LEARNING WHEN CELLULAR RESPONSES TRANSFER: A FACTORIZED MEMORY-GUIDED AGENT FOR PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

One main goal of virtual cell modeling is to predict cellular responses under unseen conditions, since the combinatorial space of perturbations and cellular contexts is too large to assay exhaustively. Existing perturbation prediction agents use gene-expression responses and knowledge graphs, but gene-level response tendencies can obscure perturbation-specific effects in their predictions. Retrieved evidence can also be unhelpful or misleading when incorporated without assessing its predictive value for the target query. We introduce the Multi-route Memory-guided Cellular Agent with Calibrated Transfer (MMCA-CT), which extracts perturbation-specific information and learns to assess the predictive value of response estimates to guide evidence selection and model reasoning. Its factorized response memory separates stable gene tendencies from perturbation-specific effects to estimate responses under unseen conditions. The agent learns to weight predictions by their estimated improvement over a gene-prior reference and to select response estimates expected to further improve predictions based on observed responses. Across five cancer cell lines, MMCA-CT improves response detection and direction prediction in both cross-perturbation and cross-cell-line settings. With QWEN3.5-9B, it yields relative gains of 15.3% in macro-per-gene AUPRC over VCWORLD for cross-perturbation response detection and 14.0% in macro-per-gene AUROC over CRISP for cross-cell-line direction prediction. Ablation studies further show that factorized response memory yields the largest accuracy gains, with additional improvements from conditional transfer estimation and utility-guided evidence selection.

1 INTRODUCTION

Perturbational transcriptomics measures how interventions change gene expression, but the growing number of perturbation–context combinations limits experimental coverage. Virtual cell models aim to extend this coverage by predicting responses under unseen conditions (Bunne et al., 2024; Zhang et al., 2025). Models trained on single-cell perturbation data learn latent representations to predict cellular responses (Lotfollahi et al., 2019; 2023; Adduri et al., 2026). In parallel, methods based on large language models (LLMs) use biological annotations and related experimental observations to predict perturbation responses with interpretable rationales (Wu et al., 2025; Wei et al., 2026).

Gene-expression responses measured in training experiments reflect both stable gene-level tendencies and effects that depend on the perturbation and cellular context. A gene's general tendency to respond provides a starting point for prediction, but does not specify how it will respond to a particular intervention (Yuan et al., 2026). Representing these stable tendencies separately from perturbation-specific effects offers a way to share information across experiments while estimating the changes relevant to an unseen condition.

Existing perturbation prediction agents retrieve observed gene-expression responses and knowledge graph facts to support prediction (Wei et al., 2026; Yuan et al., 2026). Similar molecular structures, shared drug targets, or pathway relations do not, however, establish whether the retrieved responses and model estimates improve prediction for a given perturbation and cellular context. Irrelevant retrieved information can also reduce the accuracy of LLM predictions (Yoran et al., 2024). Learning

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.

when observed and estimated gene-expression responses improve prediction provides a basis for weighting their contributions and selecting useful estimates.

We introduce MMCA-CT, a memory-guided cellular agent that predicts responses under unseen conditions by combining observed and estimated gene-expression responses. Its factorized response memory separates stable gene tendencies from perturbation-specific effects, supporting response estimation across biologically related conditions. The agent learns to weight predictions by their estimated improvement over a gene-prior reference based on training response frequencies, and selects response estimates for their additional predictive value alongside observed responses. The resulting evidence and fused prediction guide a language model, with retrieved biological knowledge providing context for its decisions and rationales.

Our contributions are:

- We develop a memory-guided cellular agent whose **factorized response memory** separates stable gene tendencies from perturbation-specific effects to estimate responses under unseen perturbations and cellular contexts.
- We develop a query-adaptive approach that links prediction weighting to evidence selection. **Conditional transfer estimation** weights response predictions by their expected gain and reliability, while **utility-guided evidence selection** builds on these weighted predictions to retain response estimates according to their expected incremental predictive value.
- We demonstrate improved macro-per-gene performance in cross-perturbation response detection and cross-cell-line direction prediction across five cancer cell lines. Component ablations show that response memory yields the largest gains in overall accuracy, with further improvements from learned prediction weighting and evidence selection.

2 METHOD

2.1 OVERVIEW AND PROBLEM FORMULATION

We consider differential expression (DE) prediction and direction (DIR) prediction for a query $q = (c, p, g, t)$, where c denotes the cell context, p the perturbation, g the gene, and $t \in \{\mathrm{DE}, \mathrm{DIR}\}$ the task. For DE, $y_{\mathrm{DE}} = 1$ denotes differential expression. DIR is defined only when $y_{\mathrm{DE}} = 1$, with $y_{\mathrm{DIR}} = 1$ for up-regulation and 0 for down-regulation. The agent returns a task decision or abstention, supported by a traceable evidence set $E(q)$.

To predict these outcomes, MMCA-CT obtains task probabilities from observed gene-expression responses, estimates from factorized response memory, and coexpression-based estimates for held-out cell lines (Figure 1). The agent weights these predictions by their estimated improvement over a gene-prior reference. It selects response estimates expected to improve the prediction obtained from observed responses. The selected evidence and fused prediction guide a language model, with facts retrieved from a public knowledge graph providing mechanistic context.

The gene-prior reference uses the positive-label frequency $\pi_{c,g,t}$ for task t, estimated across training perturbations for cell c and gene g. Hierarchical shrinkage stabilizes these estimates by smoothing cell–gene frequencies toward pooled gene-level frequencies and gene-level frequencies toward the global training frequency. For a held-out cell line, we average priors from source cell lines, weighting them by the similarity of untreated-control expression and detection-rate profiles (Appendix A.1). Calibrating the resulting frequency estimate gives the reference probability

$$p_{0,t}(c, g) = \mathrm{Cal}_t(\pi_{c,g,t}), \tag{1}$$

where Cal_t is the setting-specific calibration map. The reference $p_{0,t}(q) := p_{0,t}(c, g)$ is constant across perturbations for fixed (c, g).

Each response route $r \in \mathcal{R}$ produces a numerical response summary using a specific retrieval or response-estimation rule. Memory readouts based on molecular structure and drug targets, for example, define distinct routes over the same learned memory. Each summary provides features for task probability estimation and, together with its supporting records, forms an evidence packet. A public knowledge graph (KG) provides biological annotations for retrieval and memory readouts, together with mechanistic facts for language-model reasoning. A cell-line context KG built from

untreated controls provides baseline expression and coexpression information for cross-cell-line response estimation.

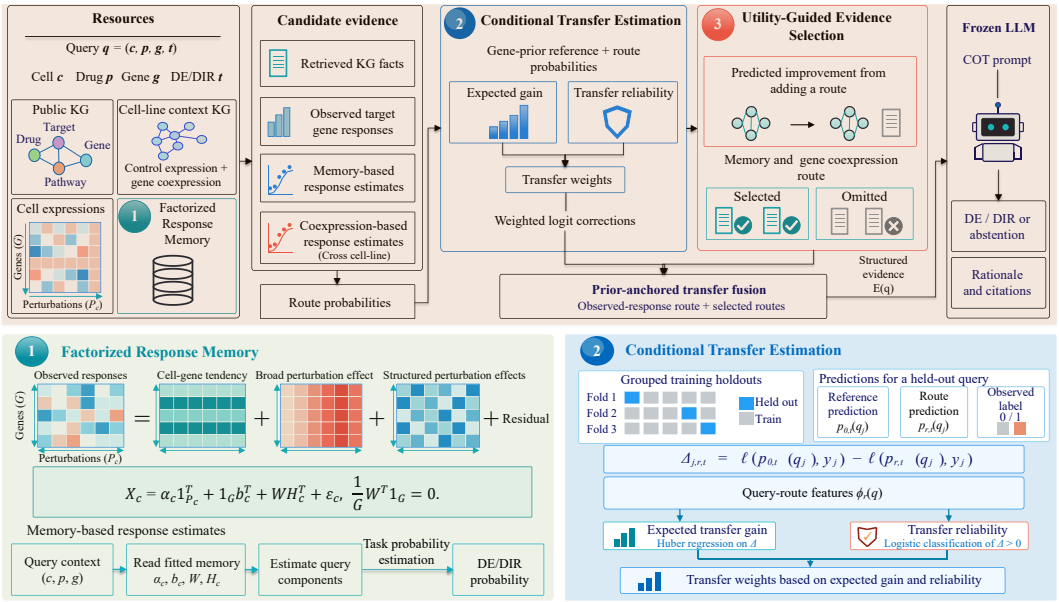

Figure 1: Overview of MMCA-CT. Observed gene-expression responses, knowledge graphs, and factorized response memory ① provide candidate evidence. Conditional transfer estimation ② weights response predictions, while utility-guided evidence selection ③ selects response estimates for fusion. The fused prediction and structured evidence guide LLM reasoning.

2.2 Factorized Response Memory and Route-Conditioned Readouts

For an unseen perturbation, the observed-response route retrieves the target gene's responses to similar training perturbations in the same cell line, ranked by molecular-structure and mechanism-of-action similarity. For a held-out cell line, this route instead pools the same gene's responses to the query perturbation across source cell lines, weighted by similarity of untreated-control expression and detection-rate profiles. The agent also retrieves facts from the public KG through drug–target and pathway relations involving the query drug and gene g (Appendix A.3).

Response-memory construction. Let $X_{cpg} \in \mathbb{R}$ denote the continuous expression effect for gene g under perturbation p in cell line c, relative to controls. For a fixed cell line, these values form a matrix $X_c \in \mathbb{R}^{G \times P_c}$, with G genes as rows and P_c training perturbations as columns, modeled as

$$X_c = \alpha_c 1_{P_c}^\top + 1_G b_c^\top + W H_c^\top + \epsilon_c, \qquad \frac{1}{G} W^\top 1_G = 0. \tag{2}$$

Here 1_n is the length-n all-ones column vector. The vectors $\alpha_c \in \mathbb{R}^G$ and $b_c \in \mathbb{R}^{P_c}$ collect stable cell–gene tendencies α_{cg} and broad perturbation effects b_{cp} shared across genes, respectively. The structured response uses a factorization of rank at most d, with gene loadings $W \in \mathbb{R}^{G \times d}$ and latent cell–perturbation states $H_c \in \mathbb{R}^{P_c \times d}$. Their rows are $W_g^\top$ and $H_{cp}^\top$, respectively, and $\epsilon_c \in \mathbb{R}^{G \times P_c}$ contains the residuals. Centering loadings across the G genes separates this effect from the broad perturbation effect. We fit the memory to training responses using Huber loss (Appendix A.2).

Response reconstruction. For an unseen perturbation, three memory routes compare it with training perturbations using molecular structure, targets and action types, or mechanism descriptions. The structure route uses a Tanimoto kernel on Morgan fingerprints (Rogers & Hahn, 2010). The target route uses a cosine kernel on drug-target and action-type feature vectors, while the mechanism route applies the same kernel to TF–IDF representations of mechanism descriptions (Appendix A.3).

