CELLCANON: AN AGENTIC CELLULAR WORLD MODEL FOR PREDICTION, EXPLANATION, AND OBSERVATION-DRIVEN UPDATING

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological discovery requires using molecular knowledge and new measurements to infer responses beyond the conditions already observed. We introduce CellCanon, the first agentic cellular world model to unify single-cell expression prediction, evidence-grounded explanation, and observation-driven updating. Its predictive core reuses condition-level biological evidence across gene-specific numerical readouts to guide single-cell expression generation. A shared world state links predictions, acquired measurements, and source-linked claims, supporting explanation and revision. On 339 held-out SciPlex3 conditions across three cell lines, CellCanon improves differential-expression AUPRC and Macro-F1 over the best baseline for each metric by 63.3% and 17.3%, respectively, and direction Macro-F1 by 7.2%. Relative to gene-wise reasoning on matched drug–cell cohorts, measured language-model token use falls by 99.77% for Predict and 98.41% for Explain. On the source-grounded benchmark, Explain correctly answers all answerable questions with supporting proofs and identifies all insufficient-evidence cases. In simulated experimental feedback, Update uses one measured source context per drug to exceed the best fine-tuned baselines in differential-expression AUPRC and Macro-F1 without changing model weights. Together, these capabilities support a cumulative account of cellular responses, with predictions and biological judgments that can be revised as evidence accumulates.

1 INTRODUCTION

Single-cell perturbation screens reveal drug-induced transcriptional changes but leave many compounds and cell contexts unmeasured (Srivatsan et al., 2020; Bunne et al., 2024). A cellular world model must combine molecular knowledge and new measurements to predict unobserved responses while keeping biological judgments inspectable and revisable.

Numerical perturbation models efficiently predict expression profiles (Lotfollahi et al., 2023; Hetzel et al., 2022; Piran et al., 2024; Qi et al., 2024). Gene-wise language reasoning combines molecular knowledge and retrieved examples to infer response labels (Wei et al., 2026). Genes in one drug–cell condition share biological background but require distinct response estimates. We reuse this evidence interpretation across gene-specific numerical readouts, with distinct supervision for continuous response, differential expression (DE), and conditional direction (DIR).

We introduce CellCanon, the first agentic cellular world model to unify single-cell expression prediction, evidence-grounded explanation, and observation-driven updating. Encoded molecular evidence and a training-response atlas feed continuous multilayer perceptron (MLP) and classification readouts, whose gene-specific outputs guide single-cell generation (Figure 2). A shared world state links these predictions to source-linked claims and acquired measurements; observation-conditioned maps update response estimates (Figure 1).

Explain evaluates target actions and regulatory paths, providing proofs or identifying insufficient evidence. It distinguishes binding, signed regulation, and solvent-control RNA detection; claim dependencies allow revision when supporting records change.

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.

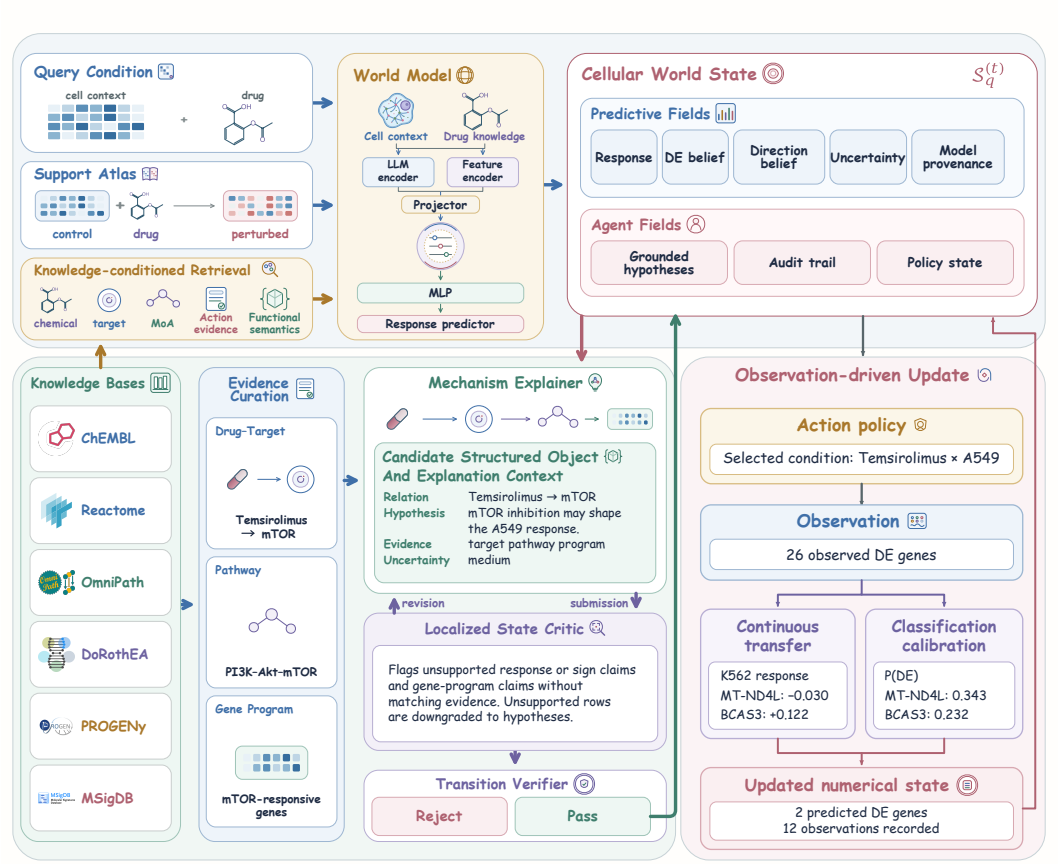

Figure 1: **An agent for prediction, explanation, and updating.** Evidence and a training-response atlas condition prediction. The world state links responses, source-linked claims, and measurements. Evidence review supports explanations; Update transfers observed responses across cell contexts with fixed predictor weights.

Update uses a measured response to revise predictions for the same drug in other, still-unmeasured cell contexts. We evaluate this process by progressively revealing existing SciPlex3 measurements to simulate experimental feedback. Separate observation-state tasks assess how measurement availability changes support for biological claims.

Experiments demonstrate improved differential-response prediction with substantially reduced language-model token use. Ablations support the contributions of biological annotations, measured cell context, and task-specific readouts.

2 Related Work

Predicting perturbation responses. Neural models predict perturbation responses (Lotfollahi et al., 2019; 2023; Hetzel et al., 2022; Piran et al., 2024; Qi et al., 2024). Optimal transport models shifts between cell distributions (Bunne et al., 2023; Driessen et al., 2026). LGR aggregates measured responses from LLM-selected neighboring drugs (Xiong et al., 2026); MapPFN supports in-context genetic-perturbation prediction without gradient updates (Sextro et al., 2026). Evaluations emphasize target and protocol choices (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026).

Agents for biological discovery. ChemCrow connects language models to chemistry tools; BioDiscoveryAgent designs genetic perturbation experiments (Bran et al., 2024; Roohani et al., 2025). PerturbAgent organizes perturbation analysis and prediction, while VCWorld studies language-based virtual-cell reasoning (Pei et al., 2026; Wei et al., 2026). PerTurboAgent updates

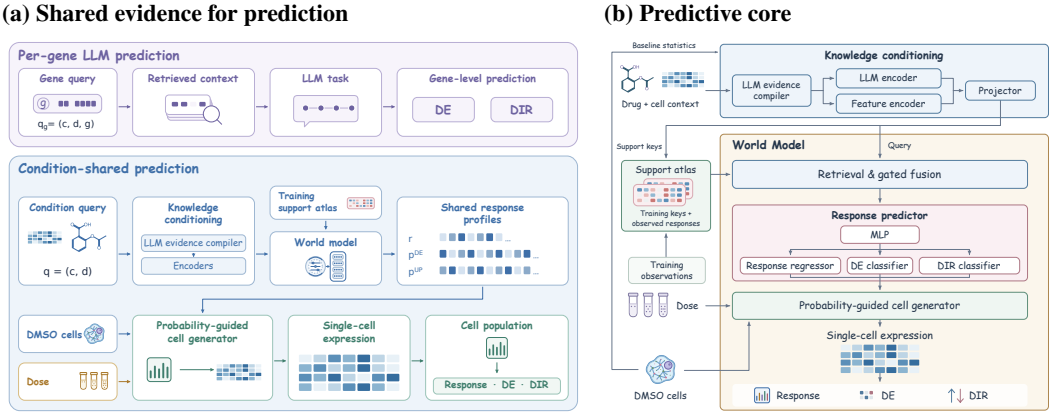

Figure 2: **Shared evidence for single-cell prediction.** **a**, Evidence compilation is shared across genes; continuous responses and DE/DIR probabilities guide generation from matched DMSO cells and doses. **b**, Text and structured features guide atlas retrieval and task readouts. Testing generated populations against DMSO gives final DE/DIR (Section B.4).

phenotype predictions and prioritizes experiments from accumulated Perturb-seq observations (Hao et al., 2025). BaiZe predicts expression and supports few-shot cross-species adaptation; BaiZe-Agent retrieves and explains saved results (Zeng et al., 2026). CellCanon links prediction, source-grounded explanation, and measurement-driven revision through a shared world state.

Evidence-grounded biological explanation. GeneAgent uses domain databases and verification for gene-set interpretation (Wang et al., 2025). Claim verification evaluates labels and rationales (Wadden et al., 2020); citation benchmarks distinguish answer correctness from citation quality (Gao et al., 2023). We evaluate judgments and supporting citations using biological evidence types, signed paths, and observation-driven changes. Molecular sources define reference answers and proofs.

3 METHOD

3.1 CELLULAR WORLD MODEL AND EVIDENCE STATE

The agentic cellular world model couples a learned response predictor and observation-conditioned update maps with a shared world state.

Let d denote a drug, c a cell context, and $q = (d, c)$ a condition. Training drugs $\mathcal{D}_{\text{tr}}$ supply the support atlas. Each query state contains a continuous response $\widehat{\mathbf{y}}_q$, DE probabilities $\mathbf{p}_q^{DE}$, and conditional up-regulation probabilities $\mathbf{p}_q^{\text{up}}$:

$$\mathcal{S}_q^{(t)} = \left(\widehat{\mathbf{y}}_q^{(t)}, \mathbf{p}_q^{DE,(t)}, \mathbf{p}_q^{\text{up},(t)}, \mathcal{E}_q, \mathcal{C}_q^{(t)}, \mathcal{O}_t\right), \tag{1}$$

where $\mathcal{E}_q$ contains source-linked evidence, $\mathcal{C}_q^{(t)}$ contains claims and their dependencies, and $\mathcal{O}_t$ records acquired measurements. The atlas supplies training responses; the observation record holds newly revealed measurements.

The knowledge base combines drug and gene descriptions, target and assay records, pathways, and signed regulatory relations, retaining each record's source identifier and evidence type (Section B). Solvent-control mean expression and detection rates provide cell-specific target context. Explanations use partial evidence or complete candidate paths.

3.2 PREDICT: LANGUAGE-SELECTED EVIDENCE MEETS AN OBSERVATION ATLAS

Biological evidence compilation. For each condition, the language model selects source-linked target evidence and qualifies its relevance and assay scope. The evidence packet combines drug knowledge and matched DMSO controls, distinguishing binding, functional action, background regulation, and control-cell expression.

Frozen text embeddings encode selected evidence and weighted target descriptions; structured features preserve evidence categories and quantitative context. Their concatenation $\mathbf{x}_q$ undergoes training-fitted projection and normalization, $\mathbf{z}_q = P_{\text{tr}}(\mathbf{x}_q)$ (Equation (6) and Section B). Source links are retained, and the representation is shared across genes.

Knowledge-conditioned retrieval. For each training drug i, the atlas stores knowledge features $\mathbf{z}_{ic}$ and normalized continuous responses, DE labels, and up-regulated DE labels. Learned maps produce a query $\mathbf{q}_{dc}$, key $\mathbf{k}_{ic}$, and observation value $\mathbf{v}_{ic}$. Retrieval combines learned compatibility with chemical, target, and mechanism-of-action similarities $s_r(d, i)$:

$$\ell_{di,c} = \frac{\mathbf{q}_{dc}^\top \mathbf{k}_{ic}}{\sqrt{H}} + \sum_{r=1}^{3} \text{softplus}(w_r) s_r(d, i), \qquad \alpha_{di,c} = \text{softmax}_{i \in \mathcal{A}_d}(\ell_{di,c}), \tag{2}$$

$$\mathbf{h}_{dc} = \text{LayerNorm}\left(\mathbf{q}_{dc} + \text{sigmoid}(W_g \mathbf{q}_{dc} + \mathbf{b}_g) \odot W_o \sum_{i \in \mathcal{A}_d} \alpha_{di,c} \mathbf{v}_{ic} \right). \tag{3}$$

A cell-specific map $\mathbf{u}_{dc} = \text{MLP}_c(\mathbf{h}_{dc})$ produces the shared readout representation. Here $\mathcal{A}_d$ contains permitted training donors. Whole-drug masking excludes the query identity and its training fold during representation fitting; test-time donors come from the training atlas.

