CellChain: Benchmarking Cross-Scale Evidence Use for Perturbation Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in foundation models and large-scale cellular perturbation datasets have driven growing interest in using LLMs to predict and reason about cellular responses. Biological reasoning in this setting involves relating molecular mechanisms, cellular responses, and functional outcomes across scales. Evaluating foundation models therefore requires assessing not only whether their predictions are correct, but also how those predictions depend on the biological evidence provided. We introduce CellChain, an experimentally grounded benchmark comprising 590 biological cases and 2,360 matched evidence-condition instances across gene-dependency prediction, cytokine-axis identification, and drug-sensitivity ranking. CellChain links perturbation experiments and external biological resources through a provenance-aware evidence graph spanning molecular, pathway, transcriptional, and phenotypic information. Task-specific conditions vary evidence coverage, composition, or assignment while preserving questions, candidate answers, reference targets, and scoring rules. Across 18 foundation models, the highest accuracies under prespecified target-evidence conditions are 57.5%, 75.5%, and 70.8%, respectively, and broader evidence coverage does not consistently improve accuracy. On the drug-sensitivity benchmark cohort, matched reference histories outperform swapped histories by an average of 13.0 percentage points in cluster-weighted accuracy. Decision-transition analysis separates corrections, newly introduced errors, and wrong-answer switches, revealing substantial answer changes hidden by small net accuracy differences. CellChain provides controlled comparisons for attributing changes in model predictions to prespecified expansions of biological context, including changes in evidence coverage, composition, and biological assignment.

1 Introduction

In recent years, researchers have developed a variety of benchmarks to evaluate large language models' biological reasoning capabilities, with an emphasis on cellular prediction tasks (Zhang et al., 2026a; Zhao et al., 2026) and pathway-centered mechanistic reasoning (Zhao et al., 2025; Kim et al., 2026). While these benchmarks assess important aspects of biological understanding, they primarily evaluate whether models produce correct outcomes, providing limited insight into how those outcomes depend on the biological information supplied in context. This limitation is particularly important for cellular perturbation reasoning, where models must connect molecular interactions, pathway processes, transcriptional responses, and functional outcomes across biological scales (Bunne et al., 2024). Existing studies further show that incorporating experimental observations or structured pathway knowledge does not necessarily lead to reliable predictions, particularly when models must interpret causal directionality or relate mechanistic information to downstream responses (Wu et al., 2025; Kim et al., 2026; Yuan et al., 2026). Evaluating biological reasoning therefore requires going beyond endpoint correctness to determine how model decisions change as specific biological evidence is introduced, combined, or reassigned under controlled conditions. *Moreover, predictive improvements alone do not establish cross-scale integration, as models may exploit target-associated cues rather than integrate query-relevant experimental evidence.*

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.

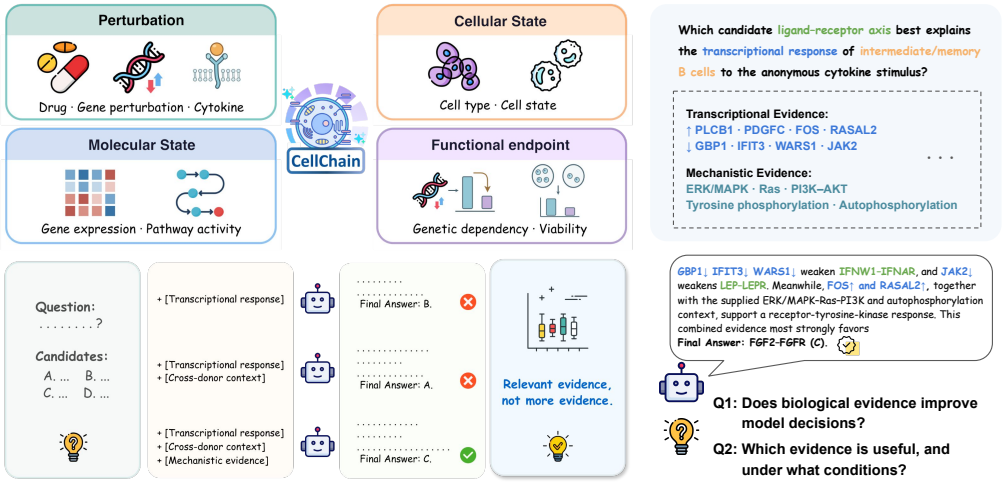


Figure 1: Overview of CellChain. CellChain integrates cross-scale biological information into structured evidence and prediction targets for evidence-grounded QA construction. The right panel illustrates a representative cytokine-axis evaluation with a held-out reference answer.

To address this gap, we introduce CellChain, **an evidence-grounded benchmark for characterizing foundation model predictions under controlled cross-scale biological evidence.** CellChain integrates large-scale perturbation experiments with external biological resources through an evidence graph connecting perturbations and cellular contexts to molecular annotations, pathway relations, transcriptional responses, and phenotypic outcomes. Reference answers are defined from experimental measurements or intervention metadata, and model-visible evidence retains its source provenance. Matched evaluation conditions preserve the question, candidate set, reference answer, and scoring rule while varying prespecified evidence-exposure configurations. This enables attribution of prediction changes to the corresponding context interventions while holding the underlying decision task fixed. This design spans genetic, cytokine, and chemical perturbations, supporting comparisons of predictive performance and evidence-dependent decisions without assuming that accuracy identifies the underlying reasoning mechanism. Table 1 situates CellChain relative to prior benchmarks.

We evaluate 18 frontier foundation models, spanning proprietary and open-source models. The highest accuracies are 57.5% for Gene Dependency Prediction, 75.5% for Cytokine-Axis Inverse Recovery, and 70.8% for Drug-Sensitivity Ranking, leaving substantial predictive headroom even under the target evidence conditions. Evidence-associated gains vary markedly across tasks and conditions, with broader context failing to yield consistent improvements. Moreover, substantial changes in model predictions can accompany nearly unchanged accuracy, revealing behavior obscured by aggregate performance alone. Model rankings also vary across tasks, with open-source models achieving competitive performance alongside proprietary systems. Together, these findings motivate evaluating both prediction accuracy and evidence-dependent decisions in biological reasoning.

2 CellChain

We present CellChain, a benchmark for evaluating foundation-model prediction behavior under controlled cross-scale biological evidence in cellular perturbation reasoning. CellChain has three key features: (1) **Cross-scale evidence integration**, connecting molecular annotations, pathway relations, transcriptional responses, and phenotypic measurements through a traceable evidence graph (§2.1); (2) **Complementary perturbation reasoning tasks**, assessing decisions across genetic, cytokine, and chemical perturbations (§2.2); and (3) **Controlled evidence exposure**, examining how information content and biological associations affect model decisions under paired evaluation conditions (§2.3). The following subsections describe data integration, task construction, evaluation protocols, quality control, and dataset characteristics.

Table 1: Comparison between CellChain and existing biological reasoning benchmarks.

Benchmark	Biological grounding			QA direction			Evidence diagnostics		
	Exp. perturb.	Multiple interventions	Joint cross-scale	Forward effect	Reverse cause	Candidate ranking	Linked provenance	Controlled evidence	Effect attribution
SOAR (Liu et al., 2025; *Briefings in Bioinformatics*)	✘	✘	✘	✘	✘	✘	✘	✘	✘
CellVerse (Zhang et al., 2026a; *NeurIPS*)	✔	✔	partial	✔	✘	✘	✘	✘	✘
SC-Arena (Zhao et al., 2026; *ICLR*)	✔	✘	partial	✔	✘	✘	partial	✘	✘
PerturbQA (Wu et al., 2025; *ICLR*)	✔	✘	partial	✔	✘	✘	partial	partial	partial
VCWorld / GeneTAK (Wei et al., 2026; *ICLR*)	✔	✘	✔	✔	✘	✘	partial	partial	partial
BioMaze (Zhao et al., 2025)	partial	partial	partial	✔	✘	✘	✔	partial	partial
PertReasonQA (Kim et al., 2026; *ICML Workshop*)	✔	✔	✔	✔	✘	✘	partial	✔	✔
AssayBench (De Brouwer et al., 2026)	✔	✘	partial	✔	✘	✔	✔	✘	✘
CellChain (Ours)	✔	✔	✔	✔	✔	✔	✔	✔	✔

Legend. ✔ = explicit feature; partial = partial/indirect coverage; ✘ = not central to the design.

