synthetic rendering provides exact ground truth and controllable visual variation, but it does not fully capture the characteristics of real scientific figures, including multi-panel layouts, heterogeneous chart types, dense legends and annotations, and varied axis conventions. As a result, performance on these benchmarks may not reflect performance on figures encountered in scientific papers.

We introduce **PlotGround**, an automated pipeline for constructing plot digitization benchmarks directly from scientific papers and their author-released source data. Our key observation is that many papers already deposit the data behind their figures, but what is missing is an explicit mapping from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

source data to the plotted quantities. Unlike synthetic benchmarks that render numerical data into plots, PlotGround starts from real scientific figures and works backward. It reconstructs the plotted quantities from the corresponding source data to provide grounded reference values. The resulting figure–data pairs also allow us to query the same quantities from either the rendered figure or the source table, isolating how the representation of evidence affects agent performance (Section 4.5).

Building on PlotGround, we construct PlotGround-1k, a human-verified benchmark of 1,119 quantitative questions from 1,066 bioRxiv preprints. Our evaluation shows strong performance among frontier models: GPT-5.6 Sol reaches 87.5% accuracy at a $\pm 5\%$ tolerance, and four of the 16 models exceed 80%. However, accuracy drops by 11–24 percentage points under a stricter $\pm 2\%$ tolerance, indicating a remaining gap in numerical precision. During human verification, we also find 8 figures whose plotted values disagree with those reconstructed from their own deposited source data (Section 4.6). These cases suggest that accurate plot digitization could help flag inconsistencies between scientific figures and their source data.

Our contributions are threefold:

- **A source-grounded construction pipeline.** We introduce PlotGround, an automated pipeline that turns real scientific figures into digitization benchmarks by grounding plotted values in author-released source data.
- **A realistic benchmark with paired evidence representations.** We construct PlotGround-1k, a human-verified benchmark of 1,119 quantitative questions over real scientific figures. Each question is linked to the source table from which its reference value is computed, enabling controlled comparisons between figures and source data as evidence.
- **A characterization of the precision gap in scientific plot digitization.** We find that frontier models perform strongly under a $\pm 5\%$ tolerance, but accuracy drops by 11–24 percentage points at $\pm 2\%$. In an agentic setting, recovering the same quantities from source tables is more precise than estimating them from figures.

2 Related Work

Scientific Agents. Recent work has increasingly explored AI agents for scientific tasks spanning literature synthesis (Asai et al., 2026), research-idea generation (Si et al., 2025; Baek et al., 2025), experiment execution (Boiko et al., 2023; Si et al., 2026), and the replication or verification of published results (Hu et al., 2025; Siegel et al., 2024). As these systems move beyond text-only interaction, they increasingly operate over heterogeneous scientific artifacts, including papers, code, tables, and figures (Khalighinejad et al., 2025; Lu et al., 2026). This creates a particular challenge for verification and reproducibility workflows, where recovering reported quantitative results may require locating and interpreting figures (Siegel et al., 2024). Our work focuses on this interface, studying the extraction of numerical information from scientific figures and how its representation affects downstream agent performance.

Plot Digitization. Plot digitization aims to recover the numerical values encoded by visual marks in a rendered chart, ranging from interactive tools (Jung et al., 2017; Rohatgi, 2017) to fully automated extraction methods (Liu et al., 2023a;b; Yang et al., 2025). However, existing extraction benchmarks often rely on automatically rendered charts or cover only a limited set of chart types (Hassan et al., 2023; Xia et al., 2025). PlotQA uses real-world tabular data but automatically generates the chart images (Methani et al., 2020); ExcelChart400K uses authentic charts collected from public Excel files but covers only bar, line, and pie charts (Luo et al., 2021). Newer benchmarks have begun to move toward real-world figures, but important limitations remain. ChartZero constructs a 1,000-chart benchmark mined from heterogeneous PDFs but restricts evaluation to line charts (Touhidul Islam et al., 2026). SPUR and CharXiv evaluate models on real scientific figures, but through multiple-choice or descriptive and reasoning questions rather than open-ended recovery of plotted values (Wang et al., 2024; Ding et al., 2026). Our benchmark addresses these gaps by evaluating open-ended numerical recovery across diverse real scientific figures, with ground-truth values derived from author-released source data.

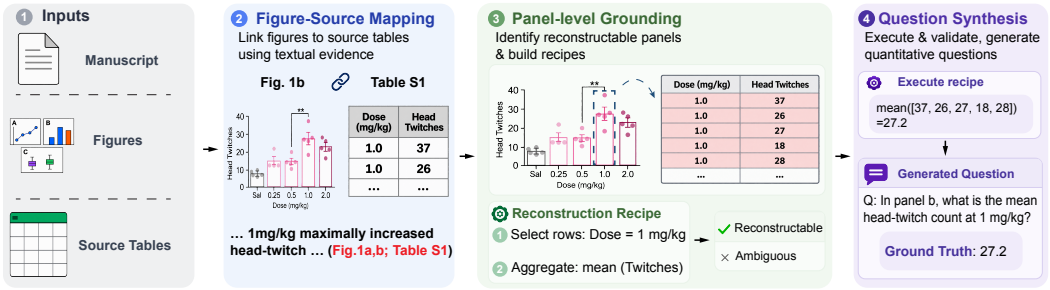


Figure 1: **Overview of the PlotGround pipeline.** Given a bioRxiv paper organized as a filesystem (manuscript, figures, and supplementary source tables), **Mapping** proposes candidate figure–table bindings from textual evidence such as captions and in-text references; **Grounding** retains only panels whose values can be reliably reconstructed and records a reconstruction recipe, rejecting ambiguous ones; **Synthesis** follows the recipe, validates the recomputed values, and writes questions using only figure-visible labels. For clarity, the example shows a single panel with question text abbreviated. Real figures are typically multi-panel and are given to models in full (Appendix A.2).

3 PLOTGROUND

3.1 PIPELINE OVERVIEW

PlotGround turns a scientific figure and its corresponding source data into grounded quantitative questions. Each question refers to a visible element in the figure, such as a bar, labeled point, or box, and asks for its numerical value. The answer is recomputed from the underlying source table rather than read from the rendered image. We evaluate digitization at the level of **individual plotted values** rather than complete data series, allowing each answer to be verified against its source data and scored independently.

PlotGround operates in three stages (Figure 1). **Mapping** identifies candidate source tables for each scientific figure. **Grounding** examines each mapped figure at the panel level, retaining only panels whose plotted values can be reliably reconstructed and recording a reconstruction recipe specifying the operations needed to recover them. **Synthesis** follows these recipes to compute reference values and generates panel-specific quantitative questions.

All three stages are executed by Claude Code (Anthropic, 2026c) using its default tool set, which allows the agent to navigate and read files, inspect spreadsheets, and execute code. We use Claude Opus 5 for the grounding stage and Claude Sonnet 5 for mapping and synthesis. Each paper is processed in an isolated working directory containing only the files associated with that paper.

3.2 FIGURE–SOURCE MAPPING

The mapping stage identifies candidate supplementary source data for each scientific figure. The agent navigates the working directory and infers figure–source correspondences from the manuscript, captions, and spreadsheet summaries.

The agent proposes figure–source bindings supported by either *direct* evidence, such as an explicit textual reference connecting a figure to a supplementary file, or *indirect* evidence, such as a match between the figure description and the contents of a candidate table. When the available evidence explicitly identifies a specific panel, the correspondence may be resolved at the panel level; otherwise, it is retained at the figure level and further disambiguated during grounding.