Continuous and classification readouts. A cell-specific MLP maps the shared representation to the continuous gene panel:

$$\widehat{\mathbf{y}}_{dc} = \boldsymbol{\mu}_c + s_y \, \text{MLP}_c^{\text{RNA}}(\mathbf{u}_{dc}), \tag{4}$$

where $\boldsymbol{\mu}_c$ and s_y are training response normalization statistics. We train the frontend and continuous head using mean squared error, then freeze the representation and fit independent regularized logistic heads:

$$p_{dcg}^{DE} = \text{sigmoid}(b_{cg}^{DE} + \boldsymbol{\beta}_{cg}^{DE\top} \mathbf{u}_{dc}), \qquad p_{dcg}^{\text{up}} = \text{sigmoid}(b_{cg}^{\text{up}} + \boldsymbol{\beta}_{cg}^{\text{up}\top} \mathbf{u}_{dc}). \tag{5}$$

Each gene has its own coefficients; direction supervision uses only training DE examples. A validation-fitted decoder estimates each condition's DE count and selects genes with the highest DE probabilities; the direction head assigns up or down. Separately supervised magnitude, DE occurrence, and direction readouts jointly guide expression generation.

Single-cell generation. Matched DMSO cells, dose-dependent parameters, and the continuous and classification readouts generate single-cell expression (Figure 2). DE confidence and direction probabilities guide bounded response corrections and gene-detection rates using support-trained experts. Continuous metrics score generated expression; differential testing against DMSO yields population DE/DIR (Section B.4).

3.3 Explain: evidence-grounded reasoning and correction

Explain assesses source support for target action, control-cell expression, regulatory paths, and source statements, returning *supported*, *contradicted*, or *insufficient evidence* judgments with citations and natural-language rationales. Control-cell expression supplies cellular context; functional and binding records characterize drug–target interactions. Candidate drug–target–regulatory paths retain evidence types and edge signs. Tools enumerate unknown-sign assignments consistently across shared edges, and the language model selects eligible paths.

Review checks citations, biological interpretation, and path consistency, providing localized diagnostics for revision. Claims depend on background records, predictions, or measurements: new measurements can make claims answerable; source withdrawal can remove support. Section F specifies the source-question and observation-state tasks.

3.4 Update: incorporating newly measured responses

Update uses a newly measured drug response to revise predictions for the same drug in unmeasured cell contexts. The environment reveals a selected source condition's response and DE/DIR labels. A support-trained kernel ridge model combines measured source-response similarity with chemical

and biological similarities to infer the target response (Equations (12) and (13)). The parent predictor and update maps remain fixed. One or two source-cell responses can be used; without a source measurement, the initial prediction is retained.

Support-trained classification corrections combine parent probabilities with source responses and source–target label relationships. A source-activity model adjusts the expected DE count, and updated probabilities rank the genes. Measurements and revised predictions enter successive states. Section G gives the transfer, correction, and count-model settings.

4 Experiments

4.1 Tasks, data, and evaluation

Drug-disjoint prediction. We evaluate 339 held-out SciPlex3 conditions in A549, K562, and MCF7 (Srivatsan et al., 2020), pooling 24-hour responses across doses. Continuous scores cover 2,000 genes; classification retains all reference DE positives and randomly samples up to 200 negatives with unadjusted $p > 0.1$ per condition (Section A). We use the DeepSeek-V4-Flash API for evidence compilation and frozen Qwen3-Embedding-0.6B embeddings (Zhang et al., 2025).

Comparators and metrics. We compare BioLORD, chemCPA, CMonge, PRnet, Squidiff (He et al., 2026), PerturbNet (Yu et al., 2025), STATE (Adduri et al., 2025), VCWorld, and Random. The first four use validation-fitted DE adapters and validation-selected decision rules. Squidiff, PerturbNet, and STATE generate cell populations, from which differential testing against matched DMSO yields DE and direction predictions (Section C). VCWorld uses DeepSeek-V4-Flash.

Continuous-response metrics include Top50 mean-expression R^2, logFC Pearson, and Δ-Pearson for changes relative to DMSO; within-condition scores are averaged across test conditions. Top50/Top100 panels follow absolute reference Wilcoxon ranks. DE and DIR use AUPRC, AUROC, Macro-F1, accuracy, precision, and recall. DIR conditions on reference-DE queries with up-regulation as positive. Joint unchanged/up/down evaluation includes detection and direction errors (Section D).

Source-grounded explanation. The benchmark contains 966 questions across 72 conditions, with answerability determined by molecular source records. Correctness requires a supported/contradicted judgment and complete proof, or abstention for insufficient evidence (Section F).

Observation-feedback protocol. We simulate experimental feedback by revealing existing SciPlex3 measurements under fixed acquisition schedules. Each of three partitions withholds one target cell context per test drug; the other two supply measurements. One observation reveals a drug–cell response at 24 hours across all four doses. CellCanon is evaluated with and without Update under identical schedules and budgets. Feedback scores use condition-level response and classification readouts, averaged across partitions. Six response models are fine-tuned on revealed source-cell populations at 48 observations and one source context per drug. Target conditions remain unobserved (Sections G and H.1).

4.2 Predicting expression and differential responses

CellCanon achieves DE AUPRC of 0.8067 and Macro-F1 of 0.8164, versus 0.4938 and 0.6961 for the strongest baselines (Table 1). Joint unchanged/up/down Macro-F1 reaches 0.7420 versus 0.5400. Its DE precision of 0.8370 and recall of 0.5739 indicate more precise but less sensitive predictions than PRnet. Shared evidence analysis supports this gene-specific accuracy with substantially lower language-model token use (Section 4.6).

CellCanon reaches logFC Pearson of 0.6944 and Δ-Pearson of 0.3254, versus 0.4306 and 0.2543 for chemCPA (Table 2 and Figure 3). Section D.2 reports additional response-focused comparisons.

The classification advantage extends across drugs (Figure 4a–c). Across 112 drugs, median DE and joint gains over PRnet are 0.1322 (95% CI 0.1074–0.1553) and 0.1456 (0.1243–0.1699), respectively; 10,000 paired drug-bootstrap resamples support robustness to test-drug composition. Joint

Table 1: **Differential-response prediction on the 2,000-gene SciPlex3 panel.** DE/joint use sampled test queries; DIR uses reference-DE queries with up-regulation as positive. Joint classes are unchanged/up/down. F1 is Macro-F1 unless specified; Acc. is accuracy. VCWorld and Random use binary ranking scores (Section C). Higher is better. In all tables, **bold** and underlining mark best and second-best scores.

Metric	BioLORD	chemCPA	CMonge	PRnet	Squidiff	PerturbNet	STATE	VCWorld	Random	CellCanon
DE										
AUPRC	0.2694	0.1345	0.3474	0.4058	0.0875	0.4870	0.4255	0.4938	0.3666	**0.8067**
AUROC	0.7734	0.4536	0.8140	0.8572	0.1279	0.8807	0.8371	0.6527	0.4998	**0.9449**
F1	0.6184	0.4422	0.6479	0.6961	0.1086	0.6732	0.6040	0.5500	0.4324	**0.8164**
Acc.	0.6886	0.4616	0.7402	0.7698	0.1144	0.8453	0.8275	0.6259	0.4996	**0.9163**
Precision	0.3125	0.2092	0.3432	0.3941	0.1056	0.5060	0.4093	0.2481	0.1555	**0.8370**
DIR										
AUPRC	0.9042	0.7772	0.7671	0.8165	0.8222	0.7476	0.8652	0.6619	0.6558	**0.9402**
AUROC	0.8810	0.6996	0.7293	0.7589	0.7822	0.6744	0.8305	0.5459	0.4991	**0.9255**
F1	0.7903	0.6753	0.6699	0.7006	0.5852	0.6425	0.7418	0.5241	0.4982	**0.8472**
Acc.	0.7903	0.6756	0.6700	0.7006	0.6482	0.6461	0.7421	0.5316	0.4991	**0.8480**
Precision	0.8571	0.7556	0.7405	0.7582	0.6121	0.7589	0.7891	0.6104	0.5409	**0.8686**
Joint response: unchanged/up/down										
F1	0.4900	0.3821	0.4686	0.5400	0.0944	0.4910	0.4885	0.3767	0.2881	**0.7420**
Acc.	0.6564	0.4210	0.6970	0.7293	0.0854	0.8254	0.8252	0.5777	0.4607	**0.9104**
Unchanged F1	0.7821	0.5461	0.8281	0.8458	0.0363	0.9103	0.9015	0.7348	0.6277	**0.9518**
Up F1	0.3614	0.4260	0.2582	0.3620	0.1277	0.1407	0.3171	0.1758	0.1258	**0.6102**
Down F1	0.3264	0.1742	0.3196	0.4121	0.1191	0.4220	0.2470	0.2196	0.1107	**0.6640**

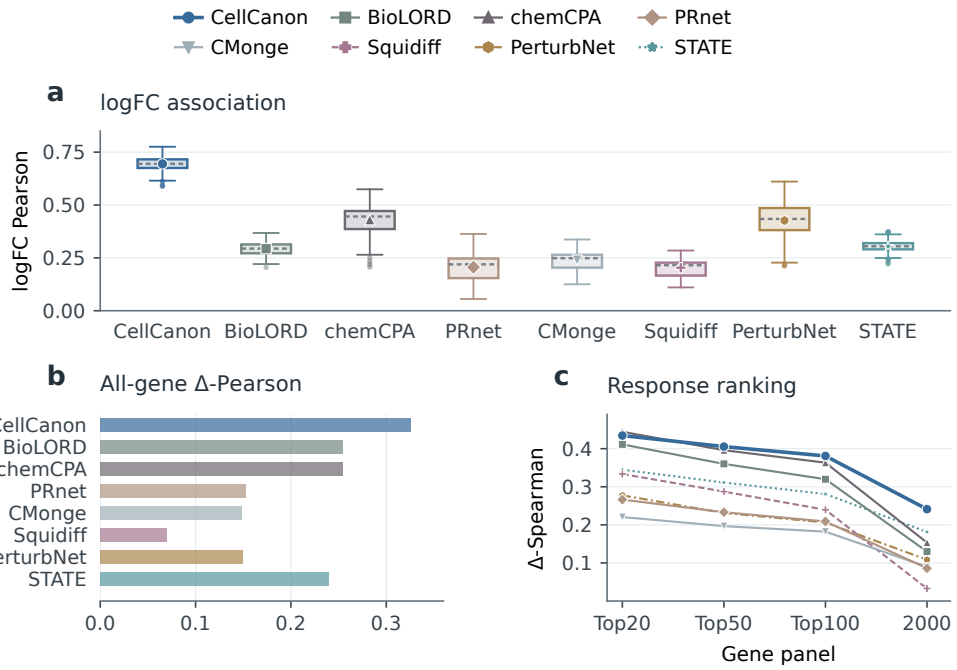

Figure 3: **Response-focused prediction.** **a**, Per-condition logFC Pearson on 2,000 genes: boxes show quartiles, gray dashed lines medians, whiskers the furthest points within 1.5 interquartile ranges, dots outliers, and model symbols means. **b**, Δ-Pearson on all 2,000 genes. **c**, Δ-Spearman across gene panels. Delta subtracts matched DMSO means. Panels use 339 conditions (333 for PerturbNet); **b–c** show means. Top-k genes follow absolute reference Wilcoxon ranks.

median gains remain positive across all four response-strength strata. DIR gains concentrate in stronger responses, while the strongest-response DE interval crosses zero.

Table 2: **Continuous-response prediction.** Scores average within-condition comparisons over evaluated test conditions. Expression R^2 and Δ-Spearman use the top 50 and 100 reference-ranked genes, respectively; logFC Pearson and Δ-Pearson use all 2,000. Delta subtracts matched DMSO means. Higher is better.