2.1 Cross-scale evidence integration

We integrate three experimental anchor datasets spanning chemical, cytokine, and genetic perturbations: Tahoe-100M (Zhang et al., 2026b), PBMC Parse10M (Oesinghaus et al., 2025), and Replogle Perturb-seq (Replogle et al., 2022). These are linked to 14 external resources providing functional endpoints, molecular annotations, pathway knowledge, cellular context, and interaction evidence, including PRISM, DepMap, PubChem, ChEMBL, BindingDB, UniProtKB, Reactome, Gene Ontology/GOA, Cellosaurus, scBaseCount, CORUM, IntAct, SIGNOR, and Open Targets (**?**Tsherniak et al., 2017; Kim et al., 2023; Zdrazil et al., 2024; Gilson et al., 2016; bat, 2025; Milacic et al., 2024; Aleksander et al., 2023; Huntley et al., 2015; Bairoch, 2018; Youngblut et al., 2026; Tsitsiridis et al., 2023; Del Toro et al., 2022; Lo Surdo et al., 2023; Ochoa et al., 2023). Table 5 summarizes their data types and roles in CellChain.

All records are represented as provenance-tagged evidence units that preserve source-specific semantics. Target-defining endpoint fields are reserved for held-out query construction and scoring, whereas prespecified annotations, measurements, and eligible reference-case evidence may be exposed under task-specific conditions. Held-out query targets and all fields used to derive them are excluded from model inputs, and curated associations are not interpreted as direct causal or experimental effects unless explicitly supported by the source.

Entity and condition alignment. We align heterogeneous sources using shared biological entities and explicit experimental context. Entity mappings establish cross-source links, while metadata such as cell type or cell line, donor, dose, and duration determine whether linked evidence is applicable to a given case. This preserves source-specific measurement semantics and avoids treating biologically related records as experimentally equivalent.

Evidence graph organization. We organize cross-scale information into a provenance-aware evidence graph linking perturbations and cellular contexts to task-relevant evidence across biological levels. The evidence graph is a provenance-linked organization of record. For Gene Dependency, CRISPR perturbations are linked to UniProtKB annotations, Reactome pathways, and Perturb-seq transcriptional responses, while DepMap gene-effect measurements remain held out. For Cytokine-Axis Recovery, cytokine-treated cellular contexts are linked to measured transcriptional responses, process-level evidence, and system context, while the applied ligand–receptor axis remains held out. For Drug-Sensitivity Ranking, compound–cell-model contexts are linked to transcriptional responses and development-set reference histories, while the query drug-sensitivity endpoint remains held out.

2.2 Benchmark Task Construction

CellChain defines three complementary perturbation prediction tasks. **Gene Dependency Prediction** predicts a category derived from held-out DepMap gene-effect scores, which quantify the effect of gene disruption on cell fitness. **Cytokine-Axis Inverse Recovery** identifies the applied cytokine through a ligand–receptor axis mapped from the recorded stimulus using frozen PBMC annotations. **Drug-Sensitivity Ranking** selects the more sensitive of two cell models exposed to the same compound, using held-out PRISM drug-response AUC as the reference endpoint.

Selecting Biological Cases. We construct eligible cases by linking perturbation measurements, cellular contexts, and task-specific reference targets. Gene cases require matched response and dependency records; Cytokine cases require treated cells with donor–cell-type-matched controls; Drug cases require complete transcriptional trajectories, QC-qualified sensitivity endpoints, and sufficient development-set reference-history coverage. These criteria ensure that the required model-visible evidence is available while reference targets remain independently defined. Additional case-selection procedures and evidence diagnostics are described in Appendix D.

Defining Prediction Targets. Gene Dependency uses three categories to distinguish strong, intermediate, and weak or absent dependency. Writing g for the gene-effect score, we retain cases with $g \leq -1.05$, $-0.75 \leq g \leq -0.55$, or $g \geq -0.25$, and exclude scores in the intervening intervals as boundary-ambiguous. The Cytokine reference is the recorded stimulus mapped to its ligand–receptor axis using the frozen PBMC annotation resource, rather than an axis inferred from the measured response. For Drug Sensitivity, lower PRISM AUC indicates greater sensitivity, and pairs require $|\Delta\mathrm{AUC}| \geq 0.15$ to exclude near-tie comparisons.

Curating QA Tasks. Each case contains a biological question, candidate answers, and a reference label. Gene presents the three dependency categories; Cytokine presents the target axis and three distractors, yielding a common four-choice inverse-identification task; Drug presents two cell models in both candidate orders, so that order-invariance can be assessed independently of task difficulty. Within a case, the question, candidate set, reference target, and scoring rule are held fixed across evidence conditions. This controls the task structure so that paired comparisons reflect the specified change in visible evidence or its assignment, rather than a different question or answer space.

2.3 Controlled Evidence Exposure

We evaluate evidence dependence through controlled-evidence contrasts between matched biological contexts. Each base case retains the same question, candidate answers, held-out query target, and scoring rule, while the prespecified model-visible evidence or its biological assignment varies.

Controlling Information Coverage. We construct matched conditions that vary the biological evidence available for reasoning. Cumulative conditions progressively add evidence types and define incremental contrasts under a fixed order. Cytokine conditions present process annotations and system context separately and jointly under a shared evidence budget, while retaining the measured response in every condition. Baseline conditions retain the information required to specify each decision problem. The process annotations are linked to the recorded cytokine stimulus, and the joint condition reallocates the evidence budget between sources.

Testing Biological Associations. We further vary whether biological evidence is assigned to the entities it describes. Specifically, in Drug-Sensitivity Ranking, matched histories associate each candidate with development-set phenotype evidence from the same cell model, whereas swapped histories exchange these histories between candidates. To avoid leakage, reference histories are fixed before model evaluation from 36 disjoint development compounds after PRISM quality control (median fit $R^2 \geq 0.5$, replicate AUC range ≤ 0.1, positive slope, and ordered asymptotes); within-compound AUC comparisons determine sensitivity percentiles without using query-compound AUC or query transcriptional responses to select reference records. Both conditions retain the query responses and the same history-comparison instruction. Accordingly, their contrast measures sensitivity to the candidate–history assignment conditional on the supplied evidence in the benchmark cohort.

Separating Evidence from Targets. To preserve a clear separation between prediction evidence and scoring targets, held-out query endpoint values and explicit answer keys are excluded from model-visible inputs. Other measurements and reference-case evidence are exposed only when prespecified for the corresponding condition and remain subject to the same task-specific masking rules. In addition, task-specific identity masking is applied consistently across matched conditions, while Cytokine target-linked process annotations remain associated with the recorded stimulus. Finally, model-visible inputs are screened during quality control for prohibited endpoint fields, answer labels, direct target-revealing information, and condition-specific leakage cues.

Table 2: Construction statistics of CellChain.

(a) Cases and coverage

Statistic	Gene	Cytokine	Drug
Source records	3,443	13,967	11,304
Eligible cases	2,092	267	737
Benchmark cases	240	110	240
Benchmark / eligible (%)	11.5	41.2	32.6
Perturbations	219	11	26
Cellular contexts	2	6	34
Donors	–	12	–

(b) Evidence and evaluation

Statistic	Gene	Cytokine	Drug
Evidence units	2,495	330	672
Evidence assignments	3,032	2,290	6,938
Biological clusters	219	11	26
Evidence conditions	4	4	4
Task instances	960	440	960
Candidate orderings	1	1	2
Model requests	960	440	1,920

2.4 Dataset Analysis

Table 2 summarizes the construction and coverage of CellChain. The benchmark comprises 590 problems derived from 17 biological resources spanning genetic, cytokine, and chemical perturbations, with evidence across molecular, pathway-process, transcriptional, and phenotypic levels. Experimental contexts include cell type, cell line, donor, dose, and duration where applicable. We apply four quality-control checks to each benchmark case. First, cross-source links must agree on the biological entity and cellular context, and unresolved mappings are excluded. Second, reference answers are recomputed from the underlying experimental endpoint or recorded intervention identity using the prespecified task rule; only cases with a unique valid answer are retained. Third, the question, candidate set, reference answer, and scoring rule remain fixed across matched evidence conditions, with only the specified evidence or associated instructions changing. Finally, model-visible inputs are screened for answer labels, prohibited query endpoint fields, and task-specific identity strings. Cases failing an applicable check in any condition are excluded as matched units. Six PhD researchers manually reviewed 150 sampled cases, of which 142 passed (94.7%).

3 Experiments

3.1 Experiment Setup

Evaluated Models. We evaluate the performance of state-of-the-art foundation models on CellChain, spanning both proprietary and open-source systems with reasoning capabilities. The proprietary models evaluated are GPT-5.6 Sol (OpenAI, 2026) and Gemini 3.8 Flash (Google DeepMind, 2026). Open-source models include DeepSeek-V4-Pro (DeepSeek, 2026a), DeepSeek-V4.1-Flash (DeepSeek, 2026b), GLM-5.2 (Z.ai, 2026a), GLM-5.3 (Z.ai, 2026b), MiniMax-M3 (MiniMax, 2026), Kimi-K3 (Moonshot AI, 2026), Qwen3.8-Flash-Next (Qwen Team, 2026c), Qwen3.8-27B (Qwen Team, 2026b), Qwen3.6-35B-A3B (Qwen Team, 2026a), GPT-OSS-20B (OpenAI, 2025), Agents-A1 (Shanghai AI Laboratory, 2026), Intern-S1-mini (Shanghai AI Laboratory, 2025), Hunyuan-1.8B-Instruct (Tencent, 2025), Hy3-FP8 (Tencent, 2026), Llama-3.3-70B-Instruct (Meta AI, 2024), and Llama-4-Scout-17B-16E-Instruct (Meta AI, 2025). This selection covers a range of model scales, providers, and reasoning configurations, accessed via standardized serving backends (API or vLLM). Detailed model specifications and versions are provided in Table 4.