3.3 PANEL-LEVEL GROUNDING

Given a candidate figure–source mapping, the grounding stage determines which panels can be reliably reconstructed from the linked source data. Grounding is intentionally conservative: a panel is retained only when its plotted quantity can be recovered with simple, well-defined operations and when the correspondence between visual elements and source-table entries is unambiguous.

The stage first applies a *reliability gate*. We retain quantitative charts whose plotted values can be recovered from the source table using basic operations such as selection, aggregation, ratios, and logarithmic transformations. Panels are rejected when reconstruction requires domain-specific analyses, lacks figure-visible identifiers for question construction, or leaks exact values through printed annotations.

For each retained panel, the agent maps visual groups to explicit rows or columns in the source table and records a structured reconstruction recipe specifying the relevant sheet, columns, row filters, and any aggregation or normalization. Every mapping must be anchored to literal evidence, such as a table header, a category-column value, or a token visible in the figure. Mappings based on domain priors, conventional ordering, or sample-ID ordinals are disallowed. Panels with ambiguous column choices (e.g., raw versus adjusted p-values), unidentified normalization references, or incomplete source-data coverage are also rejected.

3.4 Question Synthesis

The synthesis stage follows the recorded reconstruction recipe on the source table to recompute the plotted values. It then matches these values to the individually readable elements of each panel: bar heights, line or survival values at labeled x positions, the marked statistics of a box plot (median and quartiles), and the coordinates of in-figure-labeled points in scatter and volcano plots.

Before generating questions, it checks that the recipe's assumptions are supported by the source file or paper rather than by domain conventions alone. It also verifies that the recomputed values satisfy basic constraints of the plotted quantity (e.g., survival probabilities must be non-increasing over time). As a final guard against incorrect figure–source mappings, the agent reads each target off the figure and discards it only when the reading differs from the recomputed value by a large margin (at least three times the $\pm 5\%$ tolerance used for evaluation). Among the elements that pass these checks, it selects the clearest target in each panel, favoring large, uncrowded marks.

Each question asks for a single numerical value and names its target using only identifiers visible in the figure, such as the panel label, an x-tick or legend entry, and the axis label and units, e.g., "In panel b, what is the mean head-twitch count at 1 mg/kg?" (Figure 1). This ensures that questions can be answered from the figure alone without access to the source data. To reduce redundancy, synthesis emits at most two questions per figure, drawn from different panels. Additional examples, shown with their full multi-panel figures and spanning different chart types, are provided in Appendix A.2.

4 Experiments

4.1 Benchmark Construction

Paper selection. We build PlotGround on the 2025 bioRxiv corpus available through Paperclip[1], an agent-native interface to scientific literature that exposes papers in a filesystem-like environment. This representation gives coding agents direct access to the components of a paper, including the manuscript, figures, and deposited supplementary files, and allows them to navigate and reason over these artifacts programmatically. We focus on bioRxiv because source-data deposition is common in biology, making it possible to link published figures to the underlying tables from which their values can be reconstructed.

We require each paper to contain at least one machine-readable tabular supplement (`.xlsx`, `.xls`, `.csv`, or `.tsv`), as these formats provide structured source data that is well suited to direct figure reconstruction. Among 37,635 preprints with non-empty supplements, 10,540 satisfy this criterion, and all of them enter the pipeline.

Pipeline funnel. The corpus then passes through the three stages in Sections 3.2–3.4. *Mapping* identifies 28,536 figure–source bindings across 8,197 papers. To control cost, we keep at most one mapping per paper and randomly sample 7,487 of the 8,197 papers for downstream processing. *Grounding* evaluates the sampled mappings, covering 22,541 figure panels in total. Among them, 3,894 panels from 1,568 figures admit a valid reconstruction recipe. *Synthesis* generates at most

[1] `https://paperclip.gxl.ai/`

two questions per figure, and figures that do not admit a suitable question yield none. When a figure has two questions targeting panels of the same chart type, we remove one at random, leaving 1,205 candidates. We manually audit all of them and retain 1,119 verified questions from 1,066 papers, which constitute **PlotGround-1k** (see Appendix A.3 for the verification protocol). Table 1 summarizes the resulting counts.

Table 1: PlotGround pipeline funnel. Counts are given in the unit of each stage (papers, mappings, panels, or questions), since one paper can yield multiple figure–source mappings and one figure multiple panels.

Stage	Step	Retained
Collection	Index 2025 preprints with supplements	37,635 papers
Filtering	Require a tabular supplement	10,540 papers
Mapping	Bind figures to source tables (Section 3.2)	28,536 mappings (8,197 papers)
Downsampling	Randomly sample papers, one mapping per paper	7,487 mappings
Grounding	Admit reconstructable panels (Section 3.3)	3,894 panels (1,568 figures)
Synthesis	Generate and filter questions (Section 3.4)	1,205 questions
Verification	Human audit of all candidates	**1,119 questions** (1,066 papers)

4.2 Experimental Setup

We evaluate sixteen frontier multimodal models on PlotGround-1k under a shared protocol. Given the full figure image and a quantitative question targeting a specific panel (see Appendix C.1 for the full prompt), each model predicts the corresponding numerical value from the figure alone. We convert all figures to PNG, cap the longer edge at 1,600 pixels, and preserve the original aspect ratio. Following the relaxed accuracy criterion used in PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022), we measure accuracy at a relative-error tolerance of $\pm 5\%$ and additionally at a stricter $\pm 2\%$ tolerance to assess numerical precision. We also report the average cost and output-token count per question.

Models. We evaluate sixteen models spanning nine developers: OpenAI's GPT-5.6 Sol, GPT-5.6 Luna, and GPT-5.6 Terra (OpenAI, 2026); Google's Gemini 3.8 Flash, Gemini 3.7 Flash, and Gemini 3.1 Pro (Google DeepMind, 2026a;b;c); Anthropic's Claude Opus 5 and Claude Sonnet 5 (Anthropic, 2026a;b); Alibaba's Qwen3.8 Max and Qwen3.8 Flash (Qwen Team, 2026; QwenCloud, 2026); Moonshot's Kimi K3 (Kimi Team, 2026); Z.ai's GLM-5.3 Flash (Z.ai, 2026); MiniMax's M3 (MiniMax, 2026); DeepSeek's V4.1 Flash and V4 Flash Vision Exp (DeepSeek, 2026a;b); and Thinking Machines Lab's Inkling (Thinking Machines Lab, 2026). Each model is evaluated independently through OpenRouter[2] with its default settings. We do not override sampling parameters, output limits, or reasoning configuration. All models returned a numeric answer for every question.

4.3 Main Results

Table 2 shows that several frontier models perform strongly on PlotGround-1k: at a $\pm 5\%$ relative-error tolerance, four models exceed 80% accuracy, and GPT-5.6 Sol leads at 87.5%. At the stricter $\pm 2\%$ tolerance, accuracy drops by 11–24 percentage points across models, with GPT-5.6 Sol falling to 73.8%. The overall ranking remains broadly stable, suggesting that relative model performance is robust to this change in tolerance. The drop still matters when extracted values are reused in downstream calculations: errors can affect derived quantities such as ratios or fitted parameters, so precision sufficient for approximate inspection may not suffice for quantitative reuse. A breakdown of accuracy by chart type is provided in the Appendix B.1.

Higher accuracy does not require higher inference cost (Figure 2). GPT-5.6 Luna achieves 83.8% accuracy at 0.113¢ per question, approximately 4% of the cost of Claude Opus 5 while scoring 6.7 percentage points higher. The two most expensive models, Kimi K3 and Gemini 3.1 Pro, both rank in the bottom half despite averaging 5,545 and 3,337 reported output tokens per question, respectively.