Let $\mathcal{P}_c$ contain the P_c training perturbations in cell line c. For route r, $K_r \in \mathbb{R}^{P_c \times P_c}$ contains their pairwise similarities, and $k_r(p) \in \mathbb{R}^{P_c}$ contains the query perturbation's similarities to those same training perturbations. Using the fitted H_c and b_c, kernel ridge regression estimates the query perturbation's latent state and broad perturbation effect as

$$\widehat{H}_{cp}^{(r)} = H_c^\top (K_r + \lambda_r I)^{-1} k_r(p), \qquad \widehat{b}_{cp}^{(r)} = k_r(p)^\top (K_r + \lambda_r I)^{-1} b_c, \tag{3}$$

where $\lambda_r > 0$ is the regularization strength and I is the identity matrix.

For a held-out cell line $c^\star$, we fit memory across source cell lines $\mathcal{C}_{\mathrm{src}}$ with shared gene loadings W. A single memory-transfer route combines the fitted latent states and broad perturbation effects from $\mathcal{C}_{\mathrm{src}}(p)$, the source cell lines with observed responses to the query perturbation, as

$$\widehat{H}_{c^\star p}^{(r)} = \sum_{c \in \mathcal{C}_{\mathrm{src}}(p)} \widetilde{\omega}_{c|c^\star,p} H_{cp}, \qquad \widehat{b}_{c^\star p}^{(r)} = \sum_{c \in \mathcal{C}_{\mathrm{src}}(p)} \widetilde{\omega}_{c|c^\star,p} b_{cp}, \tag{4}$$

where $\widetilde{\omega}_{c|c^\star,p}$ are normalized similarity weights based on untreated-control expression and detection-rate profiles. Both settings use the decomposition in Equation 2 to reconstruct the target gene's response as

$$\widehat{X}_{cpg}^{(r)} = \widehat{\alpha}_{cg} + \widehat{b}_{cp}^{(r)} + W_g^\top \widehat{H}_{cp}^{(r)}, \tag{5}$$

where $\widehat{\alpha}_{cg} = \alpha_{cg}$ for a cell line with training responses. For a held-out cell line, we average the stable cell–gene tendencies α_{cg} from source cell lines, weighting them by similarity between untreated-control profiles (Appendix A.4).

For held-out cell lines, an additional coexpression route estimates the target gene's response from its positively coexpressed neighbors in the cell-line context KG. Built from untreated controls, the KG records baseline expression and detection rates, with coexpression estimated using CS-CORE (Su et al., 2023). The route aggregates these neighbors' responses to the query drug in source cell lines using coexpression edge weights (Appendix A.4).

DE and DIR probability estimation. A task-specific logistic regression maps each route's numerical feature vector $f_{r,t}(q)$ to the probability of the task's positive class. For memory routes, these features combine the reconstructed response, broad perturbation effect, and structured effect $(\widehat{X}_{cpg}^{(r)}, \widehat{b}_{cp}^{(r)}, W_g^\top \widehat{H}_{cp}^{(r)})$ and their products with a retrieval-support score summarizing the availability and similarity of supporting training responses. DE uses effect magnitudes, whereas DIR uses signed effects. For the observed-response route, the features summarize positive and negative task labels in the retrieved target-gene responses using the corresponding retrieval weights. For the coexpression route, the features include the target-gene response estimate obtained by aggregating neighboring-gene responses. Both routes also include features describing the amount of supporting data and disagreement among responses. With the gene-prior reference logit as a fixed offset, the route probability is

$$\operatorname{logit} p_{r,t}(q) = \operatorname{logit} p_{0,t}(q) + \nu_{r,t} + \eta_{r,t}^\top f_{r,t}(q), \tag{6}$$

where $\nu_{r,t}$ and $\eta_{r,t}$ are the intercept and coefficients fitted to training labels (Appendix A.1).

2.3 Conditional transfer estimation

For each available response route, we learn whether and by how much its prediction improves on the gene-prior reference. We obtain training targets from grouped out-of-fold comparisons, fitting the reference and route models without the held-out perturbation group or source cell line. For a held-out training query q_j with observed label y_j, the log-loss reduction is

$$\Delta_{j,r,t} = \ell(p_{0,t}(q_j), y_j) - \ell(p_{r,t}(q_j), y_j)\,. \tag{7}$$

We use the binary cross-entropy objective, denoted by $\ell(p, y)$. Positive Δ means the route assigns a higher probability to the observed label y_j than the reference on the same query.

We predict this improvement for a new query from the numerical feature vector $\phi_r(q)$, which includes the reference and route probabilities and their probability and logit differences $\big(p_{0,t}(q),\, p_{r,t}(q),\, p_{r,t}(q) - p_{0,t}(q),\, \operatorname{logit} p_{r,t}(q) - \operatorname{logit} p_{0,t}(q)\big)$. Retrieval and readout records supply the remaining features, describing the effective number of supporting cases or source cell lines, their similarity to the query, disagreement among their responses, and the query's departure from

conditions represented in training (Appendix A.5). From these features, we estimate the probability of improvement and expected gain as

$$\mathrm{Rel}_{r,t}(q) = \Pr(\Delta_{r,t} > 0 \mid \phi_r(q)), \qquad \mathrm{Gain}_{r,t}(q) = \mathbb{E}[\Delta_{r,t} \mid \phi_r(q)]. \tag{8}$$

We fit logistic regression to $\mathbf{1}[\Delta > 0]$ for transfer reliability and Huber regression to Δ for robust gain prediction, using $\phi_r(q)$ as input to both models.

For a new query, higher predicted gain and reliability increase the transfer weight $\omega_{r,t}(q) \in [0, 1]$, while limited response support and extrapolation beyond training conditions reduce it (Appendix A.5). This weight scales the route's logit correction

$$z_{r,t}(q) = \omega_{r,t}(q)\,\mathrm{clip}_{r,t}[\mathrm{logit}\, p_{r,t}(q) - \mathrm{logit}\, p_{0,t}(q)]\,. \tag{9}$$

Here $\mathrm{clip}_{r,t}$ bounds logit differences using out-of-fold training quantiles. These weighted corrections enter the fusion model used for evidence selection and final prediction in Section 2.4.

2.4 Utility-guided evidence selection and reasoning

The agent retains retrieved target-gene response cases and selects which memory or coexpression estimates to add. To evaluate these additions, a shared fusion model combines the gene-prior reference with the transfer-weighted corrections from Section 2.3. For any subset $S \subseteq \mathcal{R}$ of the query's available response routes, the resulting task probability is

$$\mathrm{logit}\, p_{\mathrm{fuse},t}(q; S) = \mathrm{logit}\, p_{0,t}(q) + \beta_{0,t} + \sum_{r \in S} \beta_{r,t} z_{r,t}(q), \qquad \beta_{r,t} \geq 0. \tag{10}$$

The task-specific intercept $\beta_{0,t}$ and route coefficients $\beta_{r,t}$ are shared across queries and learned by minimizing log loss on out-of-fold predictions.

Let S_{core} contain the observed-response route when available and be empty otherwise. During training, we measure the reduction in log loss from adding a memory or coexpression route r to this set, keeping the fusion parameters and calibration fixed. For a held-out training query q_j with observed label y_j, the utility target is

$$U_{j,r,t} = \ell(\widetilde{p}_{\mathrm{fuse},t}(q_j; S_{\mathrm{core}}), y_j) - \ell(\widetilde{p}_{\mathrm{fuse},t}(q_j; S_{\mathrm{core}} \cup \{r\}), y_j)\,. \tag{11}$$

Here $\widetilde{p}_{\mathrm{fuse},t}$ denotes the calibrated fusion probability for cross-cell-line DIR and equals $p_{\mathrm{fuse},t}$ otherwise (Appendix A.5). We train a Huber regressor on these held-out loss reductions to predict $\widehat{U}_{r,t}(q)$. Its inputs include $\phi_r(q)$, predicted gain and reliability, and $|\beta_{r,t} \cdot z_{r,t}(q)|$, the magnitude of the route's contribution to the fusion logit. We obtain these training targets and features through grouped cross-fitting (Algorithm 1, Appendix A.6).

At inference, let $\mathcal{S}(q)$ contain subsets of available memory and coexpression routes with positive predicted utility that fit the evidence budget (Appendix A.7). We select the subset with the highest total predicted utility and combine it with the observed-response route set

$$S^{\star}_{\mathrm{sup}}(q) \in \arg\max_{S_{\mathrm{sup}} \in \mathcal{S}(q)} \sum_{r \in S_{\mathrm{sup}}} \widehat{U}_{r,t}(q), \qquad S = S_{\mathrm{core}} \cup S^{\star}_{\mathrm{sup}}(q). \tag{12}$$

Evaluating the fusion model (Equation 10) on the selected set S, with the same calibration where applicable, gives the final probability $\widetilde{p}_{\mathrm{fuse},t}(q; S)$.

We add the selected response estimates and fused prediction to the retrieved evidence to form $E(q)$ (Appendix B). The agent supplies $E(q)$ to the frozen LLM with a fixed chain-of-thought (CoT) prompt (Wei et al., 2022). When the fused prediction meets the task-specific support criteria, it determines the output class, and the LLM provides a biological rationale, confidence, and evidence citations. Otherwise, the LLM considers the available predictions and biological evidence to return a task decision or abstention (Appendix B.2; Algorithm 2).

3 Experiments

3.1 Experimental setup

We evaluate DE and DIR in cross-perturbation and cross-cell-line settings using GeneTAK (Wei et al., 2026), derived from Tahoe-100M (Zhang et al., 2025). The evaluation covers 2,000 benchmark-matched highly variable genes in C32, HepG2/C3A, HOP62, Hs 766T, and PANC-1. Within each

split, methods share the labeled training pool and test queries, with response-dependent fitting and retrieval restricted to training data (Appendix C).

Setting-specific baselines include the gene-prior predictor Prior vote, data-driven perturbation models, and VCWORLD, as detailed below. QWEN3.5-9B (Qwen Team, 2026) is the primary backbone for agent comparisons, the case study, and ablations. Additional comparisons examine QWEN3.5-27B and Llama3.1-8B (Dubey et al., 2024), keeping each agent's evidence construction and reasoning prompt unchanged (Section 3.4). Model configurations are provided in Appendix E.

Evaluation considers both overall performance and macro-per-gene performance, which assesses how well models predict a given gene's responses to different perturbations (Yuan et al., 2026). The main tables present accuracy, AUROC, and AUPRC under both evaluations. Additional results include precision, recall, F1, coverage, and abstention rate (Appendix F). Metric definitions and scoring conventions are provided in Appendix D.

3.2 CROSS-PERTURBATION GENERALIZATION

We evaluate generalization to held-out perturbations within each cell line (Appendix C). Comparisons include Prior vote, GAT (Veličković et al., 2018), scVI (Lopez et al., 2018), STATE (Adduri et al., 2026), and VCWORLD (Wei et al., 2026).

MMCA-CT achieves the highest mean overall accuracy, AUROC, and AUPRC on both tasks (Table 1). DE and DIR accuracies reach 0.933 and 0.784, compared with 0.905 and 0.775 for STATE, the strongest baseline in overall accuracy.

Table 1: Cross-perturbation generalization.