Metric	BioLORD	chemCPA	CMonge	PRnet	Squidiff	PerturbNet	STATE	DMSO reference	CellCanon
Top50 R^2	**0.9197**	0.9026	0.7359	0.9039	0.8590	0.4461	0.9063	0.8880	0.9145
logFC Pearson	0.2935	0.4306	0.2392	0.2061	0.2034	0.4268	0.3046	–	**0.6944**
Δ-Pearson	0.2538	0.2543	0.1484	0.1528	0.0703	0.1500	0.2401	–	**0.3254**
Top100 Δ-Spearman	0.3199	0.3632	0.1822	0.2089	0.2396	0.2062	0.2809	–	**0.3809**

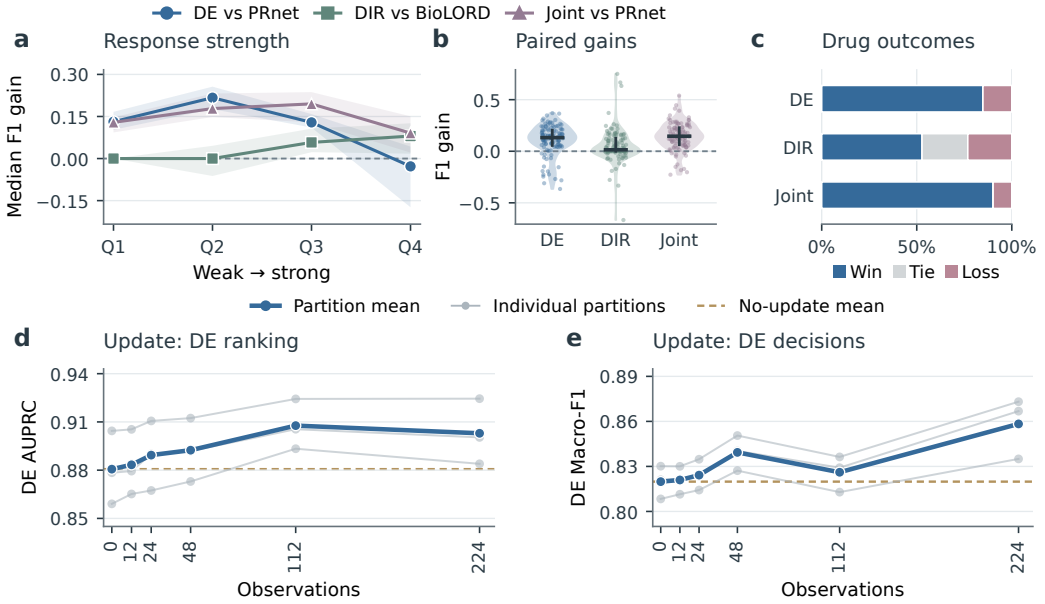

Figure 4: **Classification robustness and observation feedback. a–c**, Expression-derived Macro-F1 gains over fixed baselines (PRnet: DE/joint; BioLORD: DIR; Table 1): **a**, drug medians by response-strength quartile, with 95% paired drug-bootstrap intervals; **b**, paired drug gains, medians and interquartile ranges; **c**, win/tie/loss fractions. Scores pool three cell contexts and average equivalent identifiers (112 drugs; 108 for DIR; Section D.1). **d,e**, Updated condition-head DE AUPRC/Macro-F1: thin lines show three target-cell partitions, thick lines their mean, dashed lines the no-Update mean. Budgets 112/224 reveal one/two source contexts per canonical drug (Section G).

Beyond the 2,000-gene continuous head, CellCanon classifies 931 additional LINCS landmark genes using their support DE/DIR labels without their continuous response targets. Continuous generation for these genes uses a separate model fitted to their continuous support responses (Section J).

4.3 Updating unmeasured conditions from new observations

Six response models are fine-tuned on the revealed populations; CellCanon uses its fixed updater. With one source context per drug, CellCanon reaches DE AUPRC/Macro-F1 of 0.9077/0.8261, compared with 0.6279/0.7630 for the strongest fine-tuned comparator (Table 3). Joint unchanged/up/down Macro-F1 reaches 0.7431 versus 0.6116.

At 48 observations, CellCanon's DE Macro-F1 rises from 0.8199 to 0.8393 (Figure 4e). Section H.1 and Table 20 report fine-tuning settings, intermediate budgets, and changes from each method's initial prediction.

Table 3: **Prediction after observing one source context per drug.** Absolute scores average three source–target partitions over 2,000 genes, using Table 1 metrics. Fine-tuned comparators use differential tests on generated expression; CellCanon uses updated condition heads (Section H.1). Higher is better.

Metric	BioLORD	chemCPA	CMonge	PRnet	PerturbNet	STATE	CellCanon
DE							
AUPRC	0.1337	0.3188	0.1306	0.0948	0.6279	0.5535	**0.9077**
AUROC	0.3925	0.8069	0.3350	0.1467	0.9135	0.8663	**0.9792**
F1	0.2627	0.5786	0.1965	0.1241	0.7630	0.7003	**0.8261**
Acc.	0.2680	0.7199	0.1976	0.1338	0.8637	0.8555	**0.9185**
Precision	0.1236	0.2767	0.1023	0.1271	0.5846	0.5925	**0.9021**
DIR							
AUPRC	0.7850	0.6807	0.7836	0.8070	0.8305	**0.9368**	0.9313
AUROC	0.7425	0.6040	0.7269	0.7689	0.7607	**0.9185**	0.9167
F1	0.6691	0.5661	0.6608	0.6488	0.6889	0.8323	**0.8453**
Acc.	0.6709	0.5892	0.6693	0.6785	0.6937	0.8329	**0.8460**
Precision	0.7221	0.8084	0.6792	0.6530	0.8606	0.8750	**0.8840**
Joint response: unchanged/up/down							
F1	0.2400	0.4210	0.1859	0.1746	0.5983	0.6116	**0.7431**
Acc.	0.2460	0.6944	0.1769	0.1010	0.8365	0.8531	**0.9100**
Unchanged F1	0.3235	0.8224	0.2239	0.0327	0.9173	0.9158	**0.9528**
Up F1	0.1058	0.1092	0.1073	0.1364	0.2952	0.5220	**0.6008**
Down F1	0.2908	0.3312	0.2265	0.3548	0.5824	0.3969	**0.6756**

4.4 Evidence-linked explanations with explicit answerability

Explain correctly answers all answerable questions with complete proofs, abstains on all insufficient-evidence cases, and retrieves valid paths among the top three for all eligible conditions (Table 4). For TSA (Figure 5), functional records support HDAC1 inhibition and K562 controls establish HDAC1 expression. Reviewing unknown-sign alternatives changes two reverse-sign pathway judgments from contradicted to insufficient evidence. Observation-state and withdrawal tests track revisions as measurements arrive or sources are removed (Section F).

Table 4: **Source-grounded explanation.** Questions cover 72 conditions; path success uses the 63 eligible conditions.

Endpoint	Correct / total
Label + complete proof	561 / 561
Correct abstention	405 / 405
Valid source path@3	63 / 63

4.5 Biological context and independent readouts

Three ablations assess biological information and task-specific supervision (Table 5). Removing biological annotations lowers DE Macro-F1 from 0.8077 to 0.7117. Removing measured DMSO expression and detection lowers Δ-Pearson from 0.3074 to 0.2810. Replacing independent classification heads with continuous-response scores lowers DE Macro-F1 to 0.5278 with validation-selected thresholds, supporting distinct supervision for response magnitude and DE occurrence. Paired drug-bootstrap intervals for these decreases exclude zero (Section E).

Table 5: **Predict component ablations.** Pre-generation readouts on 339 conditions and 2,000 genes. F1 is macro-averaged; the atlas is retained in the context comparison.

Biological annotations	DE F1	DIR F1
Full predictor	0.8077	0.8393
Without annotations	0.7117	0.7900
Measured context	Δ-Pearson	Δ-Spearman
With DMSO features	0.3074	0.2383
Without DMSO features	0.2810	0.2292
Task readouts	DE F1	DIR F1
Independent heads	0.8077	0.8393
Continuous + count	0.4538	0.8134
Continuous + threshold	0.5278	0.8134

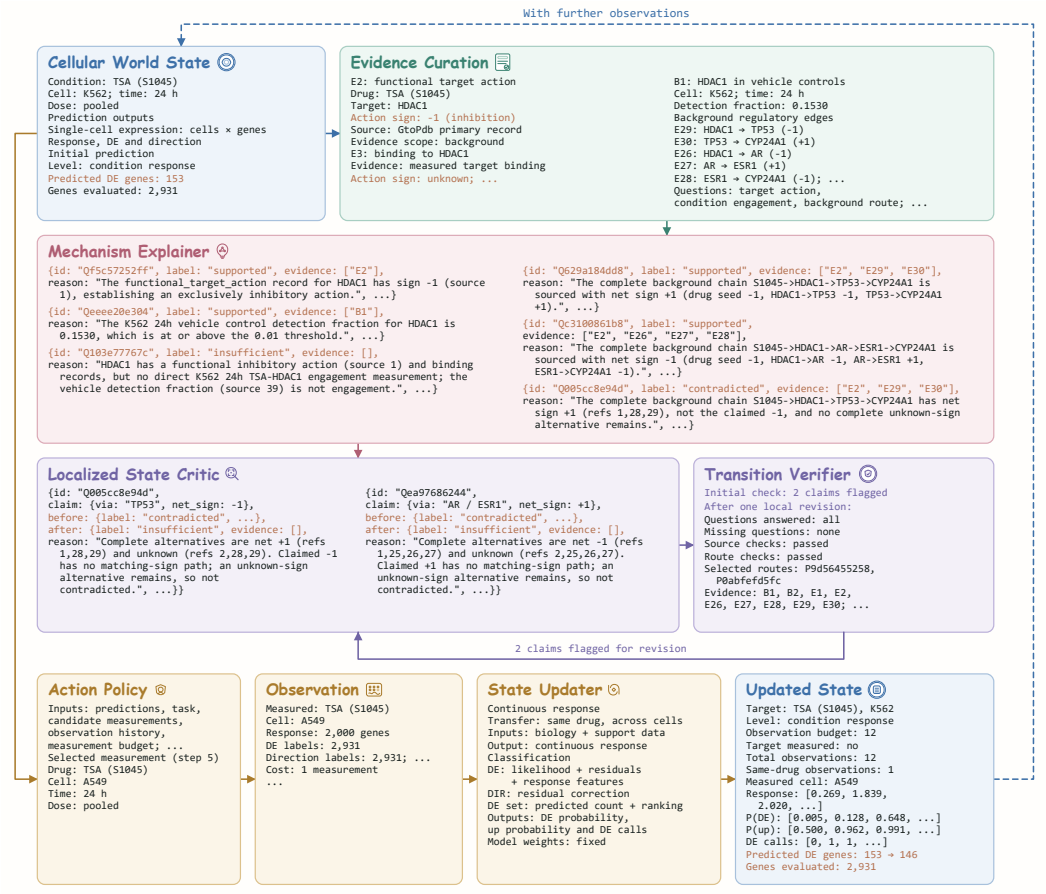

Figure 5: **Explain and Update for Trichostatin A (TSA).** HDAC1 evidence review changes two K562 pathway judgments from contradicted to insufficient evidence (purple loop). A separate Update experiment revises the K562 prediction from an A549 measurement. The blue loop indicates further observations (Section F.4).

4.6 Reusing evidence analysis reduces token use

Condition-level evidence reuse reduces repeated language-model work. Relative to VCWorld's gene-wise prediction and chain-of-thought explanation on matched drug–cell cohorts, CellCanon reduces language-model token use by 99.77% for Predict across 339 conditions and 98.41% for Explain across 72 conditions. Evidence analysis is shared across genes; explanation is invoked separately. Section I gives token totals and cohort definitions.

5 Discussion and Conclusion

CellCanon connects accurate single-cell response prediction, source-grounded explanation, and observation-driven updating in an agentic cellular world model. Shared evidence supports gene-specific readouts and generation with substantially lower language-model token use. The world state links predictions to the evidence and observations needed to check and revise them. In simulated experimental feedback, measured responses improve predictions for unobserved cell contexts of the same drug.

AI USE STATEMENT

AI tools were used to polish writing, support literature retrieval and discovery, and assist with part of the research execution. The use of language models as components of CellCanon itself is described in the main text and appendices.

REPRODUCIBILITY STATEMENT

Sections A to C document the data splits, prediction targets, model architecture, training, and baseline configurations. Sections D and E define the evaluation metrics, ablations, and paired bootstrap procedures; Sections F to H specify the Explain benchmark and Update protocol. Section I details token accounting, and Section J describes the additional gene panel. An anonymized ZIP archive containing the core code has been submitted as supplementary material. The archive provides source code and instructions for obtaining public inputs, preprocessing data, configuring software dependencies, and running the core Predict, Explain, and Update workflow.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

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*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

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. doi: 10.1016/j.cell.2024.11.015.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, Marianna Rapsomaniki, and Jannis Born. Conditional Monge Gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, 8:984–996, 2026. doi: 10.1038/s42256-026-01242-8.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, 2019. doi: 10.1101/gr.240663.118.