[2] https://openrouter.ai/

Table 2: Model accuracy on PlotGround-1k (1,119 questions). A prediction counts as correct if its relative error with respect to the gold value is within the tolerance. All models were run with their default settings, including reasoning configuration. Cost and output tokens are averaged over all questions; output tokens include reasoning tokens where the provider reports them.

Model	Acc. (±2%)	Acc. (±5%)	Avg. Cost (¢)	Avg. Output tokens
GPT-5.6 Sol	**73.8**	**87.5**	1.013	419
GPT-5.6 Luna	69.7	83.8	0.113	447
GPT-5.6 Terra	66.6	82.1	0.868	231
Gemini 3.7 Flash	65.7	81.2	0.428	909
Gemini 3.8 Flash	65.5	79.8	0.933	2,257
Qwen3.8 Max	57.7	77.9	2.945	4,269
Claude Opus 5	60.2	77.1	2.724	564
Claude Sonnet 5	55.6	74.4	0.631	105
Gemini 3.1 Pro	49.9	67.6	4.237	3,337
Qwen3.8 Flash	42.6	64.4	0.223	4,111
Kimi K3	38.2	62.5	8.371	5,545
GLM-5.3 Flash	39.8	60.2	0.105	1,495
Inkling	32.0	50.3	0.866	1,633
MiniMax M3	33.8	50.3	0.102	358
DeepSeek V4.1 Flash	29.3	47.0	0.131	904
DeepSeek V4 FV exp	17.0	28.6	0.261	3,177

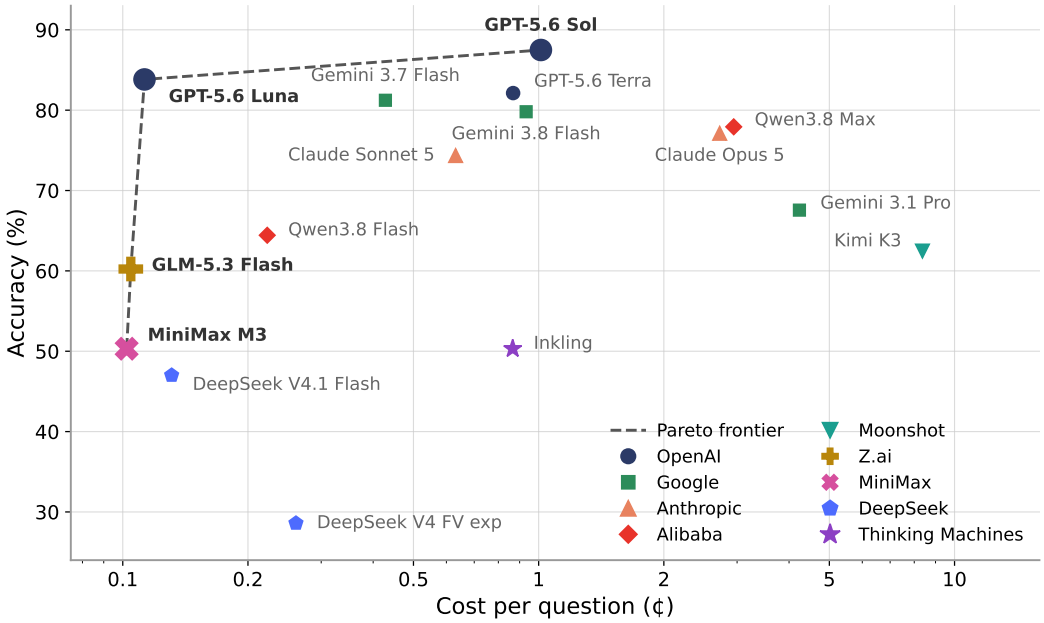


Figure 2: **Accuracy versus inference cost on PlotGround-1k.** Each point represents a model, with colors and marker shapes indicating developers. Accuracy is measured at a ±5% tolerance, and mean inference cost per question is shown in cents on a logarithmic scale. The dashed line connects Pareto-optimal models, whose labels are shown in bold. Higher cost does not necessarily yield higher accuracy: GPT-5.6 Luna achieves 83.8% accuracy at 0.11 cents per question, compared with 62.5% at 8.37 cents for Kimi K3.

4.4 Qualitative Error Analysis

To better understand where current models fail, we manually inspect representative incorrect predictions across the evaluated models. We observe several recurring error patterns that are not apparent

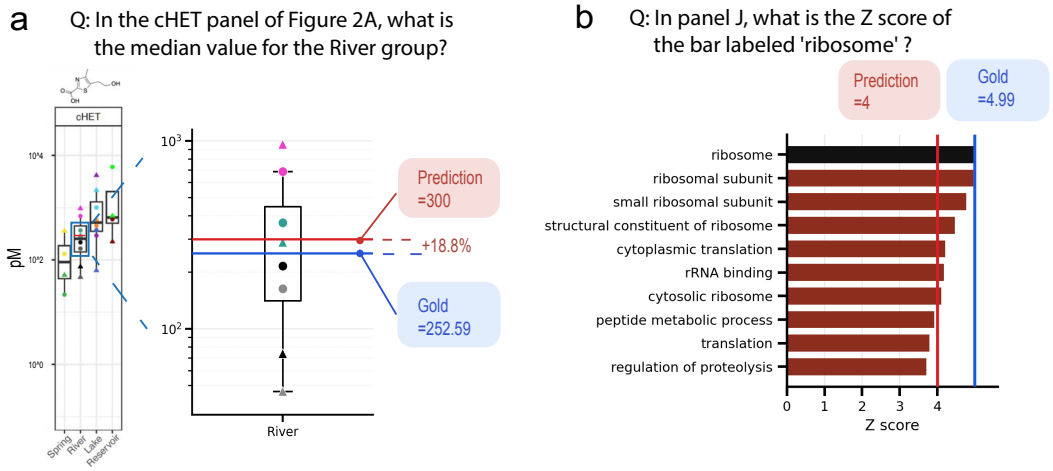


Figure 3: **Representative qualitative failure cases.** (a) On a log axis, a visually small discrepancy in position can correspond to a substantial numerical error; the model predicts 300 for a gold value of 252.59. (b) Models often snap predictions to labeled tick marks rather than interpolate continuously; for a bar with gold value 4.99, eight of sixteen models predict the final labeled tick at 4.

from aggregate accuracy alone. In particular, many failures arise from converting visual position into precise numerical values, especially when interpolation between labeled ticks is required.

Logarithmic axes. Reading a value from a logarithmic axis requires geometric interpolation between labeled ticks, which is a frequent source of error among the cases we inspected. Models generally recognize that the axis is logarithmic, but struggle to convert a target's visual position into the correct numerical value when it falls between labeled ticks. On a log scale, small positional errors can translate into much larger relative numerical errors because equal distances correspond to multiplicative rather than additive changes. For example, predicting 300 for a gold value of 252.59 yields a relative error of approximately 19% (Figure 3a), despite the prediction appearing visually close to the plotted point.

Snapping to tick marks. Predictions often concentrate on labeled tick values rather than around the true value. In a horizontal bar chart (Figure 3b) whose x-axis is labeled at 0, 1, 2, 3 and 4, the target bar has a gold value of 4.99 and visibly extends past the final gridline. Eight of sixteen models nevertheless answered exactly 4, the last labeled tick. Only three read the overshoot and answered 5, two gave intermediate values (4.2 and 4.5), and the remaining three were further off (3, 20, and 40). This behavior is consistent with models estimating the target relative to the nearest tick and then reporting that tick, or a nearby round number, rather than interpolating continuously between ticks. This suggests that accuracy may depend on where the gold value lies relative to the axis labels: values close to ticks are often recovered correctly, whereas values between ticks are more error-prone.