		Overall			Macro-per-gene		
Task	**Method**	**Accuracy ↑**	**AUROC ↑**	**AUPRC ↑**	**Accuracy ↑**	**AUROC ↑**	**AUPRC ↑**
DE	Prior vote	0.903±0.009	0.943±0.007	0.931±0.011	0.783±0.013	0.500±0.000	0.628±0.049
	GAT	0.597±0.032	0.636±0.042	0.612±0.082	0.557±0.010	0.556±0.016	0.658±0.044
	scVI	0.554±0.053	0.582±0.063	0.560±0.105	0.536±0.011	0.514±0.011	0.633±0.047
	STATE	0.905±0.013	0.905±0.008	0.930±0.010	0.774±0.013	0.498±0.004	0.627±0.049
	VCWORLD (QWEN3.5-9B)	0.686±0.010	0.687±0.012	0.613±0.042	0.608±0.015	0.497±0.008	0.641±0.047
	MMCA-CT (QWEN3.5-9B)	**0.933±0.011**	**0.953±0.008**	**0.952±0.012**	**0.809±0.013**	**0.577±0.042**	**0.739±0.014**
DIR	Prior vote	0.747±0.026	0.843±0.020	0.817±0.042	0.716±0.018	0.500±0.000	0.476±0.077
	GAT	0.602±0.022	0.640±0.027	0.590±0.077	0.544±0.014	0.560±0.016	0.510±0.070
	scVI	0.538±0.009	0.538±0.018	0.660±0.087	0.502±0.002	0.504±0.005	0.478±0.075
	STATE	0.775±0.023	0.760±0.019	0.782±0.045	0.698±0.018	0.500±0.002	0.476±0.077
	VCWORLD (QWEN3.5-9B)	0.537±0.018	0.537±0.008	0.481±0.096	0.523±0.017	0.511±0.012	0.485±0.070
	MMCA-CT (QWEN3.5-9B)	**0.784±0.026**	**0.856±0.016**	**0.829±0.047**	**0.739±0.010**	**0.562±0.036**	**0.568±0.079**

Prior vote achieves 0.903 overall DE accuracy but a macro AUROC of 0.500, as it assigns each gene its training positive-label frequency regardless of perturbation. MMCA-CT reaches 0.577 DE macro AUROC and 0.739 macro AUPRC, with the latter exceeding GAT's 0.658 and VCWORLD's 0.641. For DIR, MMCA-CT also achieves the highest macro accuracy and AUPRC. These results suggest that MMCA-CT provides perturbation-specific predictive information beyond gene-level response tendencies.

3.3 CROSS-CELL-LINE GENERALIZATION

We evaluate each cell line as a held-out target, training on the other four source cell lines. Test drugs have observed responses in the source lines, while their target-line responses are withheld from fitting and retrieval (Appendix C.3). Comparisons include scVI (Lopez et al., 2018), scGen (Lotfollahi et al., 2019), CRISP (Wang et al., 2026), and STATE (Adduri et al., 2026).

MMCA-CT achieves the highest mean overall accuracy, AUROC, and AUPRC on both tasks (Table 2). DE and DIR accuracies reach 0.731 and 0.792, compared with 0.621 and 0.644 for CRISP, the strongest baseline in overall accuracy.

Macro-per-gene gains are clearest in DIR, where MMCA-CT achieves the highest accuracy, AUROC, and AUPRC. DIR macro AUROC reaches 0.870, compared with CRISP's 0.763. For DE, MMCA-CT also achieves the highest macro accuracy, reaching 0.633 compared with CRISP's 0.579.

Table 2: Cross-cell-line generalization.

Task	Method	Overall			Macro-per-gene		
		Accuracy ↑	AUROC ↑	AUPRC ↑	Accuracy ↑	AUROC ↑	AUPRC ↑
DE	scVI	0.500±0.027	0.500±0.038	0.500±0.032	0.490±0.023	0.500±0.006	0.521±0.027
	scGen	0.495±0.033	0.495±0.051	0.597±0.043	0.480±0.031	0.490±0.008	0.515±0.037
	CRISP	0.621±0.037	0.621±0.063	0.711±0.053	0.579±0.039	**0.578±0.010**	**0.665±0.046**
	STATE	0.586±0.035	0.586±0.031	0.672±0.059	0.484±0.043	0.549±0.011	0.630±0.050
	MMCA-CT (QWEN3.5-9B)	**0.731±0.040**	**0.774±0.063**	**0.794±0.053**	**0.633±0.039**	0.576±0.010	0.644±0.046
DIR	scVI	0.504±0.028	0.506±0.023	0.505±0.023	0.500±0.018	0.500±0.026	0.509±0.023
	scGen	0.578±0.044	0.578±0.036	0.700±0.037	0.560±0.029	0.580±0.041	0.575±0.036
	CRISP	0.644±0.042	0.644±0.045	0.729±0.046	0.618±0.036	0.763±0.051	0.772±0.045
	STATE	0.643±0.054	0.643±0.045	0.728±0.046	0.636±0.036	0.722±0.051	0.717±0.045
	MMCA-CT (QWEN3.5-9B)	**0.792±0.048**	**0.867±0.045**	**0.860±0.046**	**0.787±0.036**	**0.870±0.051**	**0.849±0.045**

Across both generalization settings, learned transfer reliability outperforms biological similarity scores in identifying when predictions based on observed responses, memory-based estimates, or coexpression-based estimates improve on the gene-prior reference (Appendix Figure 4).

3.4 Robustness across language models

To assess whether predictive gains persist across language models, we evaluate MMCA-CT with QWEN3.5-9B, QWEN3.5-27B, and Llama3.1-8B in both settings. STATE provides a common baseline across backbone configurations, while VCWorld is evaluated with the same QWEN3.5 backbones in the cross-perturbation setting.

In the matched-backbone comparison, MMCA-CT exceeds VCWorld at both QWEN3.5 sizes in mean overall precision, recall, and F1, as well as macro-per-gene F1, for DE and DIR (Table S5). With QWEN3.5-9B, mean macro accuracy exceeds VCWorld by 20.1 percentage points for DE and 21.6 for DIR. Increasing the QWEN3.5 backbone from 9B to 27B yields limited gains in mean macro-per-gene performance across both tasks and settings.

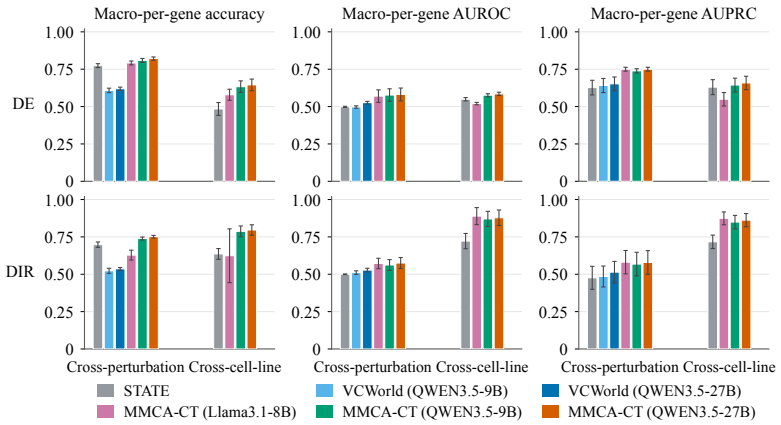

Figure 2: Macro-per-gene performance across language models. Error bars indicate standard deviations across cell lines.

With Llama3.1-8B, MMCA-CT achieves a cross-perturbation DE macro AUPRC of 0.750, compared with STATE's 0.627. The DIR ranking advantage over STATE persists across all three backbones in both settings, with higher mean macro-per-gene AUROC and AUPRC throughout (Figure 2). Complete results are provided in Appendix F, with accuracy and coverage reported in Tables S6 and S9.

3.5 Biological analysis of response predictions

We analyze DIR predictions across biological processes and individual genes using QWEN3.5-9B. For cross-cell-line evaluation, we compare MMCA-CT with Prior vote, which uses each gene's up-

regulation frequency in the other four cell lines. Each entry in Figure 3(a) corresponds to one MSigDB Hallmark gene set and one held-out cell line (Liberzon et al., 2015). It reports the DIR accuracy gain over Prior vote on identical drug–gene queries pooled across test perturbations, restricted to genes belonging to that set.

Of the 50 Hallmark gene sets, 29 have sufficient coverage in at least one cell line, yielding 130 eligible gene-set–cell-line pairs. MMCA-CT achieves higher DIR accuracy in 128 of these pairs, showing that improvements over the gene-prior baseline extend across the evaluated biological processes and target cell lines. Coverage criteria and additional baseline comparisons are provided in Appendix F.4.

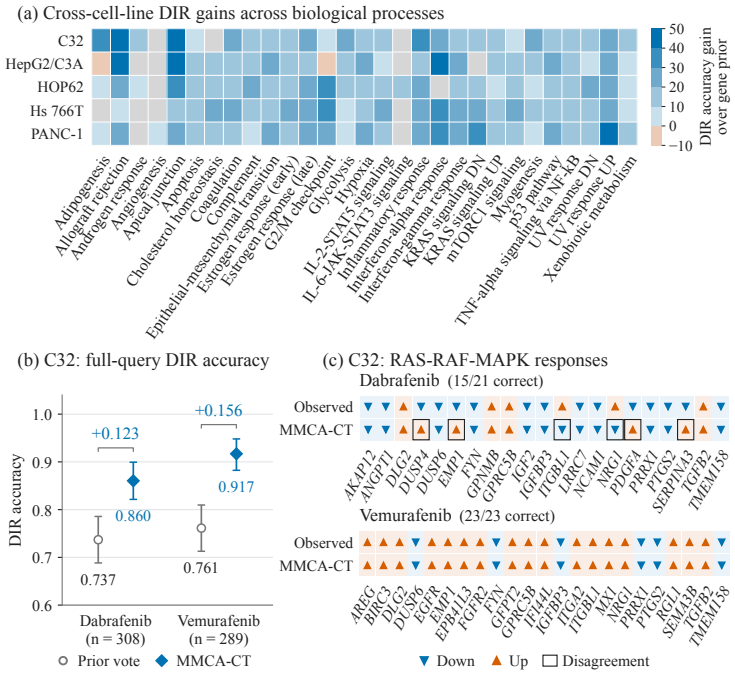

Figure 3: Process-level and gene-level DIR prediction. (a) Cross-cell-line accuracy gains over Prior vote (percentage points); gray indicates insufficient coverage. (b) Accuracy across all DIR queries for held-out BRAF inhibitors in C32. n counts queries; error bars show 95% bootstrap confidence intervals. (c) Observed and predicted directions in the C32 RAS–RAF–MAPK response panel.

C32 carries the activating BRAF V600E mutation, which promotes MAPK signaling (Bairoch, 2018; Tovar-Parra & Zammit-Mangion, 2024). We therefore examine responses to the BRAF inhibitors dabrafenib and vemurafenib in this cell line under a separate perturbation holdout, with their treated responses excluded from fitting and retrieval (Appendix F.4). Across all DIR queries for each inhibitor, MMCA-CT achieves accuracies of 0.860 and 0.917, compared with 0.737 and 0.761 for Prior vote, respectively (Figure 3(b)).

To examine gene-level predictions related to this signaling pathway, we define a RAS–RAF–MAPK response panel as the union of the Hallmark KRAS signaling up/down sets and Reactome RAF/MAP kinase cascade participants (Liberzon et al., 2015; Milacic et al., 2024). This combines transcriptional signatures of KRAS activation with annotated pathway components. The analysis includes all DIR queries whose genes belong to this panel. MMCA-CT correctly predicts 15 of 21 response directions for dabrafenib and all 23 for vemurafenib (Figure 3(c)).