Minsheng Hao, Yongju Lee, Hanchen Wang, Gabriele Scalia, and Aviv Regev. PerTurboAgent: An LLM-based agent for designing iterative Perturb-Seq experiments. In *Proceedings of the 20th Machine Learning in Computational Biology meeting*, volume 311 of *Proceedings of Machine Learning Research*, pp. 44–64, 2025. URL https://proceedings.mlr.press/v311/hao25b.html.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi, and Kam W. Leong. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 23:65–77, 2026. doi: 10.1038/s41592-025-02877-y.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa933b5abc1be30baece1d230ec575a7-Abstract-Conference.html.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8.

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, Jay Shendure, José L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023. doi: 10.15252/msb.202211517.

Kanglu Pei, Sisi Qu, Philip H. S. Torr, Jonathan G. Hedley, and Christian Schroeder de Witt. PerturbAgent: An agentic AI system for analysis and prediction of genetic perturbations. In *AAAI 2026 Workshop XAI4Science*, 2026. URL https://openreview.net/pdf?id=FfHATDCkCQ.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42:1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, and Yi Zhao. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15:9256, 2024. doi: 10.1038/s41467-024-53457-1.

Eliot Ragueneau, Chuqiao Gong, Pierre Sinquin, Cristoffer Sevilla, Deidre Beavers, Alexander Grentner, Johannes Griss, Gregory F J Hogue, Nancy T Li, Lisa Matthews, Bruce May, Marija Milacic, Helia Mohammadi, Robert Petryszak, Karen Rothfels, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Joel Weiser, Adam Wright, Marc Gillespie, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome knowledgebase 2026. *Nucleic Acids Research*, 54(D1):D673–D681, 2026. doi: 10.1093/nar/gkaf1223.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html.

Marvin Sextro, Weronika Kłos, and Gabriel Dernbach. MapPFN: Learning causal perturbation maps in context. *arXiv preprint arXiv:2601.21092*, 2026. URL https://arxiv.org/abs/2601.21092.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367:45–51, 2020. doi: 10.1126/science.aax6234.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, Tamás Korcsmáros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021. doi: 10.15252/msb.20209923.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44:1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025. doi: 10.1038/s41592-025-02748-6. URL https://www.nature.com/articles/s41592-025-02748-6.

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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/767ff070c27c8954babe74eccf3fbe91-Abstract-Conference.html.

Betty Xiong, Jan-Christian Huetter, Gabriele Scalia, Tommaso Biancalani, and Sepideh Maleki. LLM-guided retrieval for prediction of molecular perturbation responses. *arXiv preprint arXiv:2608.01734*, 2026. URL https://arxiv.org/abs/2608.01734.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 2025. doi: 10.1038/s44320-025-00131-3.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. 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. doi: 10.1093/nar/gkad1004.

Qingwu Zeng, Wenxiang Cai, Renchong Tian, Qi Wang, Dongzhan Zhou, Minting Pan, Hai Yang, Zhe Liu, Guan Ning Lin, and Zhe Wang. BaiZe: A multi-view dynamic framework for simulating and interpreting cellular responses across perturbation contexts. *bioRxiv*, 2026. doi: 10.64898/2026.07.15.738608. URL https://www.biorxiv.org/content/10.64898/2026.07.15.738608v1.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A DATA AND PREDICTION TARGETS

The SciPlex3 evaluation uses A549, K562, and MCF7 at 24 hours, with the available dose conditions pooled into a drug–cell response. Each continuous target is a 2,000-gene log-fold-change vector relative to the corresponding control. The classification panel contains these genes and 931 additional genes with differential-expression and direction labels. The additional panel contributes classification supervision and atlas values, while the continuous output dimension remains 2,000.

DE labels compare pooled 24-hour treated cells with same-cell-line DMSO using Wilcoxon without tie correction, BH adjustment over 2,931 genes, adjusted $p < 0.05$, and $|\log_2 \mathrm{FC}| > 0.25$.

Drug identity resolution precedes support-atlas construction. The predictor uses 54 support drugs and 19 validation drugs. The test set has 113 recorded identifiers corresponding to 112 canonical identities, yielding 339 identifier–cell conditions. Canonically equivalent identities are excluded together when a training query or fold is masked. The 2,000-gene evaluation includes 678,000 condition–gene values. The sampled classification masks contain 55,598 queries in this panel and 24,630 additional queries, with 8,651 and 3,777 DE positives, respectively. For each drug–cell condition, the sampled endpoint retains all DE positives and up to 200 negatives with unadjusted $p > 0.1$, sampled with seed 42. The 2,000-gene and additional-gene panels are scored separately.

The prediction target is the population response within each drug–cell condition. The shared representation combines drug knowledge and control-cell context with pooled-dose responses; the subsequent cell generator takes explicit dose inputs. All three cell lines are represented during training.

B PREDICT IMPLEMENTATION AND TRAINING

B.1 EVIDENCE COMPILATION AND ENCODING

The biological knowledge base includes records derived from ChEMBL (Zdrazil et al., 2024), Reactome (Ragueneau et al., 2026), OmniPath (Türei et al., 2021), and DoRothEA (Garcia-Alonso et al., 2019). Entity aliases connect the same compound or gene across sources.

Each condition is represented by drug descriptions, candidate targets, action or assay records, and cellular baseline features. The evidence compiler selects at most three targets and assigns each target cited evidence, relevance, and action scope. We use DeepSeek-V4-Flash through the `deepseek-flash` API with fixed language-model weights.

Qwen3-Embedding-0.6B (Zhang et al., 2025) supplies 1,024-dimensional frozen text embeddings. Text is divided into chunks of at most 768 tokens. The last valid token representation is normalized within each chunk; chunk representations are averaged with token-count weights and normalized again. The target-text aggregate weights target descriptions by relevance and control-cell detection. High, medium, low, and unknown relevance receive weights 1, 0.5, 0.1, and 0.25, multiplied by $0.25 + 0.75\sqrt{\rho}$, where ρ is the target's control detection rate; missing detection uses 0.5.

Structured features $\mathbf{s}_q$ are vectorized and normalized separately from the packet embedding $\mathbf{e}_q$ and weighted target-description embedding $\mathbf{t}_q$. After component normalization, the concatenation balances structured and semantic features:

$$\mathbf{x}_q = \left[\mathbf{s}_q/\sqrt{2}\,;\ \mathbf{e}_q/2\,;\ \mathbf{t}_q/2\right], \qquad \mathbf{z}_q = P_{\mathrm{tr}}(\mathbf{x}_q) \in \mathbb{R}^{96}. \tag{6}$$

Truncated SVD and feature normalization are fitted using the support set. Selected targets are available for 279 of the 339 test conditions; missing evidence is encoded explicitly, and all conditions are evaluated. Citation lists are normalized to the evidence budget while retaining their source identifiers.

B.2 ATLAS FRONTEND AND OUTPUT HEADS

The atlas value for a support condition concatenates normalized continuous response (2,000 entries), DE labels (2,931 entries), and the product of DE and up labels (2,931 entries). Its dimension is therefore 7,862. Linear query and key maps project knowledge inputs to 64 dimensions; a linear layer with GELU maps the observation values to the same width. The retrieval weights in Equation (2)

include three nonnegative learned coefficients for the chemical, target, and mechanism-of-action similarity channels.

A sigmoid gate combines retrieved context with the query before layer normalization. Each cell context has a 64–64–32 MLP producing the latent representation. Its continuous head is a 32–64–2,000 MLP with GELU and dropout 0.1. The final continuous layer is initialized to zero, so initialization corresponds to the training response mean. The response scale is a scalar learned from the support data through normalization, and the response mean is cell- and gene-specific.

Training minimizes MSE on normalized continuous responses with AdamW, learning rate 0.001, and gradient clipping at norm 5. The selected model uses hidden width 64, weight decay 0.01, and 25 epochs. We average predictions from three seeds. Five-fold drug masking restricts retrieval to observations outside the training query's fold. Model fitting and atlas construction use support responses, and validation selects model settings and checkpoints.

After continuous training, the encoder is frozen and regularized logistic heads are fitted on its representations. For support size n, the DE prior is $(n_{cg}^{DE} + 0.5)/(n + 1)$; the up prior is $(n_{cg}^{\mathrm{up}} + 0.5)/(n_{cg}^{DE} + 1)$. Their logits initialize the offsets. The objective combines DE binary cross-entropy over all support coordinates and direction binary cross-entropy over support DE coordinates, with an ℓ_2 penalty on coefficients and intercepts. Head fitting uses 200 epochs; the validation-selected regularization values are 10 for DE and 100 for DIR. Classification training supplies gene-specific decision functions without changing the continuous head.

B.3 Condition-level DE count decoding

The hard DE readout fits a ridge model to each validation condition's true DE fraction, separately for the core and additional panels. Its inputs are the mean and standard deviation of DE probabilities; their 90th, 95th, and 99th percentiles; the fraction at least 0.5; and cell-context indicators. Grouped five-fold validation over drugs selects the ridge coefficient from $\{0.1, 1, 10, 100\}$ by count-fraction error. At inference, the predicted fraction is bounded between one gene divided by panel size and 0.8. The rounded count determines how many highest-probability genes to select. Direction is called at up probability 0.5. All count-model parameters are fixed before test scoring.

B.4 Probability-guided single-cell generation

Panel (a) of Figure 2 expands the shared condition state using drug identity, cell line, dose, time, matched DMSO expression rows, and stable sample identifiers. Classification probabilities guide the expression distribution; population DE/DIR is then recomputed from the generated cells. All expansion operations are numerical and reuse the compiled condition evidence.

Mean-response guidance. For each cell context and gene, support responses fit down- and up-regulated mean-response experts m^- and m^+. We use the 54 support drugs and 5,627 core DE events, clip training logFC to $[-8, 8]$, and shrink each mean toward its cell-context/direction mean with eight pseudo-observations. Let p be the DE-head probability and u the up-head probability. The predicted-count threshold t is the midpoint between the smallest included and largest excluded probability, separately for the core and additional panels. Its relative decision weight and corrected core response are

$$w = \frac{p(1-t)}{p(1-t) + t(1-p)}, \tag{7}$$

$$r^* = r + \mathrm{clip}\big(\lambda w |2u - 1| \big[(1-u)m^- + um^+ - r\big], -b, b\big)\,. \tag{8}$$

Here w is a relative confidence weight for the predicted gene set. The selected mean strength is $\lambda = 1$, with the correction clipped to one logFC unit ($b = 1$). An empty gene selection uses $t = \max(0.5, \max p)$; an all-included panel uses half its minimum probability, with t bounded away from zero and one.

Detection guidance and cellular variation. Support-fitted down/up detection experts q^-, q^+ are cell- and gene-specific, pooling labeled support observations across doses with eight pseudo-

Table 6: **Classification guidance connects cell generation to population decisions.** All rows use the same 339 conditions and sampled queries from the 2,000-gene panel. The first row scores the shared condition heads; the other rows recompute DE/DIR from generated expression versus DMSO. F1 is Macro-F1.

Output	RMSE	Pearson	DE AUPRC	DE F1	DIR F1
Shared condition heads	4.4762	0.6943	0.8588	0.8077	0.8393
Control-conditioned expression	4.4770	0.6942	0.7982	0.6506	0.8203
Probability-guided expression	4.4755	0.6944	0.8067	0.8164	0.8472

observations of shrinkage. We mix them with the original detection target q_0:

$$q^* = (1 - \eta w)q_0 + \eta w\big[(1-u)q^- + uq^+\big], \qquad \eta = 1. \tag{9}$$

The generator ranks matched control cells to realize gene-specific nonzero quotas, applies control amplitude and state modulation, and projects each generated cell to log1p CP10K. Amplitude factors are 0.25 for head-selected DE genes and 1 otherwise, with state strength 0.1. Dose-relative mean patterns are reanchored to the corrected pooled core response. Auxiliary support-fitted dose predictions supply the additional 931 genes' means; detection guidance covers all 2,931 genes.

Selection and population readout. In this retrospective study, seven coupling settings are evaluated on 19 validation drugs, 57 pooled conditions, and 57,379 cells per setting. We select the highest expression-derived DE Macro-F1 subject to RMSE increasing by at most 1%, Pearson decreasing by at most 0.005, DIR Macro-F1 by at most 0.01, and DE AUPRC by at most 0.015 relative to the control-conditioned generator without probability guidance. The selected setting is then fixed for test generation.