Prompt Design. We use a structured prompt consisting of task instructions, fixed query information, model-visible evidence, candidate answers, and a structured output schema, as shown in Figure 2. Questions, candidates, reference answers, and scoring remain fixed across matched conditions, while prespecified evidence content, its biological assignment, or associated instructions vary. Explicit answer keys and query-specific endpoint measurements are excluded from model-visible prompts.

Implementation Details. Model predictions are evaluated against held-out query targets using task-specific metrics. Results are aggregated by model, task, and context arm to enable paired comparisons across conditions. We set the sampling temperature to 0.0 to reduce variability. The maximum output length is set to 8,192 tokens, with an increased budget of 32,768 tokens for reasoning-enabled evaluations. Each prompt is constructed with up to 15 evidence units, using a prespecified evidence selection and ordering procedure applied consistently across models.

Controlled Evidence Evaluation Template

SYSTEM

You are a molecular and cellular biology expert. Answer the biological question using only the fixed task inputs and evidence visible under the current exposure condition. Treat annotations and biological associations as supporting evidence rather than direct measurements of the held-out query endpoint unless explicitly stated. Do not infer endpoint values that are not provided. Return exactly one valid JSON object conforming to the required output schema.

MATCHED DECISION TASK — FIXED ACROSS CONDITIONS

Question: `{biological_question}`

Query and Fixed Observations: `{query_context_and_fixed_observations}`

Candidate Answers: `A: {candidate_A}; B: {candidate_B}; ...`

CONTROLLED EVIDENCE EXPOSURE — VARIED BY CONDITION

Visible Cross-Scale Evidence:

```
[Molecular] {molecular_evidence}
[Pathway/Process] {pathway_evidence}
[Transcriptional] {transcriptional_evidence}
[Reference-History] {reference_history_evidence}
```

Across matched evidence conditions, the visible evidence content or its biological assignment varies while the underlying decision task remains fixed.

OUTPUT CONTRACT — FIXED ACROSS CONDITIONS

`{task_specific_output_schema}`

BENCHMARK CONTROL — NOT MODEL-VISIBLE

Held-out Query Target and Target-Defining Endpoint:
`{held_out_query_target}`

Figure 2: Controlled-evidence evaluation template in CellChain, with a fixed decision task and condition-specific evidence exposure.

3.2 Evaluation Metrics

We report strict accuracy together with macro-averaged precision, recall, and F1, all normalized to $[0, 1]$. Let i index biological cases and o index candidate-order presentations: $O{=}1$ for Gene Dependency and Cytokine Axis, and $O{=}2$ (reversed candidate order) for Drug Sensitivity. Let $\hat{y}^{(a)}_{m,i,o}$ be model m's prediction under evidence condition a, and $y_{i,o}$ the reference answer for that presentation. We score each presentation and average within case:

$$s^{(a)}_{m,i,o} = \mathbb{I}\left[\hat{y}^{(a)}_{m,i,o} = y_{i,o}\right], \qquad \bar{s}^{(a)}_{m,i} = \frac{1}{O}\sum_{o=1}^{O} s^{(a)}_{m,i,o}.$$

For Drug Sensitivity, $\bar{s}^{(a)}_{m,i} \in \{0, \frac{1}{2}, 1\}$ because the two candidate-order presentations encode the same biological target with reversed answer positions.

Endpoint Performance. Endpoint accuracy averages $\bar{s}^{(a)}_{m,i}$ equally across the N cases:

$$\mathrm{Acc}_m(a) = \frac{1}{N}\sum_{i=1}^{N} \bar{s}^{(a)}_{m,i}.$$

For Drug, accuracy first averages the two candidate-order presentations within each case and then averages across cases. Macro precision, recall, and F1 use pooled presentation-level counts, retaining invalid outputs as false negatives. Cytokine and Drug macro metrics describe candidate-position labels. Endpoint accuracy is case-weighted, whereas controlled contrasts are cluster-weighted.

Controlled-Evidence Contrasts. Cases are grouped into biological clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ by perturbed gene, cytokine axis, or compound, and contrasts weight clusters equally. For conditions a, b, the within-cluster contrast is

$$\delta_{m,k}(a,b) = \frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\left(\bar{s}^{(a)}_{m,i} - \bar{s}^{(b)}_{m,i}\right),$$

and the model- and benchmark-level contrasts give equal weight to clusters and models:

$$\Delta_{m,\mathrm{Acc}}(a,b) = \frac{1}{K}\sum_{k=1}^{K}\delta_{m,k}(a,b), \qquad \overline{\Delta}_{\mathrm{Acc}}(a,b) = \frac{1}{M}\sum_{m=1}^{M}\Delta_{m,\mathrm{Acc}}(a,b).$$

Positive values indicate higher accuracy under a. For Drug, $\Delta_{\mathrm{Combined}}$ denotes Matched minus Background, since this comparison jointly adds responses, histories, and history-comparison instructions. For Cytokine, we define

$$\Gamma = A_{\mathrm{Joint}} - A_{\mathrm{Process}} - A_{\mathrm{System}} + A_{\mathrm{Response}},$$

where each A is cluster-weighted. Because the Joint condition reallocates a fixed evidence budget across Process and System evidence, Γ is interpreted as a joint-condition deviation from additivity rather than a factorial interaction effect. Negative Γ indicates that Joint performance falls below the additive expectation defined by the separate conditions, potentially reflecting both evidence combination and budget-induced composition changes.

Decision Transitions. For a transition from b to a, paired within case and presentation, let p_b, p_a be normalized outputs (invalid outputs share a common state). Under the same cluster and model weights as above:

$$F = \Pr_w(p_a \neq p_b), \qquad C = \Pr_w(p_b \neq y,\, p_a = y),$$
$$D = \Pr_w(p_b = y,\, p_a \neq y), \quad W = \Pr_w(p_a \neq p_b,\, p_b \neq y,\, p_a \neq y).$$

It follows that $F = C + D + W$ and $\overline{\Delta}_{\mathrm{Acc}} = C - D$, where W captures both switches between incorrect answers and transitions involving invalid outputs.

3.3 Main Results

Cross-scale biological reasoning remains challenging for current foundation models. Across evaluated foundation models, the highest accuracies under the prespecified target-evidence conditions are 57.5% for Gene Dependency, 75.5% for Cytokine Axis, and 70.8% for Drug Sensitivity (Table 3a). The corresponding error rates under the task-specific scoring rules are 42.5%, 24.5%, and 29.2%. The three tasks have different candidate counts and should not be ranked by raw accuracy alone; their uniform-choice levels are 33.3%, 25.0%, and 50.0%. Within each task, accuracy spans 32.5–57.5%, 25.5–75.5%, and 33.3–70.8%, showing that the observed headroom is accompanied by substantial differences in model capability rather than a single shared performance ceiling. In particular, several models remain near uniform-choice on Gene Dependency despite receiving the full cumulative evidence, whereas stronger results on Cytokine Axis and Drug Sensitivity show the corresponding decision problems contain usable predictive information. Together, these results establish a challenging evaluation across distinct perturbation modalities while motivating the condition-level analysis of how models respond to supplied evidence.

Table 3: Main results in CellChain(%).