3.6 Component ablations

We successively add factorized response memory (A), conditional transfer estimation and fusion (B), and utility-guided evidence selection (C) in the cross-perturbation setting with QWEN3.5-9B (Table 3). The Base evidence configuration uses retrieved KG facts and observed target-gene responses for LLM reasoning (Wei et al., 2026). All available memory routes are retained until C introduces selection (Appendix F.5).

Table 3: Component ablations for cross-perturbation generalization (QWEN3.5-9B).

		Overall			Macro-per-gene		
Task	**Configuration**	**Accuracy ↑**	**AUROC ↑**	**AUPRC ↑**	**Accuracy ↑**	**AUROC ↑**	**AUPRC ↑**
DE	Base evidence	0.686±0.010	0.687±0.012	0.613±0.042	0.608±0.015	0.497±0.008	0.641±0.047
	+ A	0.904±0.015	0.920±0.016	0.915±0.018	0.761±0.009	0.533±0.039	0.687±0.065
	+ B	0.921±0.011	0.932±0.008	0.922±0.012	0.788±0.013	0.561±0.042	0.728±0.014
	+ C	**0.933±0.011**	**0.953±0.008**	**0.952±0.012**	**0.809±0.013**	**0.577±0.042**	**0.739±0.014**
DIR	Base evidence	0.537±0.018	0.537±0.008	0.481±0.096	0.523±0.017	0.511±0.012	0.485±0.070
	+ A	0.747±0.028	0.799±0.017	0.728±0.076	0.707±0.015	0.535±0.017	0.538±0.075
	+ B	0.765±0.026	0.829±0.016	0.745±0.047	0.719±0.010	0.547±0.036	0.549±0.079
	+ C	**0.784±0.026**	**0.856±0.016**	**0.829±0.047**	**0.739±0.010**	**0.562±0.036**	**0.568±0.079**

A: factorized response memory; B: conditional transfer estimation and fusion; C: utility-guided evidence selection.

Factorized response memory produces the largest stepwise gains in overall accuracy, raising DE from 0.686 to 0.904 and DIR from 0.537 to 0.747. Adding conditional transfer estimation and fusion further improves all metrics in Table 3, including DE macro AUPRC from 0.687 to 0.728. Utility-guided evidence selection adds gains beyond transfer weighting, raising DIR overall AUPRC from 0.745 to 0.829 and improving all macro-per-gene metrics. Additional metrics are provided in Appendix Table S11.

4 Related Work

Latent-variable models learn gene-expression representations (Lopez et al., 2018) and predict responses across perturbations and cellular contexts (Lotfollahi et al., 2019; 2023). At the population level, transport and diffusion approaches model perturbation-induced changes in cellular-state distributions (Bunne et al., 2023; Driessen et al., 2026; He et al., 2026). Prediction under unseen conditions is supported by pretrained cell representations (Cui et al., 2024; Adduri et al., 2026; Wang et al., 2026) and chemical features for unobserved compounds (Hetzel et al., 2022; Qi et al., 2024). Biological priors further inform prediction through gene coexpression and functional relationships (Roohani et al., 2024; Alsulami et al., 2026), as well as multiple knowledge graphs (Wenkel et al., 2026).

To connect response prediction with biological interpretation, LLM-based agents use experimental observations and biological knowledge to support gene-level decisions and generate rationales linked to the supplied evidence (Wu et al., 2025). VCWorld supplies KG facts, cellular context, and related response cases (Wei et al., 2026), while CORE organizes same-gene observations into positive and negative evidence to compare related perturbations (Yuan et al., 2026). To improve reasoning quality, PBio-Agent coordinates specialized agents with iterative verification (Kim et al., 2026), while VCR-Agent verifies explanations represented as mechanistic action graphs (Jang et al., 2026). In retrieval-augmented language models, entailment-based filtering assesses whether retrieved passages support a candidate answer to the question (Yoran et al., 2024). Evidence relevance and factual support do not by themselves establish predictive benefit for a new query. MMCA-CT learns this query-specific benefit from out-of-fold comparisons, using it to guide prediction weighting and evidence selection.

5 Conclusion

MMCA-CT addresses perturbation prediction under unseen conditions by separating stable gene tendencies from perturbation-specific effects in factorized response memory. The agent learns to downweight response predictions with low estimated reliability and exclude additional estimates with no expected predictive benefit from the evidence supplied to the language model. Across five cancer cell lines, it improves macro-per-gene performance in cross-perturbation response detection and cross-cell-line direction prediction. Drug–target and protein–protein interactions may also support extension to genetic perturbations by linking perturbed genes to related response evidence. These findings suggest that learning the predictive contribution of response evidence can improve generalization to unseen perturbations and cellular contexts while keeping predictions linked to supporting biological evidence.

AI USE STATEMENT

In our work, we mainly use frontier models for writing enhancements, primarily to improve grammar and text clarity.

ETHICS STATEMENT

All datasets used in this study are publicly available. Our analyses use existing transcriptomic measurements from cultured cell lines.

REPRODUCIBILITY STATEMENT

Appendices A and B describe model fitting and evidence-grounded reasoning, with algorithm pseudocode and recorded query examples. Appendices C–E document data preparation, evaluation splits, metric definitions, and baseline and language-model configurations.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

Reem Alsulami, Robert Lehmann, Sumeer A. Khan, Vincenzo Lagani, Alberto Maillo, David Gomez-Cabrero, Narsis A. Kiani, and Jesper Tegner. Predicting and interpreting cell-type-specific drug responses in the small-data regime using inductive priors. *Nature Machine Intelligence*, 8(3): 461–473, 2026. doi: 10.1038/s42256-026-01202-2.

Amos Bairoch. The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques*, 29(2):25–38, 2018. doi: 10.7171/jbt.18-2902-002.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

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

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

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(6):984–996, 2026. doi: 10.1038/s42256-026-01242-8.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783v1*, 2024.

Simon D. Harding, Jane F. Armstrong, Elena Faccenda, Christopher Southan, Stephen P. H. Alexander, Anthony P. Davenport, Michael Spedding, and Jamie A. Davies. The IUPHAR/BPS Guide to PHARMACOLOGY in 2026. *Nucleic Acids Research*, 54(D1):D1446–D1456, 2026. doi: 10.1093/nar/gkaf1067.

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(1):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. doi: 10.52202/068431-1937.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Yunhui Jang, Lu Zhu, Jake Fawkes, Alisandra Kaye Denton, Dominique Beaini, and Emmanuel Noutahi. Towards autonomous mechanistic reasoning in virtual cells. *arXiv preprint arXiv:2604.11661*, 2026.

Hyomin Kim, Sang-Yeon Hwang, Jaechang Lim, Yinhua Piao, Yunhak Oh, Woo Youn Kim, Chanyoung Park, Sungsoo Ahn, and Junhyeok Jeon. Progressive multi-agent reasoning for biological perturbation prediction. *arXiv preprint arXiv:2602.07408*, 2026.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, et al. PubChem 2025 update. *Nucleic Acids Research*, 53 (D1):D1516–D1525, 2025. doi: 10.1093/nar/gkae1059.

Craig Knox, Mike Wilson, Christen M. Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A. Strawbridge, et al. DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024. doi: 10.1093/nar/gkad976.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):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, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025.

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(1):9256, 2024. doi: 10.1038/s41467-024-53457-1.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen model release, 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Model cards: `https://huggingface.co/Qwen/Qwen3.5-9B` and `https://huggingface.co/Qwen/Qwen3.5-27B`.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Chang Su, Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao, and Jingfei Zhang. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nature Communications*, 14(1): 4846, 2023. doi: 10.1038/s41467-023-40503-7.

David Tovar-Parra and Marion Zammit-Mangion. Comparative analysis of the effect of the BRAF inhibitor dabrafenib in 2D and 3D cell culture models of human metastatic melanoma cells. *In Vivo*, 38(4):1579–1593, 2024. doi: 10.21873/invivo.13608.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Yixuan Wang, Xinyuan Liu, Yimin Fan, Binghui Xie, James Cheng, Kam Chung Wong, Peter Cheung, Irwin King, and Yu Li. Predicting drug responses of unseen cell types through transfer learning with foundation models. *Nature Computational Science*, 6(1):39–52, 2026. doi: 10.1038/s43588-025-00887-6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800.

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.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: Using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4. Advance online publication.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

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

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

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

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J. Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F. Mosquera, et al. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024. doi: 10.1093/nar/gkad1004.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, et al. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

A Additional Method Details

A.1 Gene-prior reference and task probabilities

The gene-prior reference in Section 2.1 uses hierarchical shrinkage to stabilize cell–gene frequencies using gene-level and global training frequencies. For task t, let n^+ and n denote positive and total training-label counts, with subscripts identifying genes and cell lines. The smoothed frequencies are

$$\pi_t = \frac{n_t^+ + 1}{n_t + 2}, \qquad \pi_{g,t} = \frac{n_{g,t}^+ + \lambda_0 \pi_t}{n_{g,t} + \lambda_0}, \qquad \pi_{c,g,t} = \frac{n_{c,g,t}^+ + \lambda_0 \pi_{g,t}}{n_{c,g,t} + \lambda_0}, \tag{13}$$

with shrinkage strength $\lambda_0 = 5$ in both settings. For cross-perturbation prediction, the calibration map Cal_t in Equation 1 is an ℓ_2-regularized logistic regression on $\mathrm{logit}\, \pi_{c,g,t}$, fitted with perturbation-balanced weights and a slope that preserves prior ranking. For a held-out cell line, $\pi_{c^\star,g,t} = \sum_{c \in \mathcal{C}_{\mathrm{src}}} \omega_{c|c^\star} \pi_{c,g,t}$ uses the normalized untreated-control similarity weights in Appendix A.4, and Cal_t is the identity.

For memory routes, the feature vector $f_{r,t}(q)$ in Equation 6 is constructed from the response components as

$$\begin{aligned} \mathbf{v}_{r,\mathrm{DIR}}(q) &= \left(\widehat{X}_{cpg}^{(r)}, \widehat{b}_{cp}^{(r)}, W_g^\top \widehat{H}_{cp}^{(r)}\right)^\top, \\ \mathbf{v}_{r,\mathrm{DE}}(q) &= |\mathbf{v}_{r,\mathrm{DIR}}(q)|, \qquad f_{r,t}(q) = \mathrm{Scale}_{r,t}\begin{bmatrix} \mathbf{v}_{r,t}(q) \\ s_r(q)\mathbf{v}_{r,t}(q) \end{bmatrix}. \end{aligned} \tag{14}$$

Absolute values are elementwise, and $\mathrm{Scale}_{r,t}$ clips and standardizes the six features using training-fold statistics.

Let $n_{\mathrm{eff},r}(q)$ denote the effective number of supporting training perturbations or source cell lines. The retrieval-support score $s_r(q) \in [0,1]$ is the minimum of their maximum similarity to the query, one minus the readout extrapolation score, and $n_{\mathrm{eff},r}(q)/(n_{\mathrm{eff},r}(q)+2)$. For perturbation readouts, the extrapolation score is $1 - \mathrm{clip}_{[0,1]}[k_r(p)^\top (K_r + \lambda_r I)^{-1} k_r(p)]$. For cross-cell-line readouts, it is one minus the maximum untreated-control similarity to an eligible source cell line.