Test generation covers 113 drug identifiers, three cell lines, four doses (10, 100, 1,000, and 10,000 nM), and 24 hours: 1,356 precise conditions and 341,202 cells, each with 2,931 genes. We pool generated cells across doses into 339 drug–cell populations and compare each against its same-cell-line DMSO population. DE uses the reference Wilcoxon test without tie correction, BH adjustment over 2,931 genes, adjusted $p < 0.05$, and $|\log_2 \mathrm{FC}| > 0.25$. DE ranking uses $-\log_{10} p$. DIR ranking and hard decisions use signed generated logFC. Thus, the scores in Table 1 come from expression testing, with classifier probabilities entering through generation.

Table 6 separates shared-head classification from expression-derived decisions. Probability guidance raises expression-derived DE/DIR Macro-F1 from 0.6506/0.8203 to 0.8164/0.8472, while logFC RMSE changes from 4.4770 to 4.4755. SciPlex3 supplies unpaired treated/control populations, so these classification endpoints describe population responses. Component ablations and Update evaluate the shared condition heads before this generation step.

C External predictor configurations

The BioLORD, chemCPA, CMonge, and PRnet experiments use SciPlex3 with a 56-drug training split and seed 42. CellCanon uses 54 support drugs after canonical-identity exclusions and averages three seeds. Squidiff, PerturbNet, and STATE use the 54-drug support split, 19 validation drugs, the fixed 2,000-gene panel, and seed 42.

The BioLORD, chemCPA, CMonge, and PRnet models provide continuous response estimates. A validation-fitted logistic adapter using absolute predicted log fold change supplies their DE scores. For hard DE decisions, we enumerate validation probability thresholds and select the threshold with the highest Macro-F1 on the sampled validation queries, keeping expression predictions and adapter coefficients fixed. Selected thresholds are 0.038410 for BioLORD, 0.028758 for chemCPA, 0.041701 for CMonge, and 0.047620 for PRnet. BioLORD and CMonge use fitted direction scores. For chemCPA and PRnet, validation Macro-F1 selects between the fitted direction adapter and the native signed-logFC readout; both select signed logFC, with positive values classified as up. This rule is fixed before test scoring. PRnet uses stochastic generation from its predicted expression distribution.

CMonge training and inference. The CMonge adaptation uses a two-layer 512-unit autoencoder with a 100-dimensional latent space. We train it for 100 complete epochs (71,600 updates), selecting the checkpoint by validation reconstruction MSE. A conditional transport map uses RDKit descriptors, log dose, and cell-line indicators and trains for 500,000 updates. Checkpoint selection every 10,000 updates uses mean-expression R^2 over 228 validation drug–cell–dose conditions, with negative condition scores floored at zero for this selection criterion. The selected map is the 10,000-update checkpoint. At inference, matched DMSO cells are encoded, transported, decoded, and clipped to nonnegative expression; aggregation produces the pooled condition response. The test output contains 341,202 cells by 2,000 genes. All reported test R^2 values retain negative scores. On the sampled endpoint, CMonge's DE precision/recall are 0.3432/0.7334 and accuracy is 0.7402; conditional DIR accuracy is 0.6700.

Squidiff training and inference. We use Squidiff's drug-structure branch (He et al., 2026), conditioning on matched control expression and a 1,024-bit radius-2 feature-based Morgan fingerprint multiplied by $\log_{10}(1 + \mathrm{dose}_{\mathrm{nM}})$. The backbone has width 2,048 and three MLP blocks, with 1,000 diffusion timesteps. We apply VP output preconditioning: for $x_t = a_t x_0 + s_t \epsilon$, the network output v_θ gives $\epsilon_\theta = s_t x_t + a_t v_\theta$, retaining the original noise-prediction MSE. Training uses 100,000 updates, batch size 128, learning rate 3×10^{-5}, gradient clipping at 1, and EMA 0.999. We select the eligible live or EMA checkpoint with minimum validation population-expression MSE and generate with 50-step deterministic DDIM.

PerturbNet training and inference. PerturbNet (Yu et al., 2025) uses the pretrained 196-dimensional ChemicalVAE representation, a 10-dimensional scVI cell latent, and a conditional invertible network with 20 coupling blocks and hidden width 1,024. Molecular features, cell-line indicators, and log dose condition the flow. We train scVI for 700 epochs and the flow for 100 epochs, both with batch size 128; validation ELBO and population-expression MSE select epochs 695 and 80, respectively. Decoded counts use paired panel and full-library sizes sampled from training cells matched by cell line and dose, then transform to log1p CP10K. Molecular canonicalization and equivalent SMILES enumeration supply the chemical encoder inputs.

STATE training and inference. We train the STATE transition model (Adduri et al., 2025) from scratch with its eight-layer, 768-hidden-unit, 12-head bidirectional transformer and energy distribution loss. Its continuous perturbation input concatenates a 2,048-bit radius-2 Morgan fingerprint with normalized log dose. Matched DMSO cells form the input; each condition-homogeneous set contains 128 cells. Training uses batches of 96 sets, Adam at 10^{-4}, and BF16 network computation with FP32 loss calculation. Early stopping ends training at epoch 540, and validation energy selects epoch 390.

For these three models, generated cells are pooled across doses within each drug–cell-line condition. DE uses the Wilcoxon test against matched DMSO with adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.25$. DE ranking uses $-\log_{10} p$; direction ranking and hard decisions use signed generated logFC. Their continuous metrics use the generated population means.

For VCWorld, DeepSeek-V4-Flash receives condition evidence and returns separate DE and direction judgments with chain-of-thought rationales for each gene query. DE and up-regulation class indicators supply binary scores for AUPRC and AUROC, with tied scores retained. The same judgments supply the hard-decision metrics. CellCanon's classification heads are trained on DE/DIR labels; the expression baselines use the response-to-label readouts described above.

Random reference. Random independently samples a Bernoulli(0.5) label for each query, using the same binary value as its ranking score and hard decision. DE uses all 55,598 sampled test queries; DIR uses the 8,651 reference-DE queries, independently of the random DE decisions. For each of 100 seeds (0–99), separate task-specific random streams draw uniform values and threshold them at 0.5. We compute each metric on the complete query set within each seed and report its mean across all seeds. Tied binary scores are retained. Table 7 gives the standard deviations across random draws. For joint evaluation, the DE decision first selects unchanged versus changed, and the independent direction decision splits changed predictions into up and down. The resulting class probabilities are 0.5, 0.25, and 0.25, respectively. Joint Macro-F1 is 0.2881 ± 0.0018 and accuracy

is 0.4607 ± 0.0019, where the standard deviations describe variation over the 100 random seeds. This reference produces classification outputs only.

Table 7: **Random binary decisions.** Mean $\pm$ standard deviation over 100 random seeds on the same test queries. These deviations describe variation across random draws.

Task	AUPRC	AUROC	Macro-F1
DE	0.3666 ± 0.0032	0.4998 ± 0.0030	0.4324 ± 0.0018
DIR	0.6558 ± 0.0040	0.4991 ± 0.0052	0.4982 ± 0.0052

D METRIC DEFINITIONS AND ADDITIONAL PREDICT RESULTS

Let μ_g and $\hat{\mu}_g$ denote observed and predicted mean expression, respectively, for gene g in a drug–cell condition. Mean expression is the pooled mean of log1p CP10K across cells. For the 50 genes with the largest absolute reference Wilcoxon statistics in each condition ($G = 50$), the Predict comparison reports the coefficient of determination,

$$R^2 = 1 - \frac{\sum_g (\mu_g - \hat{\mu}_g)^2}{\sum_g (\mu_g - \bar{\mu})^2}, \qquad \bar{\mu} = G^{-1} \sum_g \mu_g. \tag{10}$$

Delta Pearson is the Pearson correlation between $\Delta_g = \mu_g - \mu_g^{\mathrm{DMSO}}$ and $\hat{\Delta}_g = \hat{\mu}_g - \mu_g^{\mathrm{DMSO}}$. R^2 uses the top 50 genes; delta Pearson uses all 2,000 evaluation genes unless a top-ranked subset is specified. Both metrics are calculated within each condition and macro-averaged across evaluated test conditions. Delta Spearman applies the same aggregation to rank correlations; correlations for constant vectors are undefined.

For the logFC targets y_g across all $G = 2{,}000$ evaluation genes, we additionally report condition-macro RMSE, $\sqrt{G^{-1} \sum_g (y_g - \hat{y}_g)^2}$, and within-condition Pearson correlation. Update MSE pools squared logFC errors within each hidden-cell partition and averages the three partition scores; its RMSE takes the square root within each partition before averaging.

DE treats differential expression as the positive class. Precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, recall is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and Macro-F1 averages class F1 scores. Accuracy is the fraction of query labels predicted correctly. AP, AUPRC, and AUROC are defined as

$$\mathrm{AP} = \sum_k (R_k - R_{k-1}) P_k, \quad \mathrm{AUPRC} = \int_0^1 P(R)\, dR, \quad \mathrm{AUROC} = \int_0^1 \mathrm{TPR}(u)\, du, \tag{11}$$

where P and R denote precision and recall, and u denotes the false-positive rate. Ranking uses expression-test significance for final Predict DE, signed generated logFC for DIR, and head probabilities for component and Update experiments. Direction mAP averages the up- and down-class AP values on reference DE genes.

DIR metrics condition on the reference DE subset and classify up versus down. DIR AUPRC, precision, and recall treat up as positive. For joint prediction, a negative DE decision yields unchanged; a positive DE decision yields up or down according to the direction readout. We compute each class F1 from the pooled confusion matrix of sampled queries in the evaluated test conditions and average the three F1 scores. This evaluation uses the same DE thresholds and direction readouts as the separate tasks. Table 1 reports the joint scores and per-class F1 values.

Table 8: **Continuous-response error.** RMSE is computed on pooled logFC across 2,000 genes within each condition and averaged across evaluated test conditions. Lower is better; the corresponding logFC Pearson scores appear in Table 2.

Method	LogFC RMSE ↓
BioLORD	11.7617
chemCPA	9.1093
CMonge	15.1668
PRnet	15.6431
Squidiff	18.5063
PerturbNet	6.7180
STATE	10.2651
CellCanon	**4.4755**

D.1 Drug-level classification robustness

Panels a–c of Figure 4 use the frozen prediction calls and sampled queries of Table 1. Within each recorded drug identifier, we pool queries across the three cell contexts and compute Macro-F1 with fixed class sets: two classes for DE and conditional DIR, and three for joint unchanged/up/down. Undefined class F1 is set to zero. DIR uses only reference-DE queries; four canonical drugs have none and are excluded from DIR analysis. Equivalent identifiers are scored separately and their scores averaged, giving equal weight to each of 112 canonical drugs (108 for DIR). Gains subtract a fixed comparator's score from CellCanon's. The comparator is the highest aggregate Macro-F1 baseline in Table 1: PRnet for DE and joint, and BioLORD for DIR. It is not reselected per drug.

Response strength is the reference Core2000 DE-gene count averaged across cell contexts and then equivalent identifiers. Four quartile strata retain tied strengths together at cutoffs of 1.3333, 5.8333, and 29.25 DE genes per context, giving 31/25/28/28 drugs (27 in the first DIR stratum). Across all eligible drugs and within each stratum, 10,000 resamples of canonical drugs with replacement give percentile 95% intervals for the median paired gain, conditional on the saved predictors. These describe variation across drugs rather than retraining uncertainty. Violin densities show the observed range, with individual scores, horizontal median bars, and vertical interquartile ranges; wins and losses require gains above 10^{-12} or below -10^{-12}, with intermediate differences treated as ties. Drug-level summaries weight drugs equally and therefore differ from the pooled-query main-table metric.

D.2 Differential-gene and expression-delta evaluation

Delta correlations additionally cover all 2,000 genes and the top 20 and top 100 genes. Gene sets are selected for evaluation, with ties resolved by gene order. The top-k sets include the highest-ranked genes regardless of their significance threshold.

CellCanon, Squidiff, PerturbNet, and STATE mean expression is calculated from generated cells. BioLORD, chemCPA, CMonge, and PRnet means are recovered from predicted logFC vectors and the common DMSO reference using the inverse target transformation. We apply the Top50 R^2 and gene-set-specific delta-correlation definitions above, using the same pooled-dose populations and reference labels. Table 2 reports the Top50 expression comparison.

Top50 expression R^2 summarizes reconstruction on response-ranked genes: CellCanon reaches 0.9145, compared with 0.9197 for BioLORD and 0.8880 for the DMSO reference. Across all 2,000 genes, CellCanon's delta Pearson/Spearman are 0.3254/0.2412. On the top-100 panel, it reaches 0.3971/0.3809, compared with 0.3888/0.3632 for chemCPA (Table 9).