(a) Endpoint performance under target evidence conditions

Model	Gene Dependency				Cytokine Axis				Drug Sensitivity			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Open-Source Models												
DeepSeek-V4-Pro	51.5	42.1	32.2	42.1	68.8	67.2	67.0	66.4	69.8	69.8	69.8	69.8
GLM-5.2	50.2	46.7	37.9	46.7	54.8	55.2	53.6	53.6	63.2	62.9	62.7	62.9
MiniMax-M3	38.8	44.2	37.6	44.2	65.4	65.8	65.3	65.5	67.7	67.7	67.7	67.7
Kimi-K3	52.3	54.2	48.0	54.2	70.7	70.4	70.3	70.0	70.0	70.0	70.0	70.0
DeepSeek-V4.1-Flash	47.5	52.9	43.4	52.9	61.1	61.6	61.1	60.9	43.7	43.8	43.7	43.8
GLM-5.3	53.9	55.0	47.4	55.0	71.4	71.6	71.1	70.9	68.7	68.5	68.6	68.5
Qwen3.6-35B-A3B	44.6	34.6	19.2	34.6	60.2	59.9	59.9	60.0	67.9	67.9	67.9	67.9
Qwen3.8-27B	44.5	34.2	18.4	34.2	59.1	59.7	59.2	59.1	70.6	70.6	70.6	70.6
GPT-OSS-20B	11.1	33.3	16.7	33.3	38.3	38.9	38.0	38.2	70.8	70.8	70.8	70.8
Agents-A1	39.0	35.0	20.6	35.0	45.8	45.8	45.3	45.5	46.0	46.0	45.8	46.0
Intern-S1-mini	16.7	32.9	17.3	32.9	42.0	40.6	40.7	40.0	44.1	40.4	42.1	40.4
Hunyuan-1.8B-Instruct	11.3	32.5	16.7	32.5	31.1	26.7	12.9	25.5	25.0	50.0	33.3	50.0
Qwen3.8-Flash-Next	34.3	47.9	39.1	47.9	79.7	75.2	76.8	74.5	69.7	69.6	69.7	69.6
Llama-3.3-70B-Instruct	11.1	33.3	16.7	33.3	39.2	33.7	29.9	32.7	33.2	33.3	33.2	33.3
Llama-4-Scout-17B-16E-Instruct	11.1	33.3	16.7	33.3	29.2	35.7	30.0	36.4	35.4	35.4	35.4	35.4
Hy3-FP8	11.1	33.3	16.7	33.3	60.8	61.2	60.4	60.9	60.6	55.4	49.2	55.4
Proprietary Models												
GPT-5.6-Sol	52.9	57.5	52.2	57.5	72.1	72.4	71.9	71.8	68.8	68.8	68.7	68.8
Gemini-3.8-Flash	68.9	53.3	43.0	53.3	75.7	76.3	75.7	75.5	70.2	70.2	70.2	70.2

(b) Controlled evidence contrasts

Model	Gene Dependency			Cytokine Axis			Drug Sensitivity		
	ΔMol.	ΔPath.	ΔTrans.	ΔProcess	ΔSystem	Γ	ΔResponse	ΔCombined	ΔAssign.
Open-Source Models									
DeepSeek-V4-Pro	+0.5	+8.2	−1.1	+41.8	+0.9	+0.0	+2.1	+12.6	+27.0
GLM-5.2	+0.0	+11.9	−0.5	+44.5	+0.0	−6.4	−6.9	+8.9	+19.7
MiniMax-M3	+0.5	+10.3	−1.8	+43.6	−6.4	+10.0	−6.2	+11.1	+25.1
Kimi-K3	+17.1	−3.4	+1.1	+60.9	+7.3	−15.5	+0.2	+22.1	+27.0
DeepSeek-V4.1-Flash	+5.7	+10.3	−0.7	+39.1	+3.6	−11.8	−9.6	−7.3	−5.3
GLM-5.3	+13.5	+3.4	+0.5	+54.5	+4.5	−8.2	−5.6	+10.7	+25.9
Qwen3.6-35B-A3B	+0.0	+0.5	+0.5	+46.4	−1.8	+0.0	+3.8	+12.1	+23.9
Qwen3.8-27B	+0.0	+0.0	+0.7	+53.6	−0.9	−4.5	−1.0	+16.1	+28.2
GPT-OSS-20B	+0.0	+0.0	+0.0	+21.8	−4.5	+1.8	+1.8	+14.7	+28.3
Agents-A1	+0.0	+1.4	+0.2	+40.9	−2.7	−10.0	−2.8	−2.4	−0.9
Intern-S1-mini	+0.0	+0.5	−0.9	+40.9	+0.9	−9.1	−4.0	−6.8	−11.8
Hunyuan-1.8B-Instruct	+0.0	+0.5	−1.1	+0.0	+0.0	+0.9	+0.0	+0.0	+0.0
Qwen3.8-Flash-Next	+0.0	+15.5	−5.0	+56.4	−4.5	−2.7	−2.4	+12.2	+26.1
Llama-3.3-70B-Instruct	+0.0	+0.0	+0.0	+27.3	−2.7	−3.6	+0.1	−10.8	−20.6
Llama-4-Scout-17B-16E-Instruct	+0.0	+0.0	+0.0	+31.8	−0.9	−10.9	+0.3	−7.6	−16.5
Hy3-FP8	+0.0	+0.0	+0.0	+57.3	+0.9	−8.2	−7.9	+0.7	+5.0
Proprietary Models									
GPT-5.6-Sol	+4.8	+13.5	+0.0	+50.9	−0.9	−6.4	−8.9	+10.6	+25.7
Gemini-3.8-Flash	+18.0	−4.3	+0.5	+50.9	+1.8	−7.3	−7.2	+13.7	+26.4

Notes. Panel (a) evaluates full cumulative evidence for Gene Dependency, joint evidence for Cytokine Axis, and the matched condition for Drug Sensitivity. Panel (b) reports: Gene, $\Delta_{\text{Mol.}} = \text{Molecular} - \text{Base}$, $\Delta_{\text{Path.}} = \text{Pathway} - \text{Molecular}$, and $\Delta_{\text{Trans.}} = \text{Transcriptional} - \text{Pathway}$; Cytokine, $\Delta_{\text{Process}} = \text{Process} - \text{Response}$, $\Delta_{\text{System}} = \text{System} - \text{Response}$, and $\Gamma = \text{Joint} - \text{Process} - \text{System} + \text{Response}$, interpreted as a joint-condition deviation from additivity under the shared evidence budget; Drug, $\Delta_{\text{Response}} = \text{QueryResponse} - \text{Background}$, $\Delta_{\text{Combined}} = \text{Matched} - \text{Background}$, and $\Delta_{\text{Assign.}} = \text{Matched} - \text{Swapped}$.

Evidence attribution reveals heterogeneous effects across scales and assignments. Paired, cluster-weighted contrasts show that broader context does not uniformly improve predictions (Table 3b; Appendix E). Across 18 models, molecular and pathway evidence improve Gene Dependency accuracy by 3.3 and 3.8 points, while the transcriptional increment is -0.4 points (95% CI $[-1.1, +0.3]$). For Cytokine Axis, process-associated evidence yields a $+42.4$-point gain, while system context contributes -0.3 points and the joint-condition deviation is -5.1 points (95% CI $[-8.9, -1.6]$). Because the process annotations are linked to the target stimulus, this gain reflects the predictive value of target-linked process context, with response–mechanism integration unresolved. For Drug Sensitivity, matched histories outperform swapped histories by 13.0 points (95% CI $[+0.5, +24.5]$), whereas Matched versus Background is smaller and uncertain ($+6.2$ points; 95% CI $[-0.3, +12.2]$). Overall, context expansion effects depend on evidence type, composition, and biological assignment, motivating controlled attribution beyond endpoint accuracy.

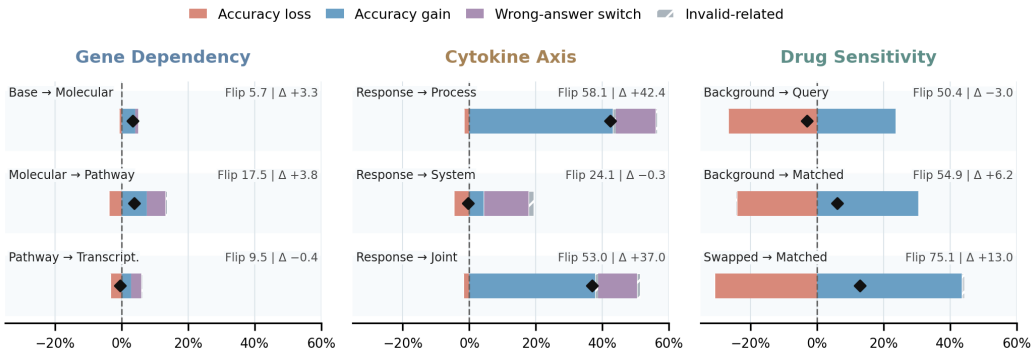


Figure 3: Decision transitions across evidence conditions. Bars show losses (leftward), gains, wrong-answer switches, and invalid-related changes; diamonds mark net Δ (pp), with flip rates annotated.

The performance gap between proprietary and open-source foundation models varies by biological task. Proprietary models achieve the highest observed accuracy on Gene Dependency and Cytokine Axis, whereas open-source models lead on Drug Sensitivity. On Gene Dependency, GPT-5.6-Sol reaches 57.5%, compared with 55.0% for the strongest open-source model, GLM-5.3. On Cytokine Axis, Gemini-3.8-Flash obtains 75.5%, followed closely by Qwen3.8-Flash-Next at 74.5%. The ordering reverses on Drug Sensitivity, where GPT-OSS-20B and Qwen3.8-27B achieve 70.8% and 70.6%, respectively, compared with 70.2% for Gemini-3.8-Flash. These differences are descriptive rankings rather than evidence of statistically significant differences between model groups. More broadly, the variation in task leaders indicates that performance differences between proprietary and open-source systems are task dependent rather than consistent across CellChain. These reversals highlight the value of CellChain' complementary task design in revealing task-specific capability differences obscured by aggregate model rankings.