The observed-response and coexpression routes use the weighted summaries defined in Appendices A.3 and A.4, respectively. The task models in Equation 6 minimize ℓ_2-regularized binary log loss on training labels, with the gene-prior reference logit held fixed.

A.2 Robust response-memory optimization

For the response memory in Equation 2, we set α_{cg} to the median observed expression effect for gene g in cell line c across training perturbations within each fold. We fit the broad perturbation effects, gene loadings, and latent cell–perturbation states by alternating optimization under Huber loss (Huber, 1964), holding the stable cell–gene tendencies fixed. Centering the gene loadings and absorbing their mean contribution into b_c enforces the centering constraint in Equation 2 without changing the reconstructed responses.

Gene loadings use ℓ_1, squared ℓ_2, and orthogonality penalties. Latent states use squared ℓ_2 regularization. In cross-cell-line fitting, they are also shrunk toward the mean state for the same perturbation across source cell lines. Table S1 lists the fitting settings. Cross-perturbation rank selection uses grouped out-of-fold gains in DE macro-per-gene AUPRC and DIR macro-per-gene AUROC, relative to the best-performing control for each task and route. Controls include the gene-prior reference, predictors based on broad perturbation effects or gene modules, and direct-response kernel ridge regression. We select the rank with the highest average of the best route gains for the two tasks. The selected ranks are 4 for C32 and Hs 766T and 32 for HepG2/C3A, HOP62, and PANC-1. Cross-cell-line prediction uses rank 4 in all five outer folds.

A.3 Route construction

For cross-perturbation prediction, the observed-response route retrieves the target gene's responses in the same cell line from training perturbations ranked by biological similarity. This ranking combines mechanism-of-action semantic similarity and molecular-fingerprint similarity with weights 0.7 and 0.3, respectively (Wei et al., 2026). Retrieved cases are weighted by the inverse square root of their

Table S1: Response-memory fitting settings.

Parameter	Value
Cross-perturbation rank candidates	$\{2, 4, 8, 16, 32, 64\}$
Cross-cell-line rank	4
Huber loss transition threshold	1
Gene-loading ℓ_1 penalty weight	0.002
Gene-loading squared ℓ_2 penalty weight	0.05
Gene-loading orthogonality penalty weight	0.01
Latent-state squared ℓ_2 penalty weight	0.1
Latent-state shrinkage weight	1

rank, with the numbers of positive and negative cases balanced as availability permits. Their weighted label contrast, the difference between total positive and negative case weights divided by their sum, enters the task model together with support and response disagreement features.

The three memory routes use the kernels defined in Section 2.2, with nonnegative features for drug targets and action types. All three use $\lambda_r = 0.1$ in Equation 3 and require a valid query representation, at least three training perturbations with positive kernel similarity, and a summed query-to-training kernel similarity of at least 0.1.

The public KG extends the protein, gene, pathway, and complex relations released with VCWORLD (Wei et al., 2026). Full InChIKeys from PubChem (Kim et al., 2025) link compounds to interaction records for human targets in GtoPdb (Harding et al., 2026) and mechanism records in ChEMBL (Zdrazil et al., 2024), supplemented by licensed DrugBank target records (Knox et al., 2024). Protein targets, families, and complexes are linked to their corresponding genes or members using shared identifiers, with ambiguous compound mappings excluded. Retrieved drug–target and pathway relations retain their sources and action annotations. These facts provide mechanistic context; response direction is inferred from observed and estimated expression changes.

A.4 CELL-LINE CONTEXT KG AND CROSS-CELL-LINE RESPONSE ROUTES

For each cell line, the context KG records baseline expression, detection rates, and coexpression among the benchmark genes from DMSO controls. CS-CORE (Su et al., 2023) estimates coexpression from sequencing depths and UMI count estimates recovered by reversing the released normalization. Coexpression estimates are aggregated by taking medians within and across plates. We retain positive coexpression edges that remain stable across subsets of the control data and plate resampling, with weights combining association strength and bootstrap stability.

To compare cell lines, the untreated-control profile u_c concatenates the cell-line context KG's per-gene median $\log(1 + \mathrm{TP10K})$ expression and detection rates. TP10K denotes counts normalized to 10^4 transcripts per cell, and each detection rate is the fraction of control cells with nonzero counts for that gene. After feature-wise standardization using means and population standard deviations computed across source cell lines, source–target similarity is $\kappa_{\mathrm{cell}}(u_{c^\star}, u_c) = (1 + \cos(u_{c^\star}, u_c))/2$, requiring at least 20 shared features. Normalizing over eligible source cell lines gives

$$\omega_{c|c^\star} = \frac{\kappa_{\mathrm{cell}}(u_{c^\star}, u_c)}{\sum_{d \in \mathcal{C}_{\mathrm{src}}} \kappa_{\mathrm{cell}}(u_{c^\star}, u_d)}, \tag{15}$$

with equal weights when all eligible similarities are zero. These weights define the stable cell–gene tendency in the target cell line $\widehat{\alpha}_{c^\star g} = \sum_c \omega_{c|c^\star} \alpha_{cg}$ and the source-averaged gene prior in Appendix A.1.

The observed-response route uses untreated-control similarity to pool the target gene's responses to the query perturbation across source cell lines. For the memory-transfer route, the weights $\widetilde{\omega}_{c|c^\star,p}$ are obtained by normalizing $\kappa_{\mathrm{cell}}(u_{c^\star}, u_c)$ over the supporting source cell lines $\mathcal{C}_{\mathrm{src}}(p)$. These weights combine their fitted latent states H_{cp} and broad perturbation effects b_{cp} in Equation 4.

For a query perturbation p, let $\mathcal{N}_{c^\star}(g)$ contain the positively coexpressed neighbors of target gene g with available source responses to p, and let e_{gh} denote their edge weights. The response of neighbor h to p, averaged across supporting source cell lines using untreated-control similarity, is $\overline{X}^{\mathrm{src}}_{ph}$. The

coexpression route estimates the target gene's response as

$$\widehat{X}^{(\mathrm{ctx})}_{c^\star pg} = \sum_{h \in \mathcal{N}_{c^\star}(g)} \frac{e_{gh}}{\sum_{h' \in \mathcal{N}_{c^\star}(g)} e_{gh'}} \overline{X}^{\mathrm{src}}_{ph}. \tag{16}$$

The task feature vector combines this estimate with the effective number of supporting neighbors, the concentration of their weights, and response disagreement. The logistic model in Equation 6 maps these features to a DE or DIR probability.

A.5 Conditional transfer estimation and fusion

Transfer features. The transfer feature vector $\phi_r(q)$ combines the reference and route probabilities with their probability and logit differences, distances from 0.5, and whether they imply different classes. Retrieval records supply effective case or source counts, similarities to the query, and the entropy of normalized retrieval weights. For retrieved cross-perturbation cases, response disagreement is the fraction of total case weight assigned to the less-supported task label.

Memory routes also use the readout extrapolation scores in Appendix A.1. For perturbation memory routes, the feature vector includes the root-mean-square error of leave-one-out kernel ridge predictions of the fitted broad perturbation effects and latent states. The gain and reliability models in Section 2.3 use perturbation-balanced weights and the same standardized features, with ℓ_2 regularization for the reliability model.

Transfer weights and fusion. We combine predicted gain and reliability with adjustments for the amount of supporting data and departure from training conditions to obtain the transfer weight

$$\omega_{r,t}(q) = \frac{n_{\mathrm{eff},r}(q)}{n_{\mathrm{eff},r}(q) + \kappa_{\mathrm{sup}}} \, h_{\mathrm{dom},r}(q) \, \mathrm{Rel}_{r,t}(q) \, \sigma\left(\frac{\mathrm{Gain}_{r,t}(q)}{T_{\mathrm{gain}}}\right), \tag{17}$$

where σ is the logistic sigmoid and $n_{\mathrm{eff},r}(q)$ measures the effective number of supporting cases or source cell lines. We compute it as $(\sum_i a_i)^2 / \sum_i a_i^2$, where $a_i \geq 0$ is the retrieval weight of supporting case or source cell line i. For the coexpression route, these effective source counts are averaged over neighbors using normalized edge weights.

The factor $h_{\mathrm{dom},r}(q) = \exp[-2d_r(q)]$ reduces the transfer weight when features in $\phi_r(q)$ fall outside their training 1st–99th-percentile intervals. Here $d_r(q)$ is the maximum deviation beyond these intervals after each deviation is normalized by the corresponding interval width. Both generalization settings use support scale $\kappa_{\mathrm{sup}} = 3$ and gain temperature $T_{\mathrm{gain}} = 0.01$.

The operator $\mathrm{clip}_{r,t}$ in Equation 9 bounds logit differences symmetrically at the 99.5th percentile of their absolute values in each task–route pair's out-of-fold training records. The shared fusion model in Equation 10 is fitted to out-of-fold transfer-weighted corrections under binary log loss, with an ℓ_2 penalty weight of 0.001. Perturbations receive equal total training weight, distributed equally over their queries and then over each query's route subsets.

Cross-cell-line DIR calibration. For cross-cell-line DIR, the observed-response route provides a consensus signal when the target gene shows the same direction of change across all source cell lines in which it is differentially expressed under the query drug. For consensus class $s \in \{0, 1\}$ supported by m source cell lines, we estimate a Laplace-smoothed probability of up-regulation using records from inner folds that each hold out one source cell line. We use this estimate to calibrate the fused prediction as

$$\begin{aligned} \widehat{\pi}_{m,s} &= \frac{N^+_{m,s} + 1}{N_{m,s} + 2}, \\ \mathrm{logit}\, \widetilde{p}_{\mathrm{fuse,DIR}}(q; S) &= (1 - \gamma) \,\mathrm{logit}\, p_{\mathrm{fuse,DIR}}(q; S) + \gamma \,\mathrm{logit}\, \widehat{\pi}_{m,s}, \end{aligned} \tag{18}$$

where $N_{m,s}$ counts held-out inner-fold queries with support pattern (m, s), and $N^+_{m,s}$ counts their up-regulated outcomes. The mixing weight is fixed at $\gamma = 0.4$. The calibrated probability enters both utility estimation and final prediction. We cap m at the largest source count represented in the calibration data. Unseen (m, s) combinations use the pooled estimate for class s. Without a valid consensus estimate, the fused probability is unchanged.

Algorithm 1 Grouped cross-fitting for transfer and utility learning

Require: Labeled training responses $\mathcal{D}_{\text{mem}}$, task t, evidence resources, and grouped response-level and meta folds.
Ensure: Fitted reference, response, transfer, fusion, and utility models.
for each response-level fold F **do**
  Fit reference and response models on $\mathcal{D}_{\text{mem}} \setminus F$.
  Record $(p_{0,t}(q_j), p_{r,t}(q_j), \phi_r(q_j), y_j)$ for queries $q_j \in F$ and available routes r.
end for
Compute log-loss reductions $\Delta_{j,r,t}$ (Equation 7).
Cross-fit gain and reliability models on the meta folds, targeting $\Delta_{j,r,t}$ and $\mathbf{1}[\Delta_{j,r,t} > 0]$, respectively.
Compute $z_{r,t}(q_j)$ from held-out gain and reliability predictions (Equation 9).
Cross-fit fusion on the meta folds over route subsets.
Evaluate S_{core} and $S_{\text{core}} \cup \{r\}$ on held-out queries using the same fitted fusion model.
if DIR consensus calibration applies **then**
  Calibrate both predictions with the same calibrator fitted without the held-out source cell line.
end if
Compute utility targets $U_{j,r,t}$ for memory and coexpression routes (Equation 11).
Cross-fit utility models on the meta folds using targets $U_{j,r,t}$.
Refit reference and response models on $\mathcal{D}_{\text{mem}}$; refit the remaining models on out-of-fold records.
return The fitted models.