Table 9: **Delta response correlations relative to DMSO.** Delta subtracts the same-cell control mean from both predicted and reference mean-log1p expression. Pearson and Spearman are computed across genes within each condition and averaged across evaluated test conditions. Both correlations are undefined for the constant zero-delta DMSO reference. Higher is better.

Method	All-gene Δ-Pearson	All-gene Δ-Spearman	Top100 Δ-Pearson	Top100 Δ-Spearman
BioLORD	0.2538	0.1298	0.3618	0.3199
chemCPA	<u>0.2543</u>	0.1534	<u>0.3888</u>	<u>0.3632</u>
CMonge	0.1484	0.0900	0.1916	0.1822
PRnet	0.1528	0.0859	0.2357	0.2089
Squidiff	0.0703	0.0328	0.2719	0.2396
PerturbNet	0.1500	0.1085	0.2405	0.2062
STATE	0.2401	<u>0.1814</u>	0.3120	0.2809
DMSO reference	–	–	–	–
CellCanon	**0.3254**	**0.2412**	**0.3971**	**0.3809**

E PREDICT COMPONENT ABLATIONS

E.1 INTERVENTIONS AND COMMON TRAINING PROTOCOL

These ablations score the shared condition predictor before probability-guided generation. All variants use the 54-support/19-validation/113-test identifier split and the same 2,000-gene evaluation. Input interventions retrain the affected frontend and independent heads, select settings on validation drugs, and average three seeds. Readout-only interventions keep continuous predictions fixed and fit their decision rules on validation drugs. Validation selects hidden width, weight decay, and DE/DIR regularization using the same search protocol across variants. Knowledge projection and normalization use the 162 support conditions. The DE count decoder is refitted on the 57 validation conditions.

Biological annotations. The knowledge-removed variant sets all 96-dimensional query/key knowledge inputs to zero throughout training and removes target and mechanism-of-action priors. Removed information includes text embeddings, target/pathway/action annotations, LLM judgments, and target-conditioned DMSO features. Chemical similarity, support-response values, ordinary training statistics, and cell-specific modules remain. This intervention measures the contribution of the combined biological-annotation channel.

Text embeddings. The text-removed variant retains LLM target selection, relevance, scope, and structured fields, but removes evidence-text and target-description embeddings. Its structured vectorization and 96-dimensional projection are refitted on training conditions.

Measured cell context. Measured-context removal deletes target control-expression and detection fields before compiling evidence for every condition, excludes those fields and their missingness indicators from text and structured features, and removes detection-based target weights. The response atlas is retained. Query features and atlas knowledge keys use the context-free encoding; training response values remain available. Drug/gene annotations, cell identity, and ordinary training means and scales remain. Context-free compilation introduces generation variation alongside context removal.

Biological-annotation deletion produces the largest classification losses among these input interventions. Table 11 reports their ranking and hard-decision metrics.

Table 10: **Predict input-component comparison.** The common 54/19/113 drug-identifier split, 2,000-gene evaluation panel, and metric definitions match Section E. Text removal retains LLM-selected structured fields. Context removal retains the response atlas. Ordinary training statistics and cell identity remain available.

Variant	RMSE ↓	DE AP	DE F1	DIR mAP	DIR F1	Joint F1
Condition heads	4.4762	0.8588	0.8077	0.9135	0.8393	0.7242
Without biological knowledge	4.4699	0.7922	0.7117	0.8593	0.7900	0.5776
Without text embeddings	4.4688	0.8640	0.8066	0.9102	0.8370	0.7228
Without measured context	4.4781	0.8611	0.8118	0.9093	0.8360	0.7281

Table 11: **Classification operating points and ranking metrics for the ablations.** DE uses 55,598 queries with 8,651 positives. DIR uses those 8,651 true-DE queries with 4,687 up-regulated positives. Its AP, PR area, precision, and recall treat up as positive; this AP is distinct from the two-direction mAP in Section E. Hard DE decisions use the validation-fitted count decoder.

Variant	Task	AP	PR area	AUROC	F1	Accuracy	Precision	Recall
Condition heads	DE	0.8588	0.8588	0.9676	0.8077	0.9163	0.8876	0.5294
	DIR	0.9292	0.9291	0.9156	0.8393	0.8398	0.8737	0.8233
Without biological knowledge	DE	0.7922	0.7921	0.9546	0.7117	0.8875	0.8396	0.3425
	DIR	0.8905	0.8904	0.8695	0.7900	0.7900	0.8616	0.7295
Without text embeddings	DE	0.8640	0.8640	0.9676	0.8066	0.9162	0.8920	0.5251
	DIR	0.9278	0.9278	0.9131	0.8370	0.8375	0.8714	0.8212
Without measured context	DE	0.8611	0.8610	0.9680	0.8118	0.9176	0.8861	0.5396
	DIR	0.9263	0.9262	0.9124	0.8360	0.8364	0.8696	0.8212

E.2 Measured cell context and response changes

With the response atlas retained, removing measured background reduces delta Pearson and Spearman on all 2,000 genes and the top-20/50/100 reference panels (Table 12). All-gene Pearson decreases by 0.0265 and Spearman by 0.0091; the top-100 decreases are 0.0241 and 0.0293. All eight paired difference intervals exclude zero. Direction mAP also decreases from 0.9135 to 0.9093. The classification operating points in Table 11 show that the benefit is endpoint-specific: DE F1 is 0.8118 without measured context versus 0.8077 for the complete model.

E.3 Independent classification readouts

We keep the complete model's continuous predictions fixed and replace the DE/DIR heads with bounded scores derived from the predicted log fold change y: $s^{DE} = |y|/(1 + |y|)$ and $s^{\mathrm{up}} = 1/2 + y/[2(1 + |y|)]$. These monotone transforms encode response magnitude and sign as ranking scores. The comparison covers the 2,000 genes with continuous outputs; the training atlas still contains its original DE/DIR observations.

The first readout variant refits the DE count decoder on validation predictions. A second variant replaces that decoder with fixed DE thresholds selected separately for each cell line to maximize sampled-validation Macro-F1. Both use validation-selected direction thresholds on true-DE queries, retaining score ties. All rows of Table 13 share identical continuous responses.

Independent heads achieve DE AUPRC 0.8588 versus 0.1429 for magnitude-based scores. Their DE Macro-F1 is 0.8077, compared with 0.4538 using the refitted count decoder and 0.5278 using fixed thresholds. Joint three-state Macro-F1 is 0.7242 versus 0.3029 and 0.4361, respectively. Direction remains more recoverable from the signed response, with Macro-F1 0.8134 versus 0.8393 for the independent direction head. These comparisons support task-specific classification readouts over direct magnitude/sign decisions from the same mean-response predictor.

Table 12: **Measured cell context improves perturbation-response correlations.** The context-removed model retains the response atlas and cell-specific modules, while excluding measured control-expression and detection information from evidence compilation and encoding. Correlations use the same 339 conditions and reference gene panels as Section D.2. Differences are context-removed minus complete model; brackets give exploratory 95% intervals from 1,000 paired canonical-drug bootstrap draws.

Gene panel	Metric	Condition heads	No context	Difference [95% interval]
All 2,000	Δ-Pearson	0.3074	0.2810	-0.0265 [-0.0410, -0.0130]
All 2,000	Δ-Spearman	0.2383	0.2292	-0.0091 [-0.0144, -0.0039]
Top20	Δ-Pearson	0.4264	0.4030	-0.0235 [-0.0423, -0.0062]
Top20	Δ-Spearman	0.3894	0.3628	-0.0265 [-0.0471, -0.0081]
Top50	Δ-Pearson	0.3961	0.3733	-0.0228 [-0.0370, -0.0093]
Top50	Δ-Spearman	0.3700	0.3420	-0.0280 [-0.0433, -0.0133]
Top100	Δ-Pearson	0.3694	0.3453	-0.0241 [-0.0374, -0.0113]
Top100	Δ-Spearman	0.3510	0.3216	-0.0293 [-0.0440, -0.0150]

Table 13: **Independent task readouts improve classification from the same continuous predictor.** All rows share frozen 2,000-gene continuous outputs (RMSE 4.4762). DE AUPRC and DE/joint F1 use 55,598 sampled queries. DIR F1 uses 8,651 true-DE queries, and DIR mAP averages up/down AP on true-DE entries. All F1 scores are macro-averages. The two continuous-score rows use validation-refitted count decoding or validation-selected fixed DE thresholds, respectively.

Readout and DE decision rule	DE AUPRC	DE F1	DIR mAP	DIR F1	Joint F1
Independent heads + count decoder	0.8588	0.8077	0.9135	0.8393	0.7242
Continuous scores + count decoder	0.1429	0.4538	0.8901	0.8134	0.3029
Continuous scores + fixed thresholds	0.1429	0.5278	0.8901	0.8134	0.4361

E.4 Paired uncertainty

We use 1,000 paired bootstrap draws over 112 canonical drug identities, keeping aliases and all associated cells and genes in the same block. Table 14 reports variant-minus-reference differences. The five principal classification intervals for annotation removal exclude zero. Context removal with the atlas retained reduces direction mAP, and replacing the independent readouts reduces DE AP, DE F1, and DIR F1; their reported intervals also exclude zero. Delta-context intervals appear alongside all four reference panels in Table 12. Text embeddings give a smaller direction-ranking effect. These exploratory intervals are conditional on the selected ensembles and use no multiple-comparison adjustment.

Table 14: **Paired differences for the reported component effects.** Differences are variant minus complete model. Percentile 95% intervals use 1,000 paired resamples of 112 canonical drug blocks, retaining all associated cells and genes. Intervals condition on the selected ensembles and are unadjusted for multiple comparisons; they describe sampling variation rather than retraining variation.

Comparison	Metric	Difference	Lower	Upper
Without biological knowledge	DE AP	-0.066687	-0.102294	-0.011355
Without biological knowledge	DE F1	-0.096053	-0.152370	-0.019405
Without biological knowledge	DIR mAP	-0.054199	-0.067440	-0.036245
Without biological knowledge	DIR F1	-0.049380	-0.068552	-0.030069
Without biological knowledge	Joint F1	-0.146592	-0.214128	-0.053563
Without text embeddings	DIR mAP	-0.003370	-0.006790	-0.000359
Without measured context	DIR mAP	-0.004185	-0.007109	-0.001334
Continuous scores + count	DE AP	-0.715933	-0.746663	-0.656455
Continuous scores + count	DE F1	-0.353891	-0.401736	-0.287063
Continuous scores + count	DIR F1	-0.025929	-0.042735	-0.012632
Continuous scores + threshold	DE F1	-0.279905	-0.317796	-0.243067

E.5 Evidence-view sensitivity

The raw-source knowledge view gives DE AP 0.8661 and direction mAP 0.9065; LLM compilation gives 0.8588 and 0.9135, respectively. Compilation therefore improves direction ranking by 0.70 percentage points while trading off DE ranking and a small amount of continuous error. The two knowledge views differ in source volume.

Table 15: **Raw-source and LLM-compiled knowledge views.** Both use Qwen text encoding and the same numerical architecture. Raw-source encoding omits LLM selection and changes information volume.

Variant	RMSE ↓	DE AP	DE F1	DIR mAP	DIR F1	Joint F1
Raw-source knowledge	4.4723	0.8661	0.8136	0.9065	0.8319	0.7339
LLM-compiled knowledge	4.4762	0.8588	0.8077	0.9135	0.8393	0.7242

F Source-grounded explanation benchmark

F.1 Reference construction and scoring

The evaluation cohort contains 24 drugs, 72 conditions, and 966 questions; the validation cohort contains six drugs and 225 questions. Each question is paired with its condition's evidence, a supported/contradicted/insufficient label, and acceptable proof bundles constructed from source rules. References define 561 answerable and 405 insufficient-evidence questions. For an answerable question, joint correctness requires the correct label, at least one complete acceptable proof, and citations drawn from the supplied evidence. Other responses receive zero. Success on an insufficient-evidence question requires a valid output with an explicit insufficient-evidence judgment.

The answerable questions comprise 240 control-detection questions, 138 target-action scope questions, 126 background-path questions, and 57 source-statement questions. Across the 966 questions, there are 487 distinct question strings and 556 question-type/argument combinations; 126 questions share a string with the validation set. Question templates are reused with condition-specific evidence. The source rules also provide a deterministic comparator (Table 16). The benchmark evaluates structured judgments and cited proofs, and the agent supplies accompanying natural-language rationales.

The explanation procedure uses the DeepSeek Flash API (`deepseek-flash`) for source analysis, question answering, and localized review. Review feedback comprises source-scope, path, and formatting diagnostics. Error analysis on the evaluation panel informed refinement of the correction procedure. Reported scores evaluate the final pipeline; reference answers were not supplied to runtime correction.