Small net accuracy changes conceal substantial decision turnover. Figure 3 decomposes prediction changes into corrections, newly introduced errors, wrong-answer switches, and invalid-related transitions. Gene Pathway-to-Transcriptional changes 9.5% of predictions for a -0.4-point net difference, Cytokine Response-to-System changes 24.1% for -0.3 points, and Drug Swapped-to-Matched changes 75.1% for a $+13.0$-point gain. In each case, substantially more individual decisions change than the corresponding net accuracy difference suggests, because improvements can be offset by newly introduced errors or switches among incorrect answers. The positive panel-level Drug assignment effect also masks substantial model heterogeneity, revealing model-specific sensitivity to biological evidence assignment beyond aggregate accuracy. Additional diagnostics reveal sensitivity beyond aggregate accuracy: under Gene Base, all foundation models produce a single predicted category, while under Drug Matched, reversing candidate order changes 20.1% of predictions despite preserving the biological target. Invalid outputs remain rare overall (187/59,760, 0.31%). Together, these results show that endpoint accuracy alone can obscure how model decisions respond to controlled evidence interventions and task-preserving presentation changes.

4 Conclusion

We present CellChain, a benchmark for evaluating foundation-model prediction behavior under controlled cross-scale biological evidence. Across 18 models, the highest accuracies are 57.5% for Gene Dependency, 75.5% for Cytokine Axis, and 70.8% for Drug Sensitivity, leaving substantial predictive headroom. Evidence effects are selective: target-linked process annotations yield the largest Cytokine gains, matched reference histories outperform swapped histories on average in Drug Sensitivity, while additional transcriptional or system context provides little net benefit. Paired analyses further show that small accuracy differences can mask substantial prediction turnover. Overall, CellChain enables fine-grained, task-specific attribution of prediction changes to controlled variations in evidence coverage, composition, and biological assignment without conflating these effects with underlying reasoning mechanisms.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with language editing, code development, and figure prototyping. They were not used as authoritative sources or to determine reference answers. The authors verified all AI-assisted text, citations, code, figures, and analyses and take full responsibility for the final work.

ETHICS STATEMENT

We use existing cellular perturbation datasets and biological resources and conduct no new human or animal experiments. Human-derived data are used only in their authorized, de-identified forms and in accordance with source licenses. CellChain is intended for research evaluation and should not be used for clinical decision-making.

REPRODUCIBILITY STATEMENT

The paper and appendix document the data resources, benchmark construction, prompts, model configurations, and evaluation procedures. Subject to source licenses and anonymous-review requirements, we will release the derived benchmark artifacts, construction and evaluation code, configurations, and detailed results.

REFERENCES

Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

Amos Bairoch. The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT*, 29(2):25, 2018.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Edward De Brouwer, Carl Edwards, Alexander Wu, Jenna Collier, Graham Heimberg, Xiner Li, Meena Subramaniam, Ehsan Hajiramezanali, David Richmond, Jan-Christian Hütter, et al. Assaybench: An assay-level virtual cell benchmark for llms and agents. *arXiv preprint arXiv:2605.10876*, 2026.

DeepSeek. Deepseek v4 model card. `https://fe-static.deepseek.com/chat/transparency/deepseek-V4-model-card-EN.pdf`, 2026a.

DeepSeek. DeepSeek-V4.1-Flash: Smarter, faster, more efficient. `https://www.deepseek.com/en/news/deepseek-v4-1-flash/`, 2026b. Official model announcement.

Noemi Del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, Livia Perfetto, Karyn How, Prashansa Ratan, Gautam Shirodkar, et al. The intact database: efficient access to fine-grained molecular interaction data. *Nucleic acids research*, 50(D1):D648–D653, 2022.

Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

Google DeepMind. Gemini 3.8 flash model card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026. Published September 2, 2026.

Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):D1057–D1063, 2015.

Dongkwan Kim, Yiming Gao, Yining Yang, and Yang Shen. Pertreason: A knowledge-grounded benchmark and framework for cell-state-conditioned mechanistic reasoning of perturbation effects. *arXiv preprint arXiv:2607.18777*, 2026.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380, 2023.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. pp. 2023–09, 2024.

Junhao Liu, Siwei Xu, Yongxian Wu, and Jing Zhang. Single-cell omics arena: evaluation of large language models for automatic cell-type annotations on single-cell omics data via rna-seq bridging. *Briefings in Bioinformatics*, 26(6):bbaf622, 2025.

Prisca Lo Surdo, Marta Iannuccelli, Silvia Contino, Luisa Castagnoli, Luana Licata, Gianni Cesareni, and Livia Perfetto. Signor 3.0, the signaling network open resource 3.0: 2022 update. *Nucleic acids research*, 51(D1):D631–D637, 2023.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

Meta AI. Llama-3.3-70B-Instruct model card. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`, 2024. Official model card.

Meta AI. Llama-4-Scout-17B-16E-Instruct model card. `https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct`, 2025. Official model card.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678, 2024.

MiniMax. Minimax-m3 model card. `https://huggingface.co/MiniMaxAI/MiniMax-M3`, 2026. Official model card, accessed September 2026.

Moonshot AI. Kimi-k3 model card. `https://huggingface.co/moonshotai/Kimi-K3`, 2026. Official model card, accessed September 2026.

David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, et al. The next-generation open targets platform: reimagined, redesigned, rebuilt. *Nucleic acids research*, 51(D1):D1353–D1359, 2023.

Lukas Oesinghaus, Sören Becker, Larsen Vornholz, Efthymia Papalexi, Joey Pangallo, Amir Ali Moinfar, Jenni Liu, Alyssa La Fleur, Maiia Shulman, Simone Marrujo, et al. A single-cell cytokine dictionary of human peripheral blood. *BioRXiv*, 2025.

OpenAI. gpt-oss-20b model. https://developers.openai.com/api/docs/models/gpt-oss-20b, 2025. Official model documentation.

OpenAI. Gpt-5.6 sol. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026. Official model documentation, accessed September 2026.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Qwen Team. Qwen3.6-35b-a3b model card. https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026a. Official model card, accessed September 2026.

Qwen Team. Qwen3.8-27b model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026b. Official model card, accessed September 2026.

Qwen Team. Qwen3.8-Flash-Next model card. https://huggingface.co/Qwen/Qwen3.8-Flash-Next, 2026c. Official model card.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature biotechnology*, 42(6):927–935, 2024.

Shanghai AI Laboratory. Intern-S1-mini model card. https://huggingface.co/internlm/Intern-S1-mini, 2025. Official model card.

Shanghai AI Laboratory. Agents-A1 model card. https://huggingface.co/InternScience/Agents-A1, 2026. Official model card.

Tencent. Hunyuan-1.8B-Instruct model card. https://huggingface.co/tencent/Hunyuan-1.8B-Instruct, 2025. Official model card.

Tencent. Hy3-FP8 model card. https://huggingface.co/tencent/Hy3-FP8, 2026. Official model card.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes–2022. *Nucleic acids research*, 51(D1):D539–D545, 2023.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation. In *International Conference on Learning Representations*, volume 2026, pp. 72795–72823, 2026.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T Whitfield, Craig Russell, Yassir El Mesbahi, et al. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, pp. 1–8, 2026.

Menghua Rachel Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations*, volume 2025, pp. 73598–73624, 2025.

Nicholas D Youngblut, Christopher Carpenter, Arshia Nayebnazar, Abhinav Adduri, Rohan Shah, Chiara Ricci-Tam, Jaanak Prashar, Rajesh Ilango, Noam Teyssier, Silvana Konermann, et al. scbasecount: An ai agent-curated, standardized, auto-updated single-cell data repository. *Cell*, 189 (19):5932–5944, 2026.

Xinyu Yuan, Xixian Liu, Jianan Zhao, Yashi Zhang, Hongyu Guo, and Jian Tang. Plausibility is not prediction: Contrastive evidence for llm-based cellular perturbation reasoning. *arXiv preprint arXiv:2606.01042*, 2026.

Z.ai. Glm-5.2 model cards. `https://huggingface.co/zai-org/GLM-5.2`, 2026a. Official open-weight model documentation, accessed September 2026.

Z.ai. GLM-5.3 model card. `https://huggingface.co/zai-org/GLM-5.3`, 2026b. Official model card.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. Cellverse: Do large language models really understand cell biology? volume 38, 2026a.

Jesse Zhang, Airol A Ubas, Valentine Svensson, Nicole Thomas, Vishvak Subramanyam, Christopher Carpenter, Aidan Winters, Richard de Borja, Noam Teyssier, Neha Thakar, et al. Tahoe-100m: Mapping drug-induced molecular phenotypes at single-cell resolution. *Cell*, 189(19):5945–5961, 2026b.