4.5 Effect of Evidence Representation

The evaluations above measure plot digitization in its purest form: a model is shown a figure and a question, without access to the surrounding paper or source data. Recent work on scientific agents, however, has begun to move beyond the paper as a single static document, exposing agents to broader collections of research artifacts (Miao et al., 2025; Liu et al., 2026). In such settings, the same quantitative evidence may also be available in structured source data. This raises a basic question for agentic scientific reasoning: how much does the representation of quantitative evidence affect an agent's ability to recover and use it?

PlotGround provides a natural way to study this question because every benchmark instance links a plotted quantity to the source table from which its value is reconstructed. We can therefore hold the

Table 3: **Effect of evidence representation on agent performance.** Claude Code (Claude Sonnet 5) answers the same 500 rewritten questions under three evidence conditions: figures, source tables, or both. Accuracy is measured at a $\pm 5\%$ relative-error tolerance; cost and latency are averaged over all questions.

Condition	Accuracy ($\pm$5%)	Avg. Cost ($)	Avg. Latency (s)
Figure (full text + figures)	90.0	0.408	139
Table (full text + source tables)	97.4	0.113	51
Both (full text + figures + source tables)	**98.2**	0.115	51

scientific question fixed while varying whether the agent has access to figures, source data, or both. In particular, we ask what the agent gains when the same quantity can be recovered from structured data rather than estimated from pixels.

Setup. We rewrite questions from PlotGround-1k in the paper's own terms. References to figures, panels, axes, and tables are removed so that the same question can be answered using either the figure or the underlying source data. For example, "In panel A, what is the log2 fold change of the point labeled CHAC1?" becomes "In the RNAseq differential expression analysis (DESeq2, compared to vehicle-treated cells) of HEK293T cells treated for 6 h with MBX-2982 (10 μM), what is the log2 fold change in expression of CHAC1?" (see Appendix C.2 for the rewriting prompt). During rewriting, we also flag and discard questions whose answers are stated directly in the manuscript text. The evaluation set consists of 500 randomly sampled questions that pass this filter.

Claude Code (Claude Sonnet 5) then answers each question under three evidence conditions. Each run takes place in a Docker container holding the paper's files, with the same prompt throughout (Appendix C.3). Only the contents of the working directory differ between conditions. All three conditions include the full text (the article XML from the bioRxiv submission package). The agent additionally receives figure images (**Figure**), supplementary data files (**Table**), or both (**Both**). The manuscript PDF is withheld throughout because it contains the rendered figures. This setting differs from the single-shot protocol of Table 2 in question wording, available context, and image resolution, so the two accuracies are not directly comparable.

Source data is more accurate, more precise, and cheaper. Access to structured source data substantially improves both accuracy and efficiency (Table 3). With the figure alone, Claude Code reaches 90.0% accuracy, but each question costs $0.41 and takes 139 s on average. The agent takes about 26 turns per question, repeatedly cropping and enlarging the target panel and, in most runs, measuring pixel positions directly from the image array. Replacing figures with source tables raises accuracy to 97.4%, while reducing cost by 72% and latency by 63%. Among answers that pass at $\pm 5\%$, the median relative error is 0.39% from the figure against 0.001% from the table.

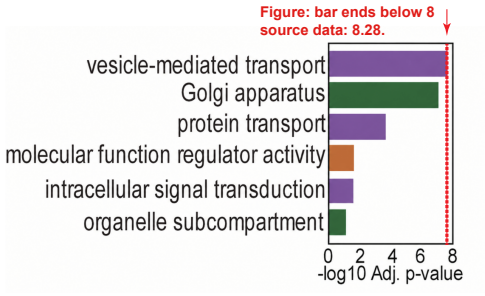
Given both, the agent predominantly relies on the table. Adding figures alongside the tables provides little additional benefit: accuracy increases only marginally to 98.2%, while cost and latency remain nearly identical to the table-only condition. The agent's working traces in the **Both** condition help explain this result. It accesses the source table for 492 of the 500 questions, but inspects a figure image for only 20. Among these 20 cases, 13 also involve table access, usually before the figure is inspected, suggesting that the image is used primarily to confirm a value already obtained from the table rather than to derive it independently. Thus, when both representations are available, the agent overwhelmingly treats structured data as the primary source of quantitative evidence.

4.6 Figure–Data Discrepancies

During verification, we found 8 candidate questions (from 8 figures) with apparent discrepancies between the plotted values and those reconstructed from the associated source data. Our review did not identify errors in the reconstruction recipes that could account for these differences. Examples included a bar ending below 8 despite a reconstructed value of 8.28 (Figure 4), and a survival

curve whose genotype labels appeared inconsistent with the source data (Supp Figure 6). We could not reconcile these differences using the available source data and manuscript descriptions, so we excluded these candidates from PlotGround-1k.

A Published GO enrichment plot



B Deposited source data

–log10 values recomputed from the deposited adjusted p-values.

GO term	Raw source value (Adj. p)	–log10(Adj. p) (reconstructed)
vesicle-mediated transport	5.27×10^{-9}	**8.28**
Golgi apparatus	2.20×10^{-8}	7.66
protein transport	1.02×10^{-4}	3.99
molecular function regulator activity	0.017797	1.75
intracellular signal transduction	0.017841	1.75
organelle subcompartment	0.058711	1.23

Figure 4: **Example of a discrepancy between a published figure and its deposited source data.** The published GO enrichment plot differs from the values reconstructed from the deposited source data. The discrepancy is clearest for the most significant term. The deposited adjusted p-value for vesicle-mediated transport (5.27×10^{-9}) corresponds to $-\log_{10} = 8.28$, but the bar shown for this term ends at approximately 7.7, short of the axis limit at 8. Axis clipping cannot explain the gap, since a clipped bar would extend to the limit.

These cases are a by-product of the pipeline rather than its goal, but they suggest a practical use. Given that frontier models can often recover plotted values accurately (Section 4.3), the same three steps (mapping a figure to its source table, reconstructing the plotted quantity, and comparing it with the value read from the figure) could be used to flag potential figure–data inconsistencies for further inspection. On the 8 discrepant figures, GPT-5.6 Sol's reading agrees with the figure rather than the table in every case. Since a model's own reading errors also trigger such flags, the check currently serves best as a filter that narrows down candidates for human review, and it becomes more precise as reading accuracy improves.

5 Conclusion

We introduced PlotGround, a pipeline for constructing plot digitization benchmarks from real scientific figures and their source data. The resulting benchmark, PlotGround-1k, contains 1,119 human-verified questions. Evaluation of 16 multimodal models reveals a remaining gap between approximate visual reading and precise numerical recovery. In our agentic evaluation, replacing figures with source tables improves accuracy while substantially reducing cost, highlighting the value of structured data for quantitative scientific workflows. We also identify figures whose plotted values disagree with their own deposited data, suggesting figure–data consistency checking as a downstream use of accurate plot digitization.