F.2 Paths and uncertain signs

A route is eligible when its endpoints, source records, and sign relationships satisfy the path criteria. Known edge signs are composed directly. Unknown signs are assigned jointly across paths so that a shared edge has one value within each completion. These alternative assignments distinguish conclusions that hold for every permitted sign from those that depend on an unresolved interaction.

The language model selects and ranks candidate paths. If eligibility review leads to a correction, the included paths are ordered by length and source identifier; otherwise the model's ordering is retained. Among the 72 evaluation conditions, 63 contain eligible reference paths. On these 63 conditions, valid path@3, precision@3, and recall@3 are each 1.000. Proposition scoring includes all 72 conditions, including the nine with no eligible reference path.

F.3 Observation-state reasoning and withdrawal

The observation-state task comprises 20 measurement–target records spanning 12 drugs and 13 unique conditions. Each record is presented before measurement, after measurement, and after withdrawal, producing 480 questions: 120 answerable and 360 insufficient. Gene identities and numerical predicates are matched to their source records, and the language model supplies a judgment, citation, and explanation. Withdrawal removes a record from the active evidence while retaining its history. The agent correctly answers all 120 answerable questions with complete proof and makes the correct insufficient-evidence judgment on all 360 insufficient questions.

An additional source-withdrawal challenge makes 16 previously answerable claims insufficient. The agent revises 14 claims correctly (87.5%). Both withdrawal tasks use source-defined reference judg-

Table 16: **Source-defined reference comparisons.** Both label accuracy and label-plus-proof accuracy use all 561 answerable questions. Insufficiency uses all 405 questions. The symbolic reference applies the source rules directly.

Method	Label accuracy	Label + proof	Correct insufficiency
TF–IDF retrieval + rules	0.5989	0.5989	405 / 405
Complete-source symbolic reference	1.0000	1.0000	405 / 405
CellCanon	1.0000	1.0000	405 / 405

ments. These explanation experiments and the numerical transfer experiment in Section 3.4 are evaluated separately.

F.4 TSA EXPLANATION AND UPDATE CASE

Figure 5 presents Trichostatin A (TSA) in two experiments. The Explain experiment uses K562 at 24 hours with pooled doses. Its evidence includes functional HDAC1 inhibition, a binding record with an unspecified action sign, and an HDAC1 RNA detection fraction of 0.1530 in vehicle controls. Two background routes connect TSA to CYP24A1 through HDAC1: TSA–HDAC1–TP53–CYP24A1 has net sign $+1$, and TSA–HDAC1–AR–ESR1–CYP24A1 has net sign -1, including the inhibitory drug–target edge. These source-linked routes provide candidate mechanisms for the condition.

Localized review changes two reverse-sign propositions from contradicted to insufficient evidence because the binding record permits alternative sign assignments. The explanation retains the supporting citations and biological rationale.

The Update example measures TSA in A549 and predicts its response in unobserved K562. The source supplies a 2,000-gene continuous response and 2,931-gene DE/DIR labels. After 12 observations, the predicted K562 DE set changes from 153 to 146 of 2,931 genes. Aggregate prediction accuracy is evaluated in Section H.

G UPDATE FITTING AND OBSERVATION PROTOCOL

G.1 TRAINING AND CROSS-CELL OBSERVATIONS

Five-fold cross-fitting over the 54 support drugs produces parent residuals for updater training. Each outer fold refits the 96-dimensional knowledge projection, normalization, atlas representation, continuous MLP, and classification heads without that fold's drug identities. Each MLP also uses internal whole-fold atlas masking. Cross-fitted predictions use three random seeds. The parent architecture and its 25-epoch schedule remain fixed. Nineteen validation drugs select updater settings before test evaluation.

Each of three test partitions assigns one hidden cell line to each test drug identity; the assignment can differ across drugs. Observations come from the other two cell lines, with target conditions and their canonical aliases excluded from acquisition. Each measured source condition supplies a 2,000-gene response vector and 2,931-gene DE/DIR labels. Budgets count distinct source conditions: 112 observations cover one source cell line for each canonical drug, and 224 cover both. Each source condition can be selected once, up to the observation budget.

The Update experiment scores the shared condition predictor's continuous and classification heads before cell generation. The three target-cell partitions contain 19,311, 18,336, and 18,711 sampled DE queries, respectively, on the 2,000-gene panel. Budgets of 12, 24, and 48 follow fixed numerical acquisition schedules; budgets of 112 and 224 supply one or two source contexts per identity. At every budget, updated and initial predictions are paired on the same hidden targets. Acquisition schedules are held constant to measure the contribution of prediction updating under cross-cell observations.

G.2 CONTINUOUS TRANSFER EQUATIONS

Let $\mathbf{r}_i$ denote the standardized source-cell response of training drug i, with concatenation when two source cells are available. A product kernel combines similarity of measured responses and drug knowledge:

$$K(i,j) = k_{\mathrm{RNA}}(\mathbf{r}_i, \mathbf{r}_j)\left[(1-\eta) + \eta k_{\mathrm{drug}}(i,j)\right]. \tag{12}$$

Here k_{RNA} is linear and k_{drug} combines chemical and biological similarities. Kernel ridge regression learns the mapping to the target-cell response matrix Y_c. With centered kernel K_0, centered query kernel $\mathbf{k}_{d,0}$, and target mean $\bar{\mathbf{y}}_c$, the transferred response is

$$\widehat{\mathbf{y}}_{dc}^{\mathrm{transfer}} = \bar{\mathbf{y}}_c + \mathbf{k}_{d,0}^{\top}(K_0 + \lambda I)^{-1}\left(Y_c - \mathbf{1}\bar{\mathbf{y}}_c^{\top}\right). \tag{13}$$

The updater uses the transferred response when a source measurement is available and retains the parent prediction otherwise.

Table 17: **Prediction after observation feedback.** Each entry averages three partitions with 113 hidden drug–cell targets per partition. The 12/24/48 schedules select a subset of source conditions; 112/224 reveal one/two other cell contexts for each of the 112 canonical drugs. F1 denotes Macro-F1.

Observation budget	MSE ↓	DE AP	DE AUROC	DE F1	DIR F1
No update	20.4360	0.8806	0.9710	0.8199	0.8402
12 observations	20.3889	0.8833	0.9715	0.8210	0.8419
24 observations	20.2662	0.8893	0.9723	0.8242	0.8410
48 observations	20.1772	0.8924	0.9726	0.8393	0.8425
112 observations	20.0427	0.9077	0.9792	0.8261	0.8453
224 observations	20.0447	0.9030	0.9777	0.8583	0.8445

Table 18: **Update classification metrics: 2,000-gene evaluation panel.** Sampled endpoint, averaged over three hidden-cell partitions. DIR AP, precision, and recall use up as the positive class. The zero row is the same model without Update.

Budget	Task	AP	AUROC	F1	Accuracy	Precision	Recall
0	DE	0.8806	0.9710	0.8199	0.9162	0.9013	0.5570
0	DIR	0.9323	0.9179	0.8402	0.8408	0.8845	0.8167
48	DE	0.8924	0.9726	0.8393	0.9230	0.8975	0.6064
48	DIR	0.9305	0.9157	0.8425	0.8432	0.8838	0.8224
112	DE	0.9077	0.9792	0.8261	0.9185	0.9021	0.5715
112	DIR	0.9313	0.9167	0.8453	0.8460	0.8840	0.8283
224	DE	0.9030	0.9777	0.8583	0.9305	0.8985	0.6553
224	DIR	0.9315	0.9168	0.8445	0.8451	0.8853	0.8248

G.3 Selected numerical update settings

Continuous transfer in Equation (12) uses $\eta = 0.5$, ridge coefficient $\lambda = 1$, and transfer strength 1. The transfer map is fitted on source and target responses from training drugs. At inference, the measured response supplies the query vector for updating the same drug in its hidden cell line.

For DE, the conditional-label likelihood updater uses smoothing 200 and strength 0.5. It is mixed equally with a cross-fitted residual update of strength 0.25. A response-feature tree with at most seven leaves contributes an additional correction of strength 0.1. For DIR, the likelihood contribution has selected strength zero; the residual contribution has strength 1 and mixture weight 0.25, with no additional response-feature tree correction. These settings are fixed across the reported budgets.

The response-count model standardizes support features and fits a ridge regressor with coefficient 10. Its source-activity count estimate contributes weight 0.25, and 1.25 times the parent's selected DE count contributes weight 0.75. The updated probabilities determine the ranked gene set. Consequently, Update includes both response-conditioned probability changes and readout recalibration. The no-observation path returns the parent's original outputs, including its hard gene selections.

The parent MLP, response-transfer map, and classification corrections remain fixed during test-time updating. Measured source responses enter as inference inputs, and the updated values are predictions for the hidden target cell line.

H Additional Update results

Table 17 summarizes continuous and classification performance; Table 18 reports sampled DE and conditional direction on the 2,000-gene evaluation panel. Metrics are computed within each of the three partitions and then averaged. At 224 observations, DE recall increases by 9.84 percentage points with precision 0.8985. DIR Macro-F1 improves, while up-positive AP and AUROC decrease slightly. DE AP is highest at 112 observations and DE Macro-F1 at 224, showing different effects on ranking and hard decisions.

The mean partition RMSE is 4.5205 without Update, 4.4918 at 48 observations, and 4.4770 at 224.

H.1 ADAPTATION TO NEWLY OBSERVED RESPONSES

This task asks whether observing a drug response in one cell context improves prediction in an unobserved context. Each partition assigns a hidden cell context to each drug, and measurements come from its other contexts. A measurement reveals the source population at 24 hours, including 10, 100, 1,000, and 10,000 nM doses. We evaluate no observations, 48 observed conditions, and one source context per drug. Source selections are fixed before adaptation; target responses are used for evaluation. Scores average the three partitions, pairing each method's adapted and unadapted predictions on the same hidden targets.

Parameter adaptation. Each comparator starts from its selected trained weights and a new optimizer independently at each observation budget. Gradient updates use exclusively the newly observed treated populations, with matched DMSO controls. We select the learning rate from $\{10^{-5}, 3 \times 10^{-5}\}$ and the number of updates from $\{100, 400\}$ using 19 validation drugs. Selection minimizes mean hidden-condition expression MSE over three validation partitions and observation budgets of 8 and 19. The selected settings are fixed for test adaptation (Table 19).

STATE updates its transition network with the energy distribution loss; PRnet updates its prediction network with Gaussian negative log-likelihood. chemCPA retains alternating reconstruction, dose-response, and adversarial optimization. PerturbNet updates its conditional flow with negative log-likelihood while holding scVI and chemical representations fixed. CMonge updates the conditional transport map with transport fitting and the Monge-gap penalty while keeping its autoencoder fixed. BioLORD updates the shared decoder and attribute networks together with nuisance codes for newly observed cells; existing cell codes remain fixed. STATE uses eight sets of 128 cells per update. CMonge uses sets of 128 cells from a single condition; the other models use mixed-condition batches of 128 cells. Optimization uses seed 42; generation uses condition- and dose-specific random streams derived from seed 42.

Table 19: **Validation-selected fine-tuning settings.** Each setting minimizes hidden-condition expression MSE across the validation observation budgets and partitions.

Method	Learning rate	Updates
STATE	10^{-5}	400
PRnet	3×10^{-5}	400
chemCPA	3×10^{-5}	400
PerturbNet	10^{-5}	400
CMonge	3×10^{-5}	400
BioLORD	10^{-5}	100

Prediction and readout. Fine-tuned comparators generate the prescribed number of cells at each target dose in the 2,000-gene panel. Adapted and unadapted predictions use matched dose-specific cell counts and random draws. Responses are pooled across doses and tested against matched DMSO by Wilcoxon rank-sum, with Benjamini–Hochberg correction over the 2,000 predicted genes. DE calls require $q < 0.05$ and absolute $\log_2$ fold change above 0.25; ranking uses $-\log_{10} p$. Direction uses signed generated logFC on reference-DE queries. PRnet Gaussian samples are clipped at zero per cell before pooling and testing. These population tests provide the comparator scores in this experiment, including their zero-observation reference.

CellCanon retains its condition-head readouts throughout observation feedback, with a 2,000-gene continuous source response and 2,931-gene source DE/DIR labels. Fine-tuning uses the observed single-cell expression in the 2,000-gene panel. All reported classification scores use the sampled 2,000-gene endpoint. Joint scores combine DE and direction into unchanged/up/down as in Section D; each metric is computed within a partition and then averaged across the three partitions. Changes are measured within each method's prediction pipeline (Table 20). Table 21 reports classification before and after observation; Table 22 reports generated-response metrics for the six fine-tuned models.