Haiteng Zhao, Chang Ma, FangZhi Xu, Lingpeng Kong, and Zhi-Hong Deng. Biomaze: Benchmarking and enhancing large language models for biological pathway reasoning. *arXiv preprint arXiv:2502.16660*, 2025.

Jiahao Zhao, Feng Jiang, Shaowei Qin, Zhonghui Zhang, Junhao Liu, Guibing Guo, Hamid Alinejad-Rokny, and Min Yang. Sc-arena: A natural language benchmark for single-cell reasoning with knowledge-augmented evaluation. 2026.

A Related Work

A.1 AI Virtual Cell Datasets and Models

AI virtual cells seek to represent cellular states and predict their responses across perturbations and contexts (Bunne et al., 2024). Large perturbation resources, including genome-scale Perturb-seq, scPerturb, and Tahoe-100M, provide experimentally grounded intervention–response measurements (Replogle et al., 2022; Peidli et al., 2024; Zhang et al., 2026b). Building on these data, single-cell foundation models learn transferable representations (Theodoris et al., 2023; Cui et al., 2024; **?**), while CPA, CellOT, GEARS, STATE, and TxPert model perturbation-induced state transitions using latent, transport-based, graph-based, or set-level architectures (Lotfollahi et al., 2023; Bunne et al., 2023; Roohani et al., 2024; Adduri et al., 2026; Wenkel et al., 2026). Language-oriented approaches such as Cell2Sentence and VCWorld further connect cellular measurements with natural-language or knowledge-grounded reasoning (Levine et al., 2024; Wei et al., 2026).

A.2 Biological Reasoning Benchmarks for LLMs

Recent benchmarks evaluate LLMs on increasingly biological forms of cellular reasoning. SOAR focuses on cell-type annotation (Liu et al., 2025); CellVerse and SC-Arena cover annotation, generation, perturbation prediction, and scientific question answering (Zhang et al., 2026a; Zhao et al., 2026); and PerturbQA and VCWorld's GeneTAK evaluate differential-expression and direction prediction for unseen perturbations (Wu et al., 2025; Wei et al., 2026). BioMaze and PertReasonQA further assess pathway reasoning and the mechanistic faithfulness of perturbation explanations (Zhao et al., 2025; Kim et al., 2026), while AssayBench extends evaluation to phenotypic CRISPR-screen prediction (De Brouwer et al., 2026).

B Experiment Setup

B.1 Configuration of Evaluated Models

Table 4 details the configuration of each evaluated model. All inferences are conducted using either the official API endpoints or local `vLLM` deployment for most open-source models. Our experiments are reproducible on a workstation equipped with two NVIDIA H200 GPUs.

Organization	Model	Release	Version	Reasoning?	Serving Backend
			Open-Source Models		
DeepSeek	DeepSeek-V4-Pro	2026-08	`deepseek-ai/DeepSeek-V4-Pro-0813`	✓	vLLM
	DeepSeek-V4.1-Flash	2026-09	`deepseek-ai/DeepSeek-V4.1-Flash`	✓	vLLM
Z.ai	GLM-5.2	2026-06	`zai-org/GLM-5.2`	✓	vLLM
	GLM-5.3	2026-08	`zai-org/GLM-5.3`	✓	vLLM
MiniMax	MiniMax-M3	2026-06	`MiniMaxAI/MiniMax-M3`	✓	vLLM
Moonshot AI	Kimi-K3	2026-07	`moonshotai/Kimi-K3`	✓	vLLM
OpenAI	GPT-OSS-20B	2025-08	`openai/gpt-oss-20b`	✓	vLLM
Shanghai AI Lab	Agents-A1	2026-06	`InternScience/Agents-A1`	✓	vLLM
	Intern-S1-mini	2025-08	`internlm/Intern-S1-mini`	✓	vLLM
Tencent	Hunyuan-1.8B-Instruct	2025-07	`tencent/Hunyuan-1.8B-Instruct`	✓	vLLM
	Hy3-FP8	2026-07	`tencent/Hy3-FP8`	✓	vLLM
Qwen	Qwen3.6-35B-A3B	2026-04	`Qwen/Qwen3.6-35B-A3B`	✓	vLLM
	Qwen3.8-27B	2026-08	`Qwen/Qwen3.8-27B`	✓	vLLM
	Qwen3.8-Flash-Next	2026-08	`Qwen/Qwen3.8-Flash-Next`	✓	vLLM
Meta	Llama-3.3-70B-Instruct	2024-12	`meta-llama/Llama-3.3-70B-Instruct`	×	vLLM
	Llama-4-Scout-17B-16E-Instruct	2025-04	`meta-llama/Llama-4-Scout-17B-16E-Instruct`	×	vLLM
			Proprietary Models		
OpenAI	GPT-5.6-Sol	2026-07	`gpt-5.6-sol`	✓	API
Google	Gemini-3.8-Flash	2026-09	`gemini-3.8-flash`	✓	API

Table 4: Details of the foundation models evaluated in CellChain.

B.2 Dataset Resources

Table 5 summarizes the experimental datasets and external biological resources used in CellChain, together with their data types and roles in benchmark construction.

Table 5: Inventory of the cross-scale data resources integrated in CellChain.

Resource	Data type	Role in CellChain
Core experimental datasets		
Tahoe-100M (Zhang et al., 2026b)	Chemical perturbation	Drug-induced single-cell transcriptional responses and perturbation conditions.
PBMC Parse10M (Oesinghaus et al., 2025)	Cytokine perturbation	Cytokine-induced responses across immune cell types and donors.
Replogle Perturb-seq (Replogle et al., 2022)	Genetic perturbation	CRISPRi single-cell transcriptional responses in K562 and RPE1 cells.
External functional endpoints and annotations		
PRISM (**?**)	Drug-response endpoint	Drug–cell response measurements and viability-related sensitivity endpoints.
DepMap (Tsherniak et al., 2017)	Genetic dependency	Gene-effect and dependency measurements across cellular models.
PubChem (Kim et al., 2023)	Chemical annotation	Compound identities, structures, and standardized chemical identifiers.
ChEMBL (Zdrazil et al., 2024)	Bioactivity annotation	Compound, target, assay, and bioactivity records.
BindingDB (Gilson et al., 2016)	Binding evidence	Protein–ligand binding affinities and interaction records.
UniProtKB (bat, 2025)	Protein annotation	Protein identifiers, functions, and molecular annotations.
Reactome (Milacic et al., 2024)	Pathway annotation	Curated biological pathways and hierarchical pathway relations.
Gene Ontology and GOA (Aleksander et al., 2023; Huntley et al., 2015)	Functional annotation	Ontology terms and evidence-coded gene-function annotations.
Cellosaurus (Bairoch, 2018)	Cellular context	Cell-line identities, aliases, cross-references, and background information.
scBaseCount (Youngblut et al., 2026)	Sample and tissue context	Standardized single-cell sample, tissue, and cellular-context metadata.
CORUM (Tsitsiridis et al., 2023)	Protein-complex annotation	Protein-complex composition and subunit relations.
IntAct (Del Toro et al., 2022)	Molecular interaction	Experimentally supported molecular interaction records.
SIGNOR (Lo Surdo et al., 2023)	Signed mechanism relation	Directed activation and inhibition relations for mechanistic reasoning.
Open Targets (Ochoa et al., 2023)	Target-association evidence	Drug–target and target–disease associations with supporting evidence.
	Total source classes	**17**

C PROMPTS

We provide one representative instantiated prompt for each CellChain QA type: Gene Dependency Prediction in Figure 4, Cytokine-Axis Inverse Recovery in Figure 5, and Drug-Sensitivity Ranking in Figure 6.

Gene Dependency Prediction Prompt

SYSTEM

You are a molecular and cellular biology expert. Evaluate the biological question using only the supplied observations and visible evidence, and select the best-supported answer from the candidates. Entity names, accessions, aliases, and database identifiers may be masked. Do not attempt to recover the hidden entity identity using external knowledge, use hidden labels, or invent unavailable dependency measurements. Return exactly one valid JSON object conforming to the required output schema, without additional explanation or formatting.

USER

Question

Infer the held-out gene-fitness dependency endpoint category for an anonymous CRISPRi perturbation in K562.

Task Requirements

General protein annotations and pathway membership provide indirect biological evidence and should not be treated as measurements of the dependency endpoint. Infer the endpoint only from the evidence visible under the current exposure condition.

Evidence Exposure Condition

`Pathway context`

Query Context

Cell line: `K562`.
Perturbation type: `CRISPRi`.
Perturbation entity: `ANON_PERTURBATION_GENE`.

Visible Evidence

The anonymous perturbation target is associated with the following Reactome pathway annotations: Signaling by EGFR, Signaling by Receptor Tyrosine Kinases, Signaling by VEGF, Signal Amplification, NOD1/2 Signaling, NLR Signaling, Hemostasis, and Immune System. Explicit gene names, protein names, accessions, aliases, and database identifiers are withheld.