Limitations. PlotGround-1k is drawn from bioRxiv preprints that deposit tabular source data, so its current coverage is concentrated in biology and may not reflect figure conventions in other fields. However, the pipeline itself is not tied to bioRxiv. Journals that publish per-figure source data, such as those in the Nature Portfolio, could support similar benchmark construction in other disciplines. Our conservative grounding criteria and preference for clearly readable targets help ensure the reliability of the pipeline, but favor figures whose plotted values are relatively easy to verify. Performance on PlotGround-1k may therefore not generalize to the full range of scientific figures. In addition, although the pipeline is automated, PlotGround-1k is fully human-verified (the audit accepted 92.9% of candidates), and reducing the need for such review remains open. Finally, our evaluation focuses on recovering individual plotted values, an atomic operation in full plot digitization. PlotGround's explicit links between figures and source data could facilitate extending this evaluation to complete data series. We leave these extensions to future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, drafting parts of the manuscript and literature discovery. We also used AI agents for benchmark construction and question rewriting, as described in Sections 3 and 4.5. We have not used generative AI tools for research ideation or for proving mathematical claims. We have reviewed all AI-assisted work. In particular, we manually reviewed the benchmark questions, as described in Appendix A.3, and verified that the cited works exist and are accurately represented in the related-work discussion. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Introducing Claude Opus 5, 2026a. URL https://www.anthropic.com/news/claude-opus-5. Accessed: 2026-09-15.

Anthropic. Introducing Claude Sonnet 5, 2026b. URL https://www.anthropic.com/news/claude-sonnet-5. Accessed: 2026-09-15.

Anthropic. Claude code. https://docs.anthropic.com/en/docs/claude-code, 2026c. Accessed: 2026-09-16.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103):857–863, 2026.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6709–6738, 2025.

Thomas Berkane, Qianyi Wang, and Maimuna S Majumder. Self-ensembling vision-language models for chart data extraction. *arXiv preprint arXiv:2605.27298*, 2026.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

DeepSeek. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026a. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Accessed: 2026-09-15.

DeepSeek. DeepSeek-V4-Flash-Vision-Exp, 2026b. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-Vision-Exp. Accessed: 2026-09-15.

Junpeng Ding, Zichen Tang, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, et al. Decoding scientific experimental images: The spur benchmark for perception, understanding, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12859–12882, 2026.

Google DeepMind. Gemini 3.1 Pro model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-09-15.

Google DeepMind. Gemini 3.7 Flash model card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed: 2026-09-15.

Google DeepMind. Gemini 3.8 Flash model card, 2026c. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Accessed: 2026-09-15.

Muhammad Yusuf Hassan, Mayank Singh, et al. Lineex: Data extraction from scientific line charts. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6213–6221, 2023.

Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, and Daniel Kang. Reprobench: can agentic ai systems assess the reproducibility of social science research? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23616–23626, 2025.

Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. Chartsense: Interactive data extraction from chart images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 6706–6717, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025957. URL https://doi.org/10.1145/3025453.3025957.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. Matvix: Multimodal information extraction from visually rich articles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3636–3655, 2025.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023a.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770, 2023b.

Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, et al. The last human-written paper: Agent-native research artifacts. *arXiv preprint arXiv:2604.24658*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. The Computer Vision Foundation, January 2021. URL https://www.microsoft.com/en-us/research/publication/chartocr-data-extraction-from-charts-images-via-a-deep-hybrid-framework/.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525. IEEE, 2020.

Jiacheng Miao, Joe R Davis, Yaohui Zhang, Jonathan K Pritchard, and James Zou. Paper2agent: Reimagining research papers as interactive and reliable ai agents. *arXiv preprint arXiv:2509.06917*, 2025.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality—all in one model, 2026. URL https://www.minimax.io/blog/minimax-m3. Accessed: 2026-09-15.

OpenAI. GPT-5.6 system card, 2026. URL https://deploymentsafety.openai.com/gpt-5-6. Accessed: 2026-09-15.

Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, 2026. URL https://qwen.ai/blog?id=qwen3.8. Accessed: 2026-09-15.

QwenCloud. Qwen3.8-Flash, 2026. URL https://www.qwencloud.com/models/qwen3.8-flash. Accessed: 2026-09-15.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728, 2024.

Ankit Rohatgi. Webplotdigitizer, 2017.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285, 2013.

Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 393–402, 2011.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. Towards execution-grounded automated ai research. *arXiv preprint arXiv:2601.14525*, 2026.

Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. Corebench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Bowen Sun and Chaowei Xiao. Automated high-precision extraction and forensic verification of data-bearing vector figures. *arXiv preprint arXiv:2606.31345*, 2026.

Thinking Machines Lab. Inkling model card, 2026. URL https://thinkingmachines.ai/model-card/inkling/. Accessed: 2026-09-15.

Md Touhidul Islam, Yasir Mahmud, Sujan Kumar Saha, Mark Tehranipoor, and Farimah Farahmandi. Chartzero: Synthetic priors enable zero shot chart data extraction. *arXiv e-prints*, pp. arXiv–2605, 2026.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K Teal. Good enough practices in scientific computing. *PLoS computational biology*, 13(6):e1005510, 2017.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 2025.

Wenjin Yang, Jie He, and Xiaotong Zhang. Efficient extraction of experimental data from line charts using advanced machine learning techniques. *Graphical Models*, 139:101259, 2025.

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost, 2026. URL https://z.ai/blog/glm-5.3-flash. Accessed: 2026-09-15.

A Benchmark Details

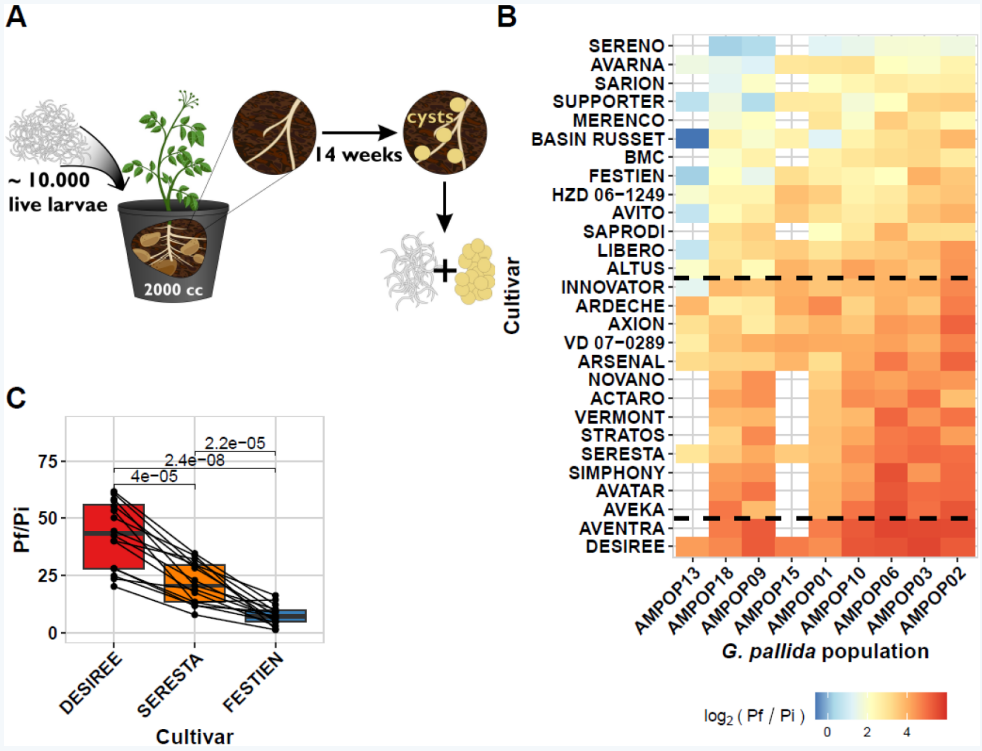
A.1 Benchmark Composition

Table 4: **Chart-type distribution of PlotGround-1k.** The benchmark contains 1,119 questions across 10 chart types. Bar, volcano, and scatter plots together account for 77.3% of the questions. The remaining types include line plots, distribution plots, and survival curves.