A.6 Two-level grouped cross-fitting

Grouped cross-fitting first produces reference and route predictions for training queries excluded from response-model fitting. These records are then divided into meta folds to learn transfer weights, fusion, and utility (Algorithm 1). Both levels group by perturbation identity in cross-perturbation prediction and by source cell line in cross-cell-line prediction.

The utility comparison holds the fusion parameters and calibration fixed between S_{core} and $S_{\text{core}} \cup \{r\}$, with the calibrator also excluding the held-out source cell line when applicable. After configuration selection, reference and response models are refitted on the training responses, and transfer, fusion, and utility models on their corresponding out-of-fold records.

A.7 Utility-guided evidence selection

Utility regressors use perturbation-balanced training weights and standardized features from Section 2.4, with route packet length included as an additional feature.

Selection maximizes the sum of predicted utilities within a budget of B tokens for $E(q)$. Let $L_{\text{core}}(q)$ denote the combined token length of the retained evidence and fused prediction, and $L_r(q)$ the token length of the evidence packet for route r. Let $\mathcal{R}_{\text{avail}}(q)$ contain available memory and coexpression routes with positive predicted utility. The feasible subsets in Equation 12 are

$$\mathcal{S}(q) = \Big\{ S_{\text{sup}} \subseteq \mathcal{R}_{\text{avail}}(q) \;\Big|\; L_{\text{core}}(q) + \sum\nolimits_{r \in S_{\text{sup}}} L_r(q) \leq B \Big\}. \tag{19}$$

The small number of routes permits exact enumeration of feasible subsets. When the budget is nonbinding, as in the reported cross-cell-line experiments, selection retains all available positive-utility routes.

B Evidence and LLM Reasoning

B.1 Structured evidence set

The evidence set $E(q)$ combines retrieved observations and KG facts with selected response estimates and the fused prediction (Table S2). Measured responses are distinguished from model estimates, and source identifiers link the LLM's citations to the supporting records.

Table S2: Structured evidence supplied to the frozen LLM.

Component	Contents
Query context	Cell line, perturbation, gene, task, and available assay context
Observed response cases	Expression effects or labels for the target gene from related training perturbations or source cell lines, with source identifiers
Retrieved KG facts	Query-relevant drug–target and pathway relations, with source identifiers and available action annotations
Selected response summaries	Response estimates from selected memory or coexpression routes, with supporting records and indicators of extrapolation beyond training conditions
Fused prediction	Predicted class, probability band, response routes used in fusion, and task-specific support status
Evidence support	Availability of supporting responses, disagreement among them, and missing evidence

Algorithm 2 Utility-guided evidence selection and frozen LLM reasoning

Require: Query q, KGs, training responses, fitted models, evidence budget B, frozen LLM with a fixed CoT prompt.
Ensure: A task decision or abstention with confidence, rationale, citations, and structured evidence $E(q)$.
Retrieve target-gene response cases and public KG facts; form E_{core} with query context.
Compute available memory and coexpression response estimates.
$S_{\mathrm{core}} \leftarrow$ observed-response route set (empty if unavailable).
Compute $p_{0,t}(q)$, $p_{r,t}(q)$, and $\phi_r(q)$ for available routes.
Compute transfer-weighted corrections $z_{r,t}(q)$ from predicted gain and reliability (Equation 9).
Predict $\widehat{U}_{r,t}(q)$ for adding each available memory or coexpression route to S_{core}.
Select $S^{\star}_{\mathrm{sup}}(q)$ to maximize total predicted utility over positive-utility routes within budget B (Equation 12).
$S \leftarrow S_{\mathrm{core}} \cup S^{\star}_{\mathrm{sup}}(q)$.
Compute $\widetilde{p}_{\mathrm{fuse},t}(q; S)$, with DIR consensus calibration when applicable.
Construct $E(q)$ from E_{core}, selected response packets, and the fused prediction.
if the fused prediction meets the task-specific support criteria **then**
    Run the LLM on $E(q)$, with the output class determined by the fused prediction.
else
    Run the LLM on $E(q)$ under the fallback and abstention rules.
end if
return The decision, confidence, rationale, citations, and $E(q)$.

B.2 Frozen LLM reasoning

Support is assessed on the selected routes S. For DE, at least one route must have a nonzero transfer-weighted logit correction. For DIR, the observed-response route provides directional support when predicted gain is positive, transfer weight is nonzero, and $\mathrm{Rel}_{r,t} \geq 0.60$. Memory and coexpression routes must also satisfy $|p_{r,t} - 0.5| \geq 0.20$. In cross-cell-line prediction, available source responses to the same drug for the same gene also provide directional support. These criteria are fixed before test evaluation.

When supported, the fused prediction determines the output class, while the frozen LLM supplies the biological rationale, confidence, and citations (Algorithm 2). Otherwise, the LLM considers supported individual route predictions, followed by observed response cases and other supported evidence from source cell lines or retrieved biological knowledge. When all active corrections are zero and neither a supported route nor an observed case is available, a prior-based fallback requires $|p_{\mathrm{fuse},t} - 0.5| \geq 0.20$. It is identified as prior-based, with confidence capped at 0.65. The LLM abstains when no supported decision is available. Task labels, cited identifiers, and the distinction between observed and derived evidence are validated; unresolved invalid outputs count as abstentions.

B.3 Reasoning prompt

The fixed CoT prompt (Wei et al., 2022) asks the LLM to check query entities and evidence coverage, compare supporting and opposing responses, and interpret them using retrieved biological knowledge (Wei et al., 2026). The rationale cites its evidence, reports disagreements, and distinguishes gene-prior

support, perturbation-specific predictions, and mechanistic interpretation. KG facts remain distinct from the LLM's inferences, and missing evidence is not interpreted as support for the negative class.

B.4 Recorded query examples

Two recorded QWEN3.5-9B examples illustrate how response predictions and KG facts inform the final output. Probability bands refer to the fused prediction; confidence is reported by the LLM. Held-out labels are used only for subsequent evaluation.

Mitoxantrone–CENPF in C32. For CENPF under Mitoxantrone (dihydrochloride) in C32, the retrieved cases show DE under Topotecan (hydrochloride) and non-DE under Indacaterol (maleate). Two-hop retrieval returns no public KG path to CENPF.

The observed-response route and all three memory routes predict DE. The supported fused prediction has a positive-class probability band of 0.80–1.00 and determines the DE output. The LLM reports confidence 0.85 and cites the fused prediction and all four response packets. The prediction matches the held-out label.

AT7519–FOS in held-out C32. For FOS under AT7519 in held-out C32, the evidence contains same-drug, same-gene responses from HOP62, HepG2/C3A, Hs 766T, and PANC-1, including an observed up-regulation label from PANC-1. The coexpression route uses 17 neighbors identified from untreated C32 controls, including EGR1 with decreased expression and KLF10 with increased expression in source responses. The public KG annotates AT7519 as a negative modulator of CDK1 and links CDK1 to FOS through an unsigned STRING association.

The observed-response, memory-transfer, and coexpression routes all predict up-regulation. The supported fused prediction has a positive-class probability band of 0.60–0.80. The LLM reports up-regulation with confidence 0.75, noting that the unsigned KG association does not establish response direction. The predicted direction matches the held-out label, and the output cites the fused prediction, the three response packets, and the public KG relation.

C Data and Evaluation Splits

C.1 Preprocessing and label construction

Response memory and evaluation use the same benchmark-matched 2,000 highly variable genes (HVGs). Single-cell counts are library-size normalized to 10^4 and transformed with $\log(1 + x)$. For each cell line and perturbation, we compute the difference in mean log-normalized expression between treated and control cells. These gene-level effects form the response-memory input X_{cpg}.

We retain the released GeneTAK labels after restricting queries to this gene universe. Benchmark DE-positive labels require a Wilcoxon rank-sum p-value below 0.05 after Benjamini–Hochberg correction and an absolute $\log_2$ fold change greater than 0.25 (Wei et al., 2026; Wilcoxon, 1945; Benjamini & Hochberg, 1995). The benchmark samples up to 200 negative genes per perturbation from genes with raw $p > 0.1$; those within the HVGs are retained without resampling. For DE-positive queries, DIR labels follow the signs of the benchmark log-fold changes. All methods use the same evaluation labels, matched to predictions after inference.

C.2 Cross-perturbation data splits

Perturbations are grouped by canonical active-moiety identity, combining formulation variants while keeping stereoisomers distinct. Entire groups are assigned to $\mathcal{D}_{\text{train}}$ (75%) or $\mathcal{D}_{\text{test}}$ (25%), so related formulations cannot span training and test sets.

Within $\mathcal{D}_{\text{train}}$, MMCA-CT uses $\mathcal{D}_{\text{mem}}$ for response-model fitting and grouped out-of-fold transfer learning, and $\mathcal{D}_{\text{sel}}$ for configuration selection. The reserved calibration partition $\mathcal{D}_{\text{cal}}$ is unused in the reported cross-perturbation configurations. These partitions contain 55%, 10%, and 10% of all perturbation groups, respectively, with $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{mem}} \dot{\cup} \mathcal{D}_{\text{sel}} \dot{\cup} \mathcal{D}_{\text{cal}}$. Baselines use the same $\mathcal{D}_{\text{train}}$ pool with their own fitting, model-selection, and calibration procedures.

Group assignment uses seed 20260711, with five response-level folds and five transfer-level meta folds within $\mathcal{D}_{\mathrm{mem}}$. Table S3 gives the resulting split by cell line and data role.

Table S3: Cross-perturbation split sizes in canonical active-moiety groups.

Cell line	**Fitting** $\mathcal{D}_{\mathrm{mem}}$	**Selection** $\mathcal{D}_{\mathrm{sel}}$	**Reserved** $\mathcal{D}_{\mathrm{cal}}$	**Training** $\mathcal{D}_{\mathrm{train}}$	**Test** $\mathcal{D}_{\mathrm{test}}$
C32	199	36	36	271	91
HepG2/C3A	199	36	36	271	91
HOP62	199	36	36	271	91
Hs 766T	199	36	36	271	91
PANC-1	199	36	36	271	91

C.3 Cross-cell-line data splits

Each of the five cell lines serves once as the held-out target $c^\star$, with the other four providing treated responses for fitting and calibration. All response-dependent components are refitted within each outer fold. Responses are pooled across doses for all methods; target inputs are limited to untreated controls, assay metadata, and pre-existing biological annotations.

Within the source pool, each of the four cell lines serves in turn as a pseudo-target, with the other three supplying observed responses for fitting and retrieval. The resulting out-of-fold records train the transfer, fusion, utility, and DIR calibration models (Appendix A.6), which remain fixed during prediction for the held-out target.