Candidate Answers

A: A strong-dependency endpoint category with a substantially negative CRISPR gene-effect value.
B: A moderate-dependency endpoint category with an intermediate negative CRISPR gene-effect value.
C: A weak-or-no-dependency endpoint category with a gene-effect value near the nonessential range.

Required Output

```
{
  "evidence_sufficiency":  "sufficient|weak|insufficient",
  "best_effort_dependency_label":
    "strong_dependency|moderate_dependency|
    weak_or_no_dependency",
  "effect_direction":
    "stronger_dependency|weaker_dependency|
    neutral_or_unknown",
  "confidence":  "low|medium|high"
}
```

Figure 4: Representative instantiated prompt for Gene Dependency Prediction under the pathway-context condition. Direct entity identifiers and database cross-references are masked, while task-relevant functional annotations remain visible. The dependency endpoint and reference answer are withheld from the model; evidence records are abbreviated for readability.

Cytokine-Axis Inverse Recovery Prompt

SYSTEM

You are a molecular and cellular biology expert. Evaluate the biological question using the supplied observations and context, and select the best-supported answer from the candidates. The queried perturbation may be anonymized. Use only the information provided in the input and do not infer information that is absent. Return exactly one valid JSON object without additional explanation or formatting.

USER

Question

Predict the held-out cytokine ligand/receptor axis for an anonymous cytokine stimulus in `CD14 Mono` cells.

Task Requirements

Use the measured single-cell response together with the visible biological-process and system-context evidence.

Observations and Context

Measured response: relative to matched PBS controls, representative top-down genes are `GBP1` (-0.371531), `CXCL10` (-0.249265), `MARCKS` (-0.246084), `CD163` (-0.236194), and `ANKRD22` (-0.219791), ...; representative top-up genes are `PLXDC2` (0.304950), `CD68` (0.257540), `SIGLEC1` (0.250829), `CD74` (0.244919), and `ACTB` (0.208748), ...
Biological-process evidence: helper T-cell enhancement of adaptive immune response; leukocyte chemotaxis; induction of positive chemotaxis; chemotaxis; T-cell activation; regulation of T-cell activation; positive regulation of T-cell proliferation; and T-cell receptor signaling.
System context: `CD14 Mono` cells from donors `Donor1`, `Donor3`, `Donor5`, `Donor6`, `Donor8`, `Donor10`, and `Donor12`.

Candidate Answers

A: ligand genes = `LEP`; receptor genes = `LEPR`.
B: ligand genes = `IFNW1`; receptor genes = `IFNAR1, IFNAR2`.
C: ligand genes = `IL1A`; receptor genes = `IL1R1, IL1RAP`.
D: ligand genes = `IL16`; receptor genes = `CD4`.

Required Output

```
{"selected_answer_label": "A|B|C|D"}
```

Figure 5: Representative instantiated prompt for Cytokine-Axis Inverse Recovery; repeated evidence records are abbreviated for readability.

Drug-Sensitivity Ranking Prompt

SYSTEM

You are a molecular and cellular biology expert. Evaluate the biological question using the supplied observations and visible evidence, and select the best-supported answer from the candidates. The queried treatment and cell-model identities may be anonymized. Use only the information provided in the input and do not infer unavailable endpoint values. Return exactly one valid JSON object without additional explanation or formatting.

USER

Question

Which candidate cell-line multidose response trajectory is more sensitive to the fixed treatment stimulus `ANON_STIMULUS`?

Task Requirements

Use the balanced response summaries and reference-history evidence. Lower held-out external AUC indicates stronger sensitivity. The held-out query AUC is not used to construct the reference histories.

Observations and Context

Dose response: At $[0.05, 0.5, 5.0]$ μM, mean absolute response trajectories are $[0.29015, 0.27761, 0.45911]$ for A and $[0.48819, 0.38289, 0.96108]$ for B.

Gene extremes: At high dose, A includes down-regulated `7SK`, `MTND2P28`, and `RNF26`, and up-regulated `ASMTL-AS1`, `RBMS3-AS3`, and `YBX1P10`; B includes down-regulated `ABHD6`, `MTND2P28`, and `TCIRG1`, and up-regulated `AFP`, `LBH`, and `USP38-DT`.

Module summaries: High-dose and low-to-high signed responses are provided for apoptosis/stress, cell cycle, DNA damage/p53, interferon, MAPK/ERK, NF-κB, PI3K/AKT/mTOR, and senescence/SASP modules.

Reference histories: Development-set sensitivity percentiles associated with the corresponding candidate cell models include 0.583333 for A and 0.062500, 0.062500, and 0.055556 for B, ... Larger percentiles indicate greater relative sensitivity within the corresponding reference compound.

Target evidence: BindingDB records link the anonymous ligand to `PLK1` ($IC_{50} = 0.87$–2.4 nM) and `BRD4` ($K_d = 77$–255 nM; $IC_{50} = 1070$ nM), with additional UniProt context for `BRD4`.

Candidate Answers

A: First anonymous cell-line multidose response trajectory.
B: Second anonymous cell-line multidose response trajectory.

Required Output

```
{"preferred_answer_label": "A|B"}
```

Figure 6: Representative instantiated prompt for Drug-Sensitivity Ranking under the matched-reference-history condition; repeated evidence records are abbreviated for readability.

D EVIDENCE ATTRIBUTION

The controlled-evidence analysis relies on task-specific evidence constructions and paired evaluation conditions. We therefore report additional details on case eligibility, benchmark selection, and evidence construction, together with the condition-level measurements underlying the contrasts in the main text.

Case Eligibility and Benchmark Selection. Across tasks, a case is considered eligible only when its biological entity and experimental context can be aligned, the required task-specific measurements and evidence are available, and the reference target maps to a unique candidate answer. The question, candidate set, reference target, and scoring rule are then held fixed across matched evidence conditions, with prohibited query-endpoint fields excluded from model-visible inputs. Failure of any applicable criterion excludes the corresponding case and all of its matched conditions.

Source-record and candidate-case counts use different units and therefore do not define a direct attrition sequence. For Gene Dependency, 2,448 endpoint-labeled candidates entered the evidence-availability and boundary-margin checks, of which 2,092 satisfied the eligibility criteria. We then sampled 40 cases from each dependency-category–cell-line stratum to construct the 240-case benchmark set. For Cytokine Axis, 267 cases satisfied the task-specific eligibility criteria, from which 110 cases were selected with ten cases per target axis. For Drug Sensitivity, 737 candidate pairs satisfied the curve-quality, complete-dose-trajectory, and AUC-separation criteria; 240 pairs were selected for benchmark evaluation to increase compound coverage while limiting reuse of drug–cell-model units. Differences between eligible and benchmark case counts therefore reflect benchmark-set construction rather than additional case-level quality-control failures. The frozen construction artifacts do not provide mutually exclusive counts for every earlier rejection criterion, so we do not further partition the remaining source records by failure reason.

Cytokine Process Evidence. The Cytokine conditions retain the measured transcriptional response while varying the additional process and system context. Process annotations are linked to the recorded cytokine stimulus, whereas system evidence describes the corresponding cellular context. The Process–Response contrast therefore measures the predictive value associated with the target-linked process context supplied to the model. Response–mechanism integration is not separately identified by this contrast. The Joint condition combines Process and System evidence under the same overall evidence budget by reallocating the available evidence units between the two sources. Accordingly,

$$\Gamma = A_{\text{Joint}} - A_{\text{Process}} - A_{\text{System}} + A_{\text{Response}}$$

is treated as a joint-condition deviation from additivity under this budget constraint.

Drug Reference-History Construction. Drug reference histories are constructed from 36 development compounds that are disjoint from the query compounds. Development records are retained after PRISM quality control requiring median fit $R^2 \geq 0.5$, replicate AUC range ≤ 0.1, positive slope, and ordered asymptotes. Within each development compound, relative sensitivity is represented by within-compound AUC-based percentiles for the corresponding cell models. Query-compound AUC values and query transcriptional responses are not used to construct or select these reference-history records.

The Matched and Swapped conditions contain the same query-response information and use the same history-comparison instruction. In the Matched condition, each reference history remains associated with the cell model from which it was derived; in the Swapped condition, the histories are exchanged between the two candidates. The Matched–Swapped contrast therefore measures sensitivity to candidate–history assignment conditional on the supplied query and reference-history evidence.

Drug Reference-History Signal Diagnostic. After selecting the 240-case Drug benchmark cohort, we evaluated whether its development-set reference histories contained predictive signal using a simple evidence-only rule. For each candidate cell model, the rule averages its sensitivity percentiles across available development-compound histories and selects the candidate with the higher mean. Predictions are scored against the held-out query label, and accuracy is averaged equally across the 26

query compounds. The rule attains 64.2% compound-weighted accuracy, exceeding the prespecified 60% cohort-level threshold.