Chart type	Questions	Share (%)
Bar	423	37.8
Volcano	268	23.9
Scatter	174	15.5
Box	112	10.0
Line	86	7.7
Violin	27	2.4
Bubble	15	1.3
MA plot	6	0.5
Strip	5	0.4
Survival	3	0.3
Total	1,119	100.0

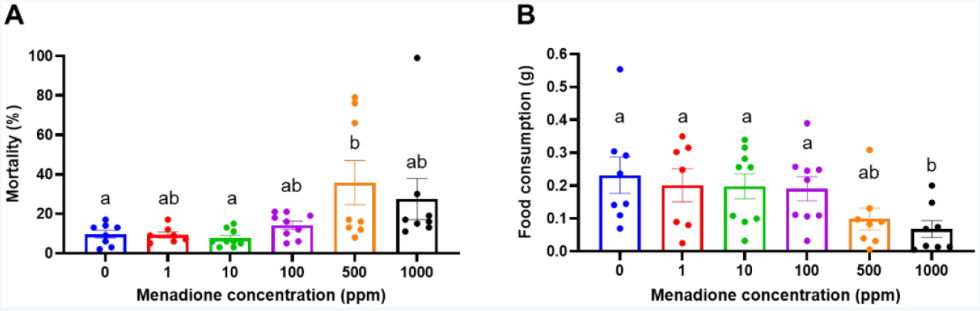
A.2 Benchmark Examples

Box Plot



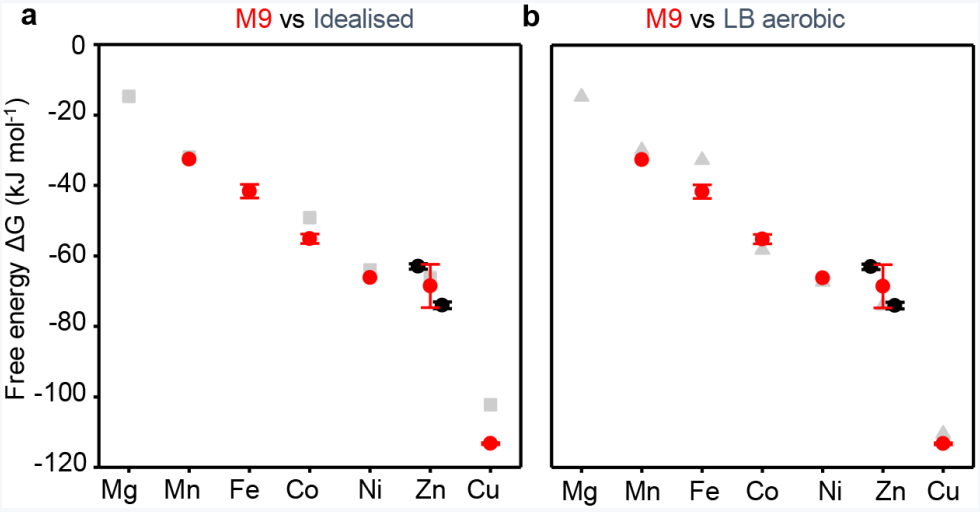
Question: In panel C, what is the median (center line) of the DESIREE box? (y-axis: 'Pf/Pi')
Answer: 43.147

Bar Plot



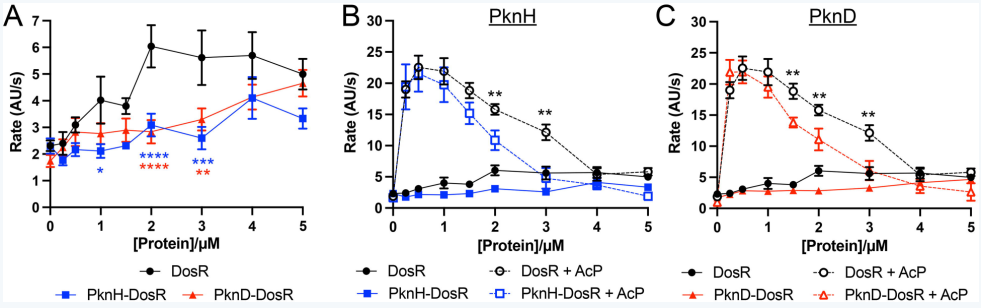
Question: In panel B, what is the bar height for the Menadione concentration of 500 ppm? (y-axis: 'Food consumption (g)')
Answer: 0.098525

Scatter Plot



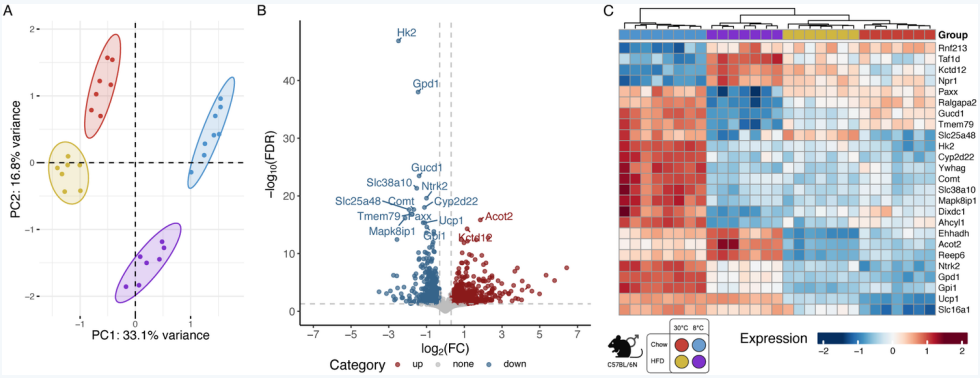
Question: In panel a, what is the free energy ΔG (kJ mol^{-1}) of the Idealised (gray square) data point at the Mg position?
Answer: -14.698

Line Plot



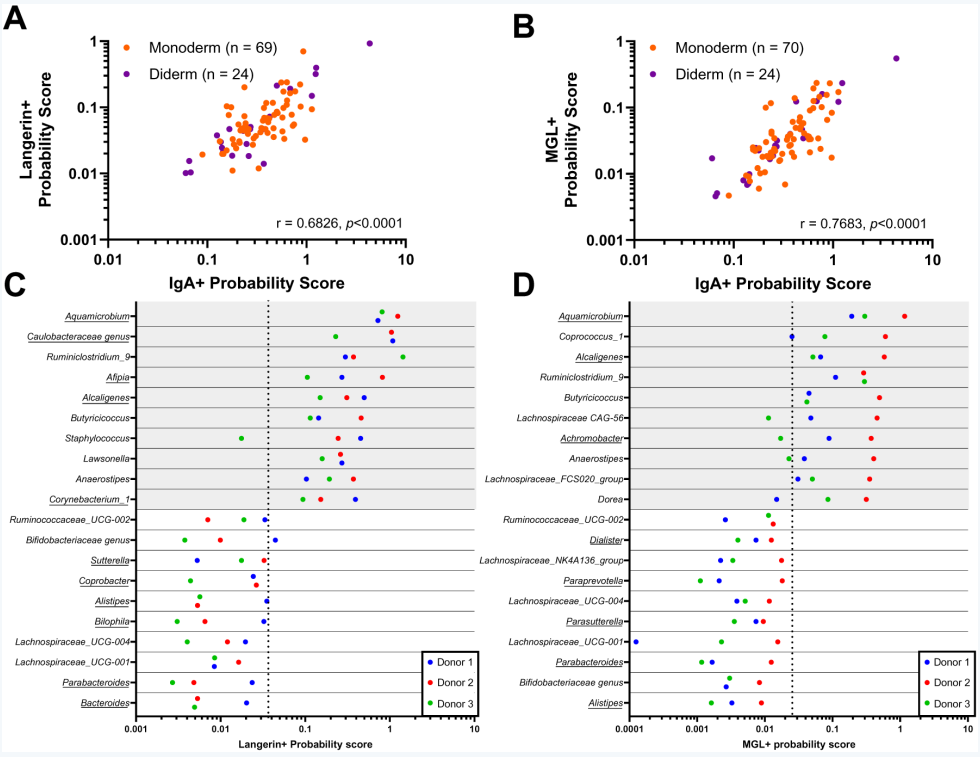
Question: In panel A, what is the Rate (AU/s) of the DosR series at [Protein]/μM = 2?
Answer: 6.04