D Evaluation Metrics

Task labels and ranking scores. Let $\mathcal{A}$ denote the answered queries in an evaluation group, with ground-truth labels y_j, final binary predictions $\widehat{y}_j$, and positive-class scores $\widehat{p}_j$. The positive class is differential expression for DE and up-regulation for DIR. Classification metrics use $(y_j, \widehat{y}_j)$, whereas ranking metrics use $(y_j, \widehat{p}_j)$. Overall and macro-per-gene evaluation use the same per-query scores.

Expression-based baselines use their binary DE/DIR predictions as ranking scores, $\widehat{p}_j = \widehat{y}_j$. Prior vote uses the empirical positive-label frequency defined in Appendix E. For MMCA-CT, valid answers based on a supported fused prediction or the permitted prior-based fallback use the fusion probability, with cross-cell-line DIR calibration when applicable. Other valid LLM answers use the reported confidence for a positive prediction and one minus that confidence for a negative prediction. Other predictors without a valid continuous score use $\widehat{p}_j = \widehat{y}_j$. For a probability-only predictor, a score of at least 0.5 gives a positive prediction.

Classification metrics. Accuracy is the fraction of all N queries answered correctly; abstentions and invalid outputs count as incorrect. Precision, recall, and F1 follow the confusion-matrix definitions (Wei et al., 2026), with unanswered positive queries counted as false negatives and no unanswered query counted as a predicted positive. Precision, recall, or F1 is zero when its denominator is zero.

Coverage and abstention. Coverage is $|\mathcal{A}|/N$, and the abstention rate is one minus coverage. Tables S6 and S9 report coverage alongside all-query accuracy.

Overall ranking metrics. We pool answered queries within each cell line and compute areas under the ROC and precision–recall curves from the scores defined above (Wei et al., 2026). AUROC counts positive–negative score ties as one half. Both metrics are reported only when the answered queries contain both task labels.

Macro-per-gene metrics. Within each cell line and task, we group queries by gene and compute metrics across perturbations within each group (Yuan et al., 2026). Let $\mathcal{G}_t^\star$ contain genes whose answered queries include both task labels. For metric M, let M_g denote its value for gene g, computed using the rules below. The macro-per-gene metric is

$$M_{\text{macro-gene}} = \frac{1}{|\mathcal{G}_t^\star|} \sum_{g \in \mathcal{G}_t^\star} M_g. \tag{20}$$

Classification metrics include all queries for each eligible gene, with abstentions handled as above; AUROC and AUPRC use its valid prediction scores $\widehat{p}_j$. Per-gene AUPRC uses average precision, with tied scores grouped together (Yuan et al., 2026). AUROC, AUPRC, accuracy, and F1 use the same eligible-gene set, determined separately for each method, with equal weight per gene. DIR follows this procedure for DE-positive queries using up- and down-regulation labels.

Aggregation and uncertainty. Results are summarized by the unweighted mean and sample standard deviation across cell lines. Each cell line contributes one overall value or one macro-per-gene average, corresponding to one held-out target in cross-cell-line evaluation. Figure 4 uses 95% t-based intervals across cell lines. Figure 3(b) uses the bootstrap procedure defined below.

Case-study agreement. For each cell line and inhibitor, directional agreement is the fraction of DIR queries in the response panel predicted correctly, counting abstentions as incorrect. Figure 3(b) instead uses all DIR queries for each inhibitor, matched between Prior vote and MMCA-CT. Paired bootstrap resampling draws these query identifiers with replacement, separately for each inhibitor; intervals use the 2.5th and 97.5th percentiles. Appendix F.4 defines the vote and response panel.

E Baseline and Language-Model Configurations

For expression-based baselines, predicted profiles are compared with matched controls. The Wilcoxon-based label-construction procedure in Appendix C converts these profiles to binary DE calls (Wei et al., 2026). DIR follows the sign of the predicted perturbation–control effect on DE-positive queries. Scoring conventions follow Appendix D.

In cross-perturbation evaluation, Prior vote uses the unweighted mean of valid training labels for each cell line and gene in the complete $\mathcal{D}_{\text{train}}$, restricting DIR votes to DE-positive queries. This positive-label frequency is constant across test perturbations for each cell–gene pair. Values of at least 0.5 predict the positive class; genes without valid training labels remain unanswered.

GAT, scVI, and STATE retain their respective fitting and model-selection procedures within the common perturbation-disjoint split. For cross-cell-line evaluation, scVI, scGen, CRISP, and STATE instead use the four source cell lines in each outer fold. Their target inputs follow Appendix C.3, while response-dependent fitting uses source data only.

VCWorld uses the same QWEN3.5 backbone as MMCA-CT at each model size, with its own evidence construction. All language-model backbones remain frozen. Backbone comparisons retain each agent's evidence construction and reasoning prompt.

F Additional Results and Analyses

F.1 Cross-perturbation generalization

Table S4 extends the main comparison to all backbones. At both QWEN3.5 sizes, MMCA-CT exceeds VCWorld in mean precision, recall, and both overall and macro-per-gene F1 for DE and DIR (Table S5). All three MMCA-CT backbones answer nearly all DE queries, while DIR coverage is lower, particularly with Llama3.1-8B (Table S6).

F.2 Cross-cell-line generalization

Tables S7 and S8 report performance across all backbones and additional classification metrics, respectively. Across all three backbones, MMCA-CT has higher mean overall and macro-per-gene F1 than the four baselines on both tasks. Both QWEN3.5 backbones answer nearly all target queries (Table S9). Llama3.1-8B has lower coverage and greater between-cell-line variation for DIR. All four expression-based baselines answer the full query set.

F.3 Transfer-reliability diagnostic

We compare learned transfer reliability with retrieval weights and similarity scores for identifying route predictions that improve on the gene-prior reference. The analysis uses held-out query–route

Table S4: Cross-perturbation performance across model configurations.

Task	Method	Overall			Macro-per-gene		
		Accuracy ↑	AUROC ↑	AUPRC ↑	Accuracy ↑	AUROC ↑	AUPRC ↑
DE	Prior vote	0.903±0.009	0.943±0.007	0.931±0.011	0.783±0.013	0.500±0.000	0.628±0.049
	GAT	0.597±0.032	0.636±0.042	0.612±0.082	0.557±0.010	0.556±0.016	0.658±0.044
	scVI	0.554±0.053	0.582±0.063	0.560±0.105	0.536±0.011	0.514±0.011	0.633±0.047
	STATE	0.905±0.013	0.905±0.008	0.930±0.010	0.774±0.013	0.498±0.004	0.627±0.049
	VCWORLD (QWEN3.5-9B)	0.686±0.010	0.687±0.012	0.613±0.042	0.608±0.015	0.497±0.008	0.641±0.047
	VCWORLD (QWEN3.5-27B)	0.691±0.011	0.711±0.013	0.628±0.045	0.619±0.011	0.526±0.009	0.652±0.046
	MMCA-CT (Llama3.1-8B)	0.921±0.012	**0.973±0.008**	0.962±0.011	0.792±0.013	0.570±0.042	**0.750±0.013**
	MMCA-CT (QWEN3.5-9B)	0.933±0.011	0.953±0.008	0.952±0.012	0.809±0.013	0.577±0.042	0.739±0.014
	MMCA-CT (QWEN3.5-27B)	**0.952±0.014**	0.972±0.010	**0.963±0.013**	**0.821±0.011**	**0.581±0.043**	**0.750±0.013**
DIR	Prior vote	0.747±0.026	0.843±0.020	0.817±0.042	0.716±0.018	0.500±0.000	0.476±0.077
	GAT	0.602±0.022	0.640±0.027	0.590±0.077	0.544±0.014	0.560±0.016	0.510±0.070
	scVI	0.538±0.009	0.538±0.018	0.660±0.087	0.502±0.002	0.504±0.005	0.478±0.075
	STATE	0.775±0.023	0.760±0.019	0.782±0.045	0.698±0.018	0.500±0.002	0.476±0.077
	VCWORLD (QWEN3.5-9B)	0.537±0.018	0.537±0.008	0.481±0.096	0.523±0.017	0.511±0.012	0.485±0.070
	VCWORLD (QWEN3.5-27B)	0.553±0.009	0.569±0.010	0.512±0.095	0.536±0.009	0.528±0.012	0.514±0.072
	MMCA-CT (Llama3.1-8B)	0.669±0.044	**0.881±0.013**	**0.858±0.045**	0.628±0.033	0.572±0.035	**0.581±0.078**
	MMCA-CT (QWEN3.5-9B)	0.784±0.026	0.856±0.016	0.829±0.047	0.739±0.010	0.562±0.036	0.568±0.079
	MMCA-CT (QWEN3.5-27B)	**0.795±0.023**	0.869±0.019	0.845±0.045	**0.751±0.009**	**0.575±0.037**	0.579±0.079

Table S5: Additional cross-perturbation metrics.

Task	Method	Overall			Macro-per-gene
		Precision ↑	Recall ↑	F1 ↑	F1 ↑
DE	Prior vote	0.905±0.014	0.913±0.011	0.909±0.009	0.575±0.063
	GAT	0.580±0.073	0.620±0.032	0.599±0.053	0.561±0.026
	scVI	0.535±0.087	0.654±0.040	0.588±0.067	0.566±0.036
	STATE	0.909±0.020	0.900±0.012	0.905±0.016	0.535±0.061
	VCWORLD (QWEN3.5-9B)	0.651±0.038	0.726±0.018	0.686±0.021	0.589±0.046
	VCWORLD (QWEN3.5-27B)	0.676±0.041	0.765±0.022	0.718±0.028	0.604±0.050
	MMCA-CT (Llama3.1-8B)	0.919±0.012	0.933±0.016	0.926±0.012	0.559±0.059
	MMCA-CT (QWEN3.5-9B)	0.919±0.012	0.943±0.015	0.931±0.012	0.612±0.062
	MMCA-CT (QWEN3.5-27B)	**0.932±0.011**	**0.955±0.017**	**0.943±0.013**	**0.633±0.059**
DIR	Prior vote	0.759±0.060	0.667±0.068	0.708±0.047	0.367±0.089
	GAT	0.562±0.084	0.590±0.020	0.573±0.045	0.453±0.045
	scVI	0.536±0.091	0.570±0.010	0.550±0.049	0.449±0.051
	STATE	0.757±0.038	0.672±0.052	0.712±0.045	0.335±0.083
	VCWORLD (QWEN3.5-9B)	0.496±0.093	0.479±0.020	0.486±0.045	0.401±0.035
	VCWORLD (QWEN3.5-27B)	0.523±0.092	0.562±0.016	0.540±0.049	0.436±0.045
	MMCA-CT (Llama3.1-8B)	**0.825±0.047**	0.578±0.034	0.679±0.031	0.369±0.072
	MMCA-CT (QWEN3.5-9B)	0.773±0.052	0.684±0.049	0.726±0.041	0.462±0.089
	MMCA-CT (QWEN3.5-27B)	0.811±0.047	**0.698±0.045**	**0.749±0.042**	**0.474±0.090**

records from both QWEN3.5-9B generalization evaluations. Route and reference probabilities, reliability, and comparison scores are recorded before held-out labels are used to compute the log-loss reduction Δ in Equation 7.

Cross-perturbation comparisons use retrieval weights for the observed-response route and kernel similarities for memory routes (Appendix A.3). Cross-cell-line comparisons use untreated-control similarities for the observed-response and memory-transfer routes, and coexpression edge weights for the coexpression route (Appendix A.4).