With one source context per drug, STATE improves DE ranking, DE and DIR decisions, and response-change Pearson correlation. PerturbNet improves classification while response-change

Table 20: **Adaptation with one source condition per drug.** Entries are absolute score changes after adaptation minus each method's own unadapted baseline, averaged over three partitions on the 2,000-gene panel. F1 denotes Macro-F1; Acc. denotes accuracy. Protocol and absolute scores appear in Section H.1.

Method	Δ DE AUPRC	Δ DE F1	Δ DIR F1	Δ DE Acc.
Fine-tuning: generated-expression readouts				
STATE	+0.0872	+0.0808	+0.0716	+0.0297
PRnet	+0.0001	+0.0065	+0.0552	+0.0076
chemCPA	-0.0014	+0.0865	-0.0785	-0.0346
PerturbNet	+0.1023	+0.0839	+0.0557	+0.0229
CMonge	-0.0031	-0.0064	-0.0042	-0.0056
BioLORD	-0.0011	-0.0912	-0.1143	-0.1306
Fixed updater: condition heads				
CellCanon	+0.0271	+0.0063	+0.0051	+0.0023

Table 21: **Classification after feedback adaptation.** Scores are averaged over three source–target partitions on the 2,000-gene panel. None denotes the unadapted model; the other rows use 48 source conditions or one source condition per drug. DE uses sampled gene queries, and DIR uses reference-DE genes. F1 denotes Macro-F1 and Acc. denotes accuracy. Higher is better. Methods are ranked within each observation budget.

Method	Sources	DE AUPRC	DE F1	DE Acc.	DIR F1	DIR Acc.
Fine-tuning: generated-expression readouts						
STATE	None	0.4663	0.6195	0.8259	0.7606	0.7613
	48	0.5314	0.6702	0.8434	0.7912	0.7930
	One source/drug	0.5535	0.7003	0.8555	0.8323	0.8329
PRnet	None	0.0947	0.1176	0.1262	0.5936	0.6566
	48	0.0945	0.1216	0.1310	0.5441	0.6282
	One source/drug	0.0948	0.1241	0.1338	0.6488	0.6785
chemCPA	None	0.3202	0.4921	0.7545	0.6446	0.6449
	48	0.3286	0.5335	0.7409	0.5870	0.5926
	One source/drug	0.3188	0.5786	0.7199	0.5661	0.5892
PerturbNet	None	0.5256	0.6791	0.8407	0.6331	0.6371
	48	0.6463	0.7634	0.8744	0.6899	0.6930
	One source/drug	0.6279	0.7630	0.8637	0.6889	0.6937
CMonge	None	0.1337	0.2030	0.2031	0.6649	0.6700
	48	0.1322	0.1995	0.1999	0.6743	0.6809
	One source/drug	0.1306	0.1965	0.1976	0.6608	0.6693
BioLORD	None	0.1348	0.3539	0.3986	0.7834	0.7860
	48	0.1285	0.2920	0.3142	0.7365	0.7415
	One source/drug	0.1337	0.2627	0.2680	0.6691	0.6709
Fixed updater: condition heads						
CellCanon	None	**0.8806**	**0.8199**	**0.9162**	**0.8402**	**0.8408**
	48	**0.8924**	**0.8393**	**0.9230**	**0.8425**	**0.8432**
	One source/drug	**0.9077**	**0.8261**	**0.9185**	**0.8453**	**0.8460**

Table 22: **Continuous prediction after fine-tuning.** Scores are averaged over three source–target partitions. Top50 R^2 measures mean expression; logFC Pearson and Δ-Pearson use all 2,000 genes. Δ denotes expression changes from matched DMSO, and Top100 Δ-Spearman uses the reference Top100 gene panel. Source budgets follow Table 21. Higher is better. Methods are ranked within each observation budget.

Method	Sources	Top50 R^2	logFC Pearson	All 2,000 Δ-Pearson	Top100 Δ-Spearman
STATE	None	0.8998	0.3045	0.2417	0.2800
	48	**0.9146**	0.3144	**0.2758**	**0.3115**
	One source/drug	**0.9230**	0.3214	**0.3017**	**0.3384**
PRnet	None	0.8475	0.2386	0.0977	0.2149
	48	0.8535	0.2394	0.0941	0.2045
	One source/drug	0.1368	0.2411	0.0724	0.1183
chemCPA	None	0.8948	**0.4315**	**0.2436**	**0.3547**
	48	0.8668	0.4330	0.1675	0.2568
	One source/drug	-1.6496	0.3859	0.0843	0.1605
PerturbNet	None	0.4234	0.4255	0.1531	0.2162
	48	0.4883	**0.4377**	0.1503	0.2106
	One source/drug	0.2069	**0.4185**	0.1332	0.1948
CMonge	None	0.7119	0.2406	0.1383	0.1685
	48	0.7828	0.2407	0.1351	0.1660
	One source/drug	0.8401	0.2408	0.1331	0.1646
BioLORD	None	**0.9145**	0.2946	0.2428	0.3106
	48	0.8952	0.2999	0.1858	0.2558
	One source/drug	0.6793	0.2631	0.0580	0.0802

Pearson decreases from 0.1531 to 0.1332. CMonge and BioLORD decline in the four classification metrics summarized in Table 20. The 48-observation results further show that increasing the amount of feedback need not improve every endpoint.

I LANGUAGE-MODEL TOKEN USE

Table 23: **Token use on matched condition cohorts.** VCWorld's gene-wise prediction and chain-of-thought requests supply a shared reference; CellCanon Predict and Explain are compared separately. Counts sum recorded input and output tokens from API calls, including corrections. Savings are relative to the corresponding VCWorld cohort.

Method / stage	Input	Output	Total	Saved
Predict: 339 conditions; 55,598 reference gene queries				
VCWorld	310,614,134	109,667,313	420,281,447	–
CellCanon Predict	882,462	99,131	981,593	99.77%
Explain: 72 conditions; 12,399 reference gene queries				
VCWorld	69,060,037	24,334,813	93,394,850	–
CellCanon Explain	1,347,957	141,533	1,489,490	98.41%

We match conditions by drug identifier and cell line. The Predict comparison covers all 339 test conditions and 55,598 sampled gene queries from the 2,000-gene evaluation panel. The Explain comparison uses the 72 evaluated conditions and their 12,399 corresponding reference gene queries. VCWorld's DE and direction requests include their chain-of-thought rationales; their combined token use supplies the reference for each separate CellCanon stage. Token savings are $100(1 - T_{\text{stage}}/T_{\text{reference}})$ percent.

Token totals include input and output tokens from each API call once, including successful correction calls. Calls without reported usage are excluded. Tables 23 and 24 reports the matched-cohort comparison and stage totals.

Table 24: **Language-model usage.** Token counts include successful correction requests. Input and output are totals; the last column gives total tokens per condition.

Stage	Conditions	Input	Output	Total	Per condition
Predict: test	339	882,462	99,131	981,593	2,896
Predict: all splits	558	1,403,786	155,684	1,559,470	2,795
Explain: evaluated cohort	72	1,347,957	141,533	1,489,490	20,687

Explain's 1,489,490 tokens cover source analysis, localized review, and path qualification for the 72-condition cohort. Observation-state experiments are evaluated separately. Numerical Update reuses compiled evidence and requires no additional language-model calls.

Evidence compilation is shared across genes. A fixed condition therefore incurs the same language-model token usage for continuous and classification predictions across the 2,000-gene evaluation panel. Text encoding, atlas retrieval, training, and numerical inference contribute additional computational cost.

J PREDICTION BEYOND THE CONTINUOUS-HEAD GENE PANEL

The additional panel contains the 931 measured LINCS L1000 landmark genes (Subramanian et al., 2017) outside the 2,000 highly variable genes: all 978 landmarks map to measured SciPlex3 genes, and 47 overlap the primary panel. Selection uses gene identities and landmark membership. This panel tests response prediction for biologically defined genes beyond the continuous head's variance-selected output set.

With support DE/DIR label supervision for these genes, the independent classification heads achieve AUPRC of 0.7674/0.9212 and Macro-F1 of 0.7908/0.8535 on 24,630 DE queries and 3,777 true-DE direction queries. Their continuous response targets are absent from the main regression loss. For expression generation, an auxiliary dose-response model is fitted to their continuous support responses (Section B.4); the generated Extra931 responses have logFC Pearson 0.5344 and Δ-Pearson of 0.2583 over 339 conditions. This separates classification with label supervision from continuous generation with response supervision. The combined 2,931-gene panel contains 80,228 sampled queries and 12,428 DE positives.

J.1 PREDICTION AND OBSERVATION FEEDBACK

Table 25: **Classification on the additional and combined gene panels.** DIR uses the true DE subset; its AUPRC, precision, and recall treat up-regulation as the positive class. F1 is two-class Macro-F1. Additional: 24,630/3,777; combined: 80,228/12,428.

Panel	Task	AUPRC	AUROC	F1	Accuracy	Precision	Recall
Additional 931	DE	0.6842	0.8777	0.7466	0.8957	0.8082	0.4194
Additional 931	DIR	0.8879	0.9051	0.8030	0.8144	0.8714	0.6671
All 2,931	DE	0.7710	0.9280	0.7965	0.9100	0.8298	0.5270
All 2,931	DIR	0.9254	0.9186	0.8377	0.8378	0.8692	0.8012

The three-state Macro-F1 is 0.7210 on the combined sampled panel. The prediction results use generated-expression testing; observation-feedback results use the shared condition heads.

Table 26: **Update classification metrics: 931 additional genes.** Sampled endpoint, averaged over three hidden-cell partitions. DIR AP, precision, and recall use up as the positive class. The zero row is the same model without Update.

Budget	Task	AP	AUROC	F1	Accuracy	Precision	Recall
0	DE	0.8179	0.9357	0.8136	0.9051	0.8581	0.5683
0	DIR	0.9157	0.9488	0.8599	0.8676	0.8194	0.8376
48	DE	0.8420	0.9392	0.8430	0.9166	0.8641	0.6421
48	DIR	0.9163	0.9490	0.8651	0.8724	0.8235	0.8472
112	DE	0.8618	0.9553	0.8182	0.9074	0.8685	0.5748
112	DIR	0.9160	0.9490	0.8644	0.8714	0.8210	0.8488
224	DE	0.8578	0.9527	0.8587	0.9240	0.8664	0.6823
224	DIR	0.9188	0.9504	0.8675	0.8745	0.8261	0.8506

Table 27: **Update classification metrics: all 2,931 genes.** Sampled endpoint, averaged over three hidden-cell partitions. DIR AP, precision, and recall use up as the positive class. The zero row is the same model without Update.

Budget	Task	AP	AUROC	F1	Accuracy	Precision	Recall
0	DE	0.8618	0.9615	0.8179	0.9127	0.8867	0.5606
0	DIR	0.9274	0.9284	0.8491	0.8495	0.8665	0.8219
48	DE	0.8774	0.9635	0.8405	0.9209	0.8859	0.6176
48	DIR	0.9264	0.9273	0.8523	0.8527	0.8671	0.8286
112	DE	0.8944	0.9729	0.8235	0.9150	0.8910	0.5723
112	DIR	0.9271	0.9282	0.8540	0.8543	0.8665	0.8335
224	DE	0.8895	0.9711	0.8585	0.9285	0.8876	0.6639
224	DIR	0.9279	0.9288	0.8543	0.8546	0.8689	0.8313

J.2 Sensitivity to the additional gene panel

Table 28: **Sensitivity to the additional 931 classification genes.** All scores use the same 2,000 core genes. The middle row retrains only core heads with the original frontend. The final row removes additional-gene DE/UP values from the atlas and retrains the frontend and heads; the atlas value dimension falls from 7,862 to 6,000. Continuous output remains 2,000 throughout. External biological knowledge and the original label definitions are retained.

Variant	RMSE ↓	DE AP	DE F1	DIR mAP	DIR F1	Joint F1
Complete model	4.4762	0.8588	0.8077	0.9135	0.8393	0.7242
Full atlas, core-only heads	4.4762	0.8588	0.8077	0.9135	0.8390	0.7242
Core-only atlas and heads	4.4715	0.8584	0.8056	0.9102	0.8326	0.7216

Removing the additional 931 genes from atlas response values, classification supervision, and output preserves core DE AP within 0.0005 of the complete model, while reducing direction Macro-F1 by 0.0067. Retraining only core heads with the original frontend changes scores little. The core-only atlas also changes projection size and initialization: frontend-plus-continuous parameters fall from 939,027 to 819,859 per seed. Consequently this comparison measures the combined input-panel and architecture change. External knowledge may still mention genes outside the core panel, and DE labels retain their original multiple-testing definition.