Volcano Plot



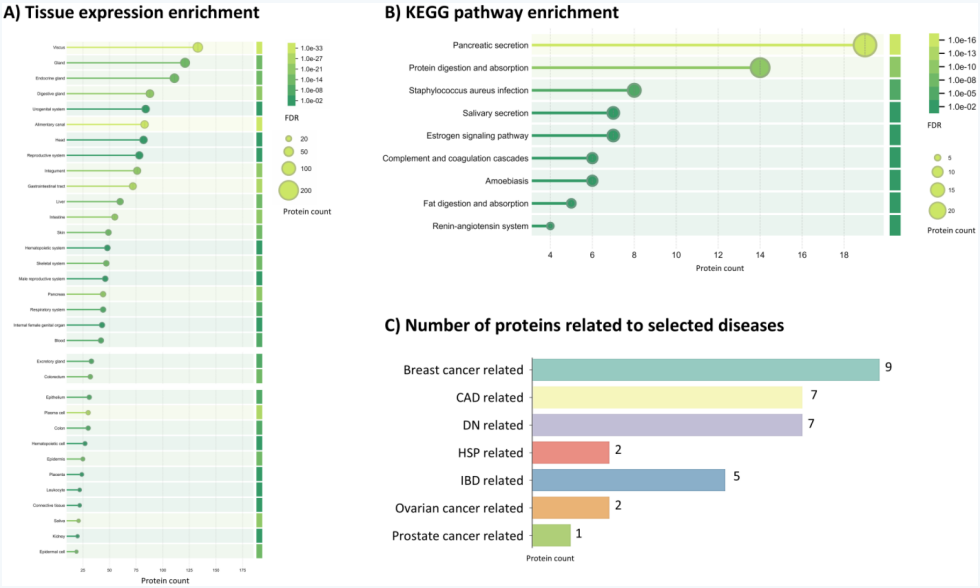
Question: In panel B, what is the -log10(FDR) value (y-axis) of the point labeled ”Hk2” in the volcano plot?
Answer: 46.86007339557034

Strip Plot



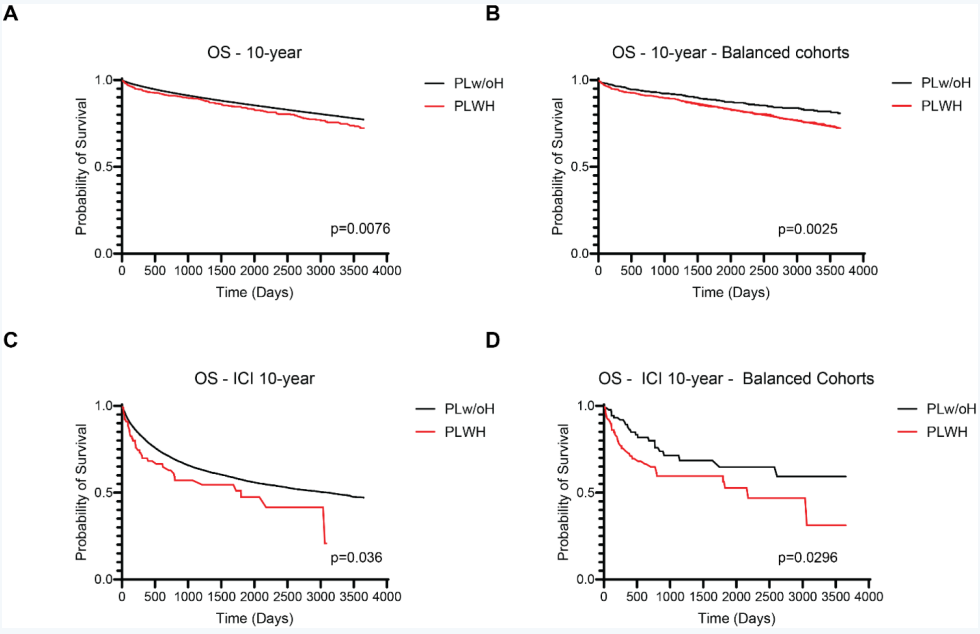
Question: In panel D, what is the MGL+ probability score for the Donor 2 (red) dot in the Aquamicrobium row?
Answer: 1.165311

Bubble Plot



Question: In panel A, what is the protein count for the bubble labeled 'Viscus'?
Answer: 132

Survival Curves



Question: In panel A, what is the probability of survival at the rightmost end of the PLWH curve?
Answer: 0.7229

A.3 Human Verification Protocol

Four reviewers manually reviewed all 1,205 candidate questions. Three reviewers each checked 300 questions, and the fourth checked the remaining 305. Reviewers used WebPlotDigitizer (Rohatgi, 2017) to recover the target value from the figure and then compared it with the generated answer. A candidate passed the initial check if the recovered value was within 2% relative error of the generated answer. Candidates were flagged for further review if the discrepancy exceeded 2%, the target could not be uniquely identified, or the figure could not be read reliably.

A fifth reviewer adjudicated all 131 flagged candidates using WebPlotDigitizer and by examining the reconstruction recipe and underlying source data. Candidates were retained unchanged when the original recipe and answer were confirmed to be correct. When an error in the recipe could be resolved, the recipe was corrected and the answer was recomputed from the source data and checked for consistency with the figure. Reasons for exclusion included unverifiable reconstruction recipes, figures in which the target value could not be read reliably, and disagreements between figures and their deposited source data. The 2% threshold therefore triggered further review rather than automatic exclusion. This process yielded 1,119 verified questions.