The diagnostic is applied to the fixed benchmark cohort and does not remove individual cases. Because its evaluation uses held-out query labels after benchmark-set selection, we interpret the resulting accuracy as a post-selection measure of reference-history informativeness. The Drug evidence contrasts reported in the main text therefore characterize the benchmark cohort rather than all 737 eligible candidate pairs.

Condition-Level Measurements. Table 6 reports biological-cluster-weighted accuracy under each evidence condition. These values provide the condition-level measurements underlying the controlled-evidence contrasts in Table 3b. Gene Dependency contrasts weight perturbed genes equally, Cytokine Axis contrasts weight cytokine axes equally, and Drug Sensitivity contrasts weight query compounds equally. For Drug Sensitivity, both candidate-order presentations remain paired within each case, and invalid outputs receive zero score under the fixed denominator.

Table 6: Biological-cluster-weighted accuracy across four evidence conditions per task (%). Differences of unrounded condition values reproduce Table 3b; displayed values may differ slightly after rounding. This weighting differs from the case-weighted endpoint accuracies in Table 3a. Drug retains both candidate orderings, and invalid outputs score zero.

Model	Gene Dependency				Cytokine Axis				Drug Sensitivity			
	Base	Mol.	Path.	Trans.	Resp.	Proc.	System	Joint	Back.	Query	Matched	Swapped
Open-Source Models												
DeepSeek-V4-Pro	36.1	36.5	44.7	43.6	23.6	65.5	24.5	66.4	50.8	52.9	63.4	36.5
GLM-5.2	36.1	36.1	47.9	47.5	15.5	60.0	15.5	53.6	49.4	42.6	58.4	38.6
MiniMax-M3	36.1	36.5	46.8	45.0	18.2	61.8	11.8	65.5	49.6	43.4	60.7	35.7
Kimi-K3	36.1	53.2	49.8	50.9	17.3	78.2	24.5	70.0	41.5	41.8	63.6	36.6
DeepSeek-V4.1-Flash	36.1	41.8	52.1	51.4	30.0	69.1	33.6	60.9	50.1	40.5	42.7	48.0
GLM-5.3	36.1	49.5	53.0	53.4	20.0	74.5	24.5	70.9	49.9	44.4	60.7	34.8
Qwen3.6-35B-A3B	36.1	36.1	36.5	37.0	15.5	61.8	13.6	60.0	50.0	53.8	62.1	38.2
Qwen3.8-27B	36.1	36.1	36.1	36.8	10.9	64.5	10.0	59.1	47.9	46.9	64.0	35.8
GPT-OSS-20B	36.1	36.1	36.1	36.1	19.1	40.9	14.5	38.2	49.4	51.2	64.2	35.8
Agents-A1	36.1	36.1	37.4	37.7	17.3	58.2	14.5	45.5	50.0	47.2	47.6	48.5
Intern-S1-mini	36.1	36.1	36.5	35.6	7.3	48.2	8.2	40.0	50.0	46.0	43.2	55.0
Hunyuan-1.8B-Instruct	28.3	28.3	28.8	27.6	24.5	24.5	24.5	25.5	50.0	50.0	50.0	50.0
Qwen3.8-Flash-Next	36.1	36.1	51.6	46.6	25.5	81.8	20.9	74.5	50.0	47.6	62.2	36.1
Llama-3.3-70B-Instruct	36.1	36.1	36.1	36.1	11.8	39.1	9.1	32.7	50.0	50.1	39.2	59.8
Llama-4-Scout-17B-16E-Instruct	36.1	36.1	36.1	36.1	16.4	48.2	15.5	36.4	50.0	50.3	42.4	58.9
Hy3-FP8	36.1	36.1	36.1	36.1	10.9	68.2	11.8	60.9	50.0	42.1	50.7	45.7
Proprietary Models												
GPT-5.6-Sol	36.1	40.9	54.3	54.3	28.2	79.1	27.3	71.8	50.2	41.3	60.8	35.1
Gemini-3.8-Flash	36.1	54.1	49.8	50.2	30.0	80.9	31.8	75.5	50.0	42.8	63.7	37.2

E Biological-Cluster Bootstrap Confidence Intervals

We estimate uncertainty using 10,000 biological-cluster bootstrap replicates with the task-specific clusters defined in Section 3.2. In each replicate, clusters are sampled with replacement while retaining all associated cases, candidate-order presentations, evidence conditions, and model predictions jointly. For Drug Sensitivity, both candidate-order presentations of each case remain within the same compound cluster. Endpoint accuracy averages presentation scores within each case and weights cases equally, whereas controlled contrasts weight biological clusters equally. We report percentile-based 95% confidence intervals using the 2.5th and 97.5th percentiles of the bootstrap distribution.

Table 7 reports model-level case-weighted endpoint accuracies and their 95% confidence intervals under the prespecified target-evidence conditions. Table 8 reports the corresponding aggregate controlled-evidence effects across the 18-model panel, including incremental, joint-deviation, and assignment contrasts. The latter averages paired case-score differences within biological clusters and then averages the resulting model-level effects across the fixed model panel.

These intervals are marginal under a single perturbation-clustering scheme. Shared donors, controls, cell models, and reference histories may induce dependence across clusters, and the intervals do not

establish robustness to such crossed dependence. The model panel is fixed, so they do not quantify uncertainty over model families or future models. Intervals containing zero should not be interpreted as evidence of equivalence.

Table 7: Case-weighted endpoint accuracy with 95% biological-cluster bootstrap confidence intervals for the formal 18-model panel (%).

Model	**Gene Dependency**	**Cytokine Axis**	**Drug Sensitivity**
	Open-Source Models		
DeepSeek-V4-Pro	42.1 [35.6, 48.7]	66.4 [42.7, 87.3]	69.8 [61.4, 77.5]
GLM-5.2	46.7 [40.0, 53.4]	53.6 [30.0, 76.4]	62.9 [54.2, 71.4]
MiniMax-M3	44.2 [37.7, 50.8]	65.5 [45.5, 83.6]	67.7 [59.6, 75.1]
Kimi-K3	54.2 [47.4, 60.8]	70.0 [47.3, 89.1]	70.0 [61.6, 77.7]
DeepSeek-V4.1-Flash	52.9 [46.4, 59.5]	60.9 [39.1, 81.8]	43.8 [37.5, 50.2]
GLM-5.3	55.0 [48.1, 61.8]	70.9 [50.0, 89.1]	68.5 [59.9, 76.5]
Qwen3.6-35B-A3B	34.6 [28.4, 40.9]	60.0 [36.4, 81.8]	67.9 [60.1, 75.2]
Qwen3.8-27B	34.2 [28.0, 40.5]	59.1 [36.4, 80.9]	70.6 [62.4, 78.1]
GPT-OSS-20B	33.3 [27.2, 39.7]	38.2 [22.7, 53.6]	70.8 [62.7, 78.2]
Agents-A1	35.0 [28.9, 41.4]	45.5 [22.7, 68.2]	46.0 [38.8, 53.8]
Intern-S1-mini	32.9 [26.9, 39.3]	40.0 [20.9, 59.1]	40.4 [35.2, 45.9]
Hunyuan-1.8B-Instruct	32.5 [25.8, 39.3]	25.5 [19.1, 32.7]	50.0 [50.0, 50.0]
Qwen3.8-Flash-Next	47.9 [41.3, 54.6]	74.5 [53.6, 90.9]	69.6 [61.3, 77.1]
Llama-3.3-70B-Instruct	33.3 [27.2, 39.7]	32.7 [19.1, 48.2]	33.3 [26.7, 40.6]
Llama-4-Scout-17B-16E-Instruct	33.3 [27.2, 39.7]	36.4 [21.8, 50.0]	35.4 [28.0, 43.7]
Hy3-FP8	33.3 [27.2, 39.7]	60.9 [40.9, 80.0]	55.4 [51.6, 59.0]
	Proprietary models		
GPT-5.6-Sol	57.5 [51.0, 64.0]	71.8 [50.9, 90.0]	68.8 [60.3, 76.3]
Gemini-3.8-Flash	53.3 [46.6, 60.0]	75.5 [50.9, 94.5]	70.2 [61.7, 78.0]

Table 8: Aggregate controlled-evidence effects across the formal 18-model panel (percentage points).

Task	**Effect**	**Aggregate Δ_{Acc}**	**95% CI**
Gene Dependency	Molecular − Base	+3.3	[+2.2, +4.5]
	Pathway − Molecular	+3.8	[+1.7, +5.9]
	Transcriptional − Pathway	−0.4	[−1.1, +0.3]
Cytokine Axis	Process − Response	+42.4	[+28.2, +56.0]
	System − Response	−0.3	[−2.0, +1.4]
	Joint deviation	−5.1	[−8.9, −1.6]
Drug Sensitivity	Query Response − Background	−3.0	[−6.6, +1.0]
	Matched − Background	+6.2	[−0.3, +12.2]
	Matched − Swapped	+13.0	[+0.5, +24.5]