B Additional Results

B.1 Accuracy by Chart Type

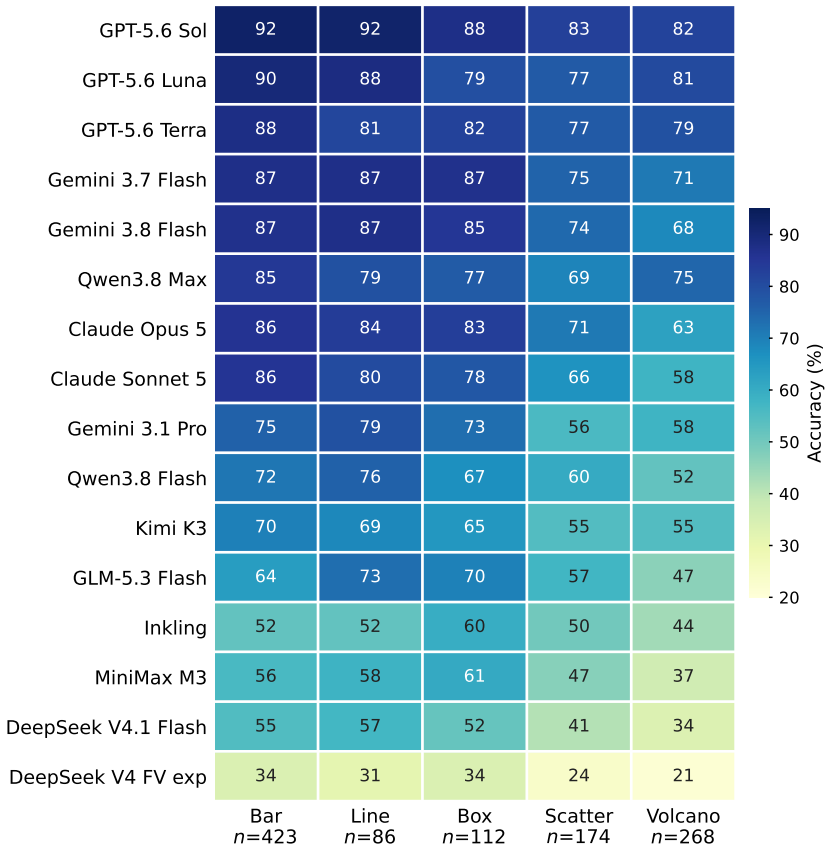
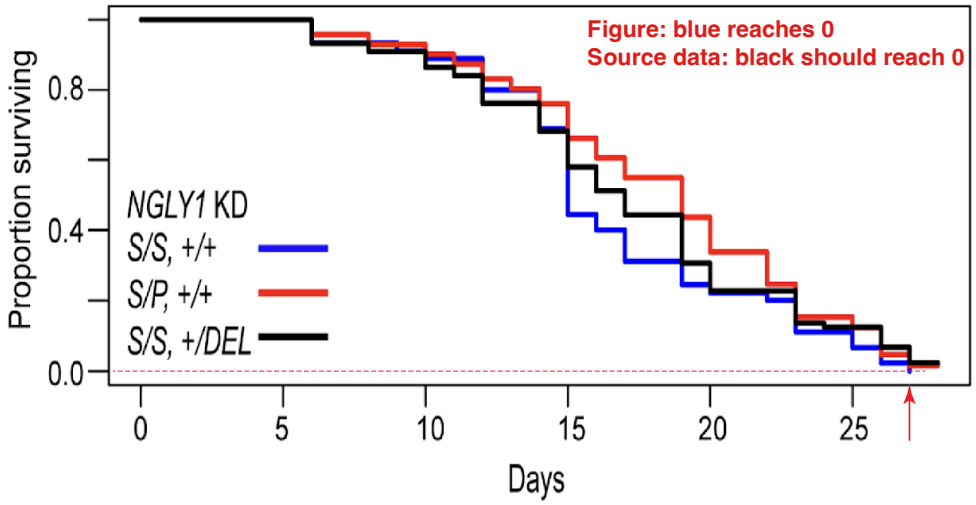


Figure 5: **Accuracy by chart type.** Per-model accuracy at a ±5% relative tolerance for the models in Table 2. Columns show the five most common chart types, ordered by mean accuracy across models. Among these five types, either scatter or volcano plots yield the lowest accuracy for each of the 16 models. This may reflect the difficulty of locating a labelled target point among many others.

B.2 Figure–Source Data Discrepancy Example

A Published survival curves



B Deposited source data

Deaths in the final days (deposited individual-level records)

Genotype	N	Deaths on day		Censored (DEAD = 0)	Survival at Day 28
		26	27		
S/S, +/+	89	5	4	2	**0.022**
S/P, +/+	71	5	2	3	**0.015**
S/S, +/DEL	45	2	1	**0**	**0.000**

Figure 6: **Additional example of a discrepancy between a published figure and its deposited source data.** The published survival curves differ from the values reconstructed from the deposited individual-level records. Although the discrepancy is visible throughout the trajectory, it is clearest at the endpoint: the published figure shows the blue curve ($S/S, +/+$) reaching zero survival first, whereas the deposited records imply zero survival for $S/S, +/DEL$. Survival at day 28 is the Kaplan–Meier estimate; two of the three censored $S/P, +/+$ animals were censored on day 21, so the estimate (0.015) differs from the simple proportion.

C PROMPTS USED IN THE STUDY

C.1 PLOT DIGITIZATION EVALUATION PROMPT

Each model receives a single user message containing the figure image followed by the question and a fixed answer-format suffix. No system prompt is used. Figures are converted from TIF to PNG and downscaled so that the longest edge is at most 1600 px. The suffix constrains the output format only: it gives no hint about how to read the figure (axes, ticks, interpolation) and explicitly permits step-by-step reasoning, so that models without internal reasoning are not forced to answer cold.

[figure image]

`{question}`

Work through the figure step by step, then end your reply with your final answer as a single JSON object on its own line:
`{"answer": <number, or null>}`

C.2 QUESTION REWRITING PROMPT

Question Rewriting Prompt

You are inside the directory of a bioRxiv preprint (full text XML and PDF, figure images as TIF, and supplementary data files under `content/supplements/`). The directory is read-only; write any scratch files to `/tmp`.

We have a benchmark question that asks for one numerical value plotted in a figure of this paper. The question is currently phrased in terms of the drawing ("panel", "bar", "point", "y-axis", ...). Your job is to REWRITE it so that it asks for the same value in the paper's own scientific terms, without referring to the figure or to any table.

The value

Original question: `{question}`
Figure file: `{figure_file}` (panel `{panel_id}`, chart type: `{chart_type}`)
Gold answer: `{gold}` `{unit}`

How the gold was computed from the deposited source data:
`{recipe}`

Rules for the rewritten question

1. Do not mention figures, panels, plots, bars, points, lines, axes, colors, legends, labels, or anything about how the value is drawn. Do not mention tables, sheets, columns, rows, or supplementary files either.
2. Describe the value the way the paper would in its text or methods: the experiment or analysis, the sample / group / condition / gene / term, the measured quantity, and the statistic (a single measurement, a mean, a median, a count, a fold change, ...). Use the terminology of the paper itself (main text, methods, figure caption). Read the paper to get this right.
3. Do NOT copy internal identifiers that appear only in the source data file (sheet names, cryptic column headers such as `study_count` or `padj`, sample codes) unless the paper text also uses them.
4. The rewritten question must identify exactly ONE value in the paper, and it must be the gold value above. If the same quantity was measured under several conditions, time points, replicates, or normalizations, spell out which one. Check the paper and the data for such collisions and rule them out in the wording rather than with a figure reference.
5. Keep the statistic identical to the original (if the original asks for a median, ask for the median; if it asks for a mean of replicates, ask for the mean). State the unit if the gold has one.
6. One sentence, or two if needed for disambiguation. Ask for a single number.

Before answering

Open the source data file and follow the recipe to confirm that it yields the gold value; report the recomputed value. Read the figure caption and the relevant text so the wording matches the paper.

Then check whether the paper ITSELF already states the answer: search the full text XML (abstract, results, methods, figure captions, any in-text tables) for the gold value, including rounded forms (e.g.

209.84 -> "210", "~2.1 x 10^2"), other units or scalings (pg/mL vs ng/mL, fraction vs percent) and the value embedded in a range or a "mean +/- sd". If the text gives the value closely enough that a reader could answer the question without the figure or the data (within a few percent), set value_in_text to true and quote the sentence. This does not change how you write the question; such questions are simply excluded downstream.

Return the result as a JSON object matching the required schema.

C.3 Evidence-Representation Agent Prompt

You are in the working directory of a bioRxiv preprint. It contains files from the paper: the full text as XML and, depending on what is available, figure images and/or supplementary data files. Use whatever is in the directory to answer the question below about this paper.

Question: {question}

Work through it step by step, then end your reply with your final answer as a single JSON object on its own line:
{"answer": <number, or null>}