Within each cell line, task, and route, query–route pairs are divided into reliability quintiles Q1–Q5, retaining groups with five nonempty bins. The positive-gain rate in each bin is the fraction with $\Delta > 0$. Rates are averaged equally over routes within each cell line and then over the five cell lines; 95% confidence bands use a t distribution with four degrees of freedom. For the direct comparison, AUROC treats positive and nonpositive gain as the two classes. It is computed for each cell line and route with both classes present, then averaged equally within each setting and task.

Table S6: Cross-perturbation coverage and accuracy.

Task	Method	Coverage ↑	Abstention rate ↓	Accuracy ↑
DE	Prior vote	1.000±0.000	0.000±0.000	0.903±0.009
	GAT	1.000±0.000	0.000±0.000	0.597±0.032
	scVI	1.000±0.000	0.000±0.000	0.554±0.053
	STATE	1.000±0.000	0.000±0.000	0.905±0.013
	VCWORLD (QWEN3.5-9B)	0.940±0.004	0.060±0.004	0.686±0.010
	VCWORLD (QWEN3.5-27B)	0.997±0.000	0.003±0.000	0.691±0.011
	MMCA-CT (Llama3.1-8B)	0.988±0.002	0.012±0.002	0.921±0.012
	MMCA-CT (QWEN3.5-9B)	0.993±0.001	0.007±0.001	0.933±0.011
	MMCA-CT (QWEN3.5-27B)	0.990±0.001	0.010±0.001	**0.952±0.014**
DIR	Prior vote	0.998±0.003	0.002±0.003	0.747±0.026
	GAT	1.000±0.000	0.000±0.000	0.602±0.022
	scVI	1.000±0.000	0.000±0.000	0.538±0.009
	STATE	1.000±0.000	0.000±0.000	0.775±0.023
	VCWORLD (QWEN3.5-9B)	0.931±0.009	0.069±0.009	0.537±0.018
	VCWORLD (QWEN3.5-27B)	0.980±0.007	0.020±0.007	0.553±0.009
	MMCA-CT (Llama3.1-8B)	0.822±0.037	0.178±0.037	0.669±0.044
	MMCA-CT (QWEN3.5-9B)	0.899±0.026	0.101±0.026	0.784±0.026
	MMCA-CT (QWEN3.5-27B)	0.902±0.026	0.098±0.026	**0.795±0.023**

Table S7: Cross-cell-line performance across model configurations.

		Overall			Macro-per-gene		
Task	Method	Accuracy ↑	AUROC ↑	AUPRC ↑	Accuracy ↑	AUROC ↑	AUPRC ↑
DE	scVI	0.500±0.027	0.500±0.038	0.500±0.032	0.490±0.023	0.500±0.006	0.521±0.027
	scGen	0.495±0.033	0.495±0.051	0.597±0.043	0.480±0.031	0.490±0.008	0.515±0.037
	CRISP	0.621±0.037	0.621±0.063	0.711±0.053	0.579±0.039	0.578±0.010	**0.665±0.046**
	STATE	0.586±0.035	0.586±0.031	0.672±0.059	0.484±0.043	0.549±0.011	0.630±0.050
	MMCA-CT (Llama3.1-8B)	0.591±0.035	**0.806±0.044**	**0.831±0.045**	0.579±0.037	0.520±0.008	0.549±0.045
	MMCA-CT (QWEN3.5-9B)	0.731±0.040	0.774±0.063	0.794±0.053	0.633±0.039	0.576±0.010	0.644±0.046
	MMCA-CT (QWEN3.5-27B)	**0.740±0.040**	0.791±0.055	0.799±0.054	**0.645±0.039**	**0.585±0.011**	0.658±0.045
DIR	scVI	0.504±0.028	0.506±0.023	0.505±0.023	0.500±0.018	0.500±0.026	0.509±0.023
	scGen	0.578±0.044	0.578±0.036	0.700±0.037	0.560±0.029	0.580±0.041	0.575±0.036
	CRISP	0.644±0.042	0.644±0.045	0.729±0.046	0.618±0.036	0.763±0.051	0.772±0.045
	STATE	0.643±0.054	0.643±0.045	0.728±0.046	0.636±0.036	0.722±0.051	0.717±0.045
	MMCA-CT (Llama3.1-8B)	0.600±0.212	**0.896±0.056**	**0.893±0.050**	0.624±0.180	**0.889±0.057**	**0.874±0.043**
	MMCA-CT (QWEN3.5-9B)	0.792±0.048	0.867±0.045	0.860±0.046	0.787±0.036	0.870±0.051	0.849±0.045
	MMCA-CT (QWEN3.5-27B)	**0.798±0.049**	0.882±0.046	0.869±0.045	**0.796±0.035**	0.878±0.052	0.862±0.044

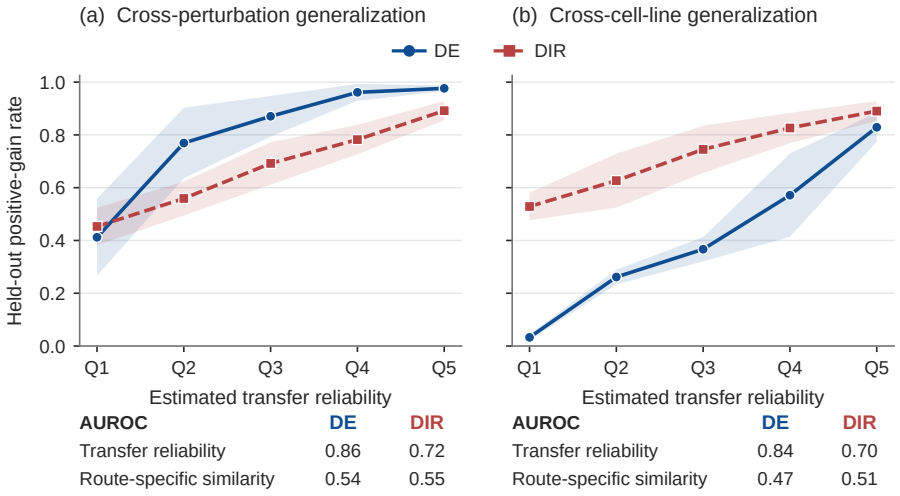

Figure 4: Transfer reliability and held-out predictive benefit.

Table S8: Additional cross-cell-line metrics.

		Overall			Macro-per-gene
Task	**Method**	**Precision ↑**	**Recall ↑**	**F1 ↑**	**F1 ↑**
DE	scVI	0.500±0.031	0.420±0.031	0.457±0.031	0.230±0.018
	scGen	0.494±0.038	0.400±0.043	0.442±0.039	0.210±0.024
	CRISP	0.632±0.035	0.579±0.060	0.604±0.044	0.439±0.030
	STATE	0.615±0.039	0.459±0.061	0.526±0.049	0.200±0.033
	MMCA-CT (Llama3.1-8B)	0.727±0.024	0.707±0.074	0.716±0.049	0.460±0.033
	MMCA-CT (QWEN3.5-9B)	0.727±0.024	0.735±0.077	0.730±0.049	0.458±0.030
	MMCA-CT (QWEN3.5-27B)	**0.735±0.025**	**0.748±0.078**	**0.741±0.050**	**0.467±0.031**
DIR	scVI	0.504±0.027	0.504±0.030	0.504±0.028	0.420±0.035
	scGen	0.596±0.043	0.589±0.047	0.592±0.044	0.500±0.056
	CRISP	0.729±0.066	0.459±0.042	0.563±0.051	0.328±0.070
	STATE	0.660±0.061	0.591±0.051	0.624±0.055	0.506±0.070
	MMCA-CT (Llama3.1-8B)	**0.846±0.072**	0.549±0.242	0.634±0.189	0.554±0.132
	MMCA-CT (QWEN3.5-9B)	0.802±0.051	0.774±0.061	0.788±0.052	0.667±0.070
	MMCA-CT (QWEN3.5-27B)	0.811±0.052	**0.787±0.061**	**0.799±0.053**	**0.679±0.068**

Table S9: Cross-cell-line coverage and accuracy.

Task	**Method**	**Coverage ↑**	**Abstention rate ↓**	**Accuracy ↑**
DE	scVI	1.000±0.000	0.000±0.000	0.500±0.027
	scGen	1.000±0.000	0.000±0.000	0.495±0.033
	CRISP	1.000±0.000	0.000±0.000	0.621±0.037
	STATE	1.000±0.000	0.000±0.000	0.586±0.035
	MMCA-CT (Llama3.1-8B)	0.799±0.019	0.201±0.019	0.591±0.035
	MMCA-CT (QWEN3.5-9B)	0.999±0.000	0.001±0.000	0.731±0.040
	MMCA-CT (QWEN3.5-27B)	0.986±0.001	0.014±0.001	**0.740±0.040**
DIR	scVI	1.000±0.000	0.000±0.000	0.504±0.028
	scGen	1.000±0.000	0.000±0.000	0.578±0.044
	CRISP	1.000±0.000	0.000±0.000	0.644±0.042
	STATE	1.000±0.000	0.000±0.000	0.643±0.054
	MMCA-CT (Llama3.1-8B)	0.733±0.286	0.267±0.286	0.600±0.212
	MMCA-CT (QWEN3.5-9B)	0.998±0.001	0.002±0.001	0.792±0.048
	MMCA-CT (QWEN3.5-27B)	0.985±0.003	0.015±0.003	**0.798±0.049**

Positive-gain rates increase across reliability quintiles in both settings. For DE, learned reliability achieves mean AUROC of 0.86 compared with 0.54 for the comparison scores in cross-perturbation evaluation, and 0.84 compared with 0.47 in cross-cell-line evaluation. Learned reliability also achieves higher AUROC for DIR in both settings (Figure 4).

F.4 Biological analysis of response predictions

Process-level evaluation. The process-level analysis groups held-out QWEN3.5-9B DIR predictions using all 50 MSigDB Hallmark gene sets (Liberzon et al., 2015). In cross-cell-line evaluation, Prior vote uses the empirical frequency of up-regulation among available same-gene DIR labels from the other four cell lines, predicting up-regulation when this frequency is at least 0.5. The comparison uses identical drug–gene queries with an available prior and counts abstentions as incorrect. Each heatmap entry gives the accuracy difference in percentage points for one gene set and target cell line.

An entry requires at least 100 paired queries spanning 10 distinct genes and 10 distinct drugs. Figure 3(a) includes all 29 gene sets with at least one eligible cell line and masks the remaining entries. The other 21 sets have no eligible cell line. Because gene sets overlap, entries are not independent replicates.

Cross-cell-line gains remain predominantly positive when the minimum counts for queries, genes, and drugs are changed to $(50, 5, 5)$ or $(200, 20, 20)$, and when the fitted gene-prior reference replaces Prior vote.

Additional process-level comparisons. Applying the same analysis to cross-perturbation predictions, with Prior vote computed from each cell line's training responses, gives lower accuracy for MMCA-CT in all 128 eligible entries. In cross-cell-line evaluation, we also compare with voting over source responses for the same drug and gene. Against this drug-specific comparator, MMCA-CT has 32 positive, 5 tied, and 88 negative accuracy differences across 125 eligible entries. The available-query subset differs from that used for Prior vote.

BRAF inhibitor case study. The case study uses a separate holdout of dabrafenib, vemurafenib, encorafenib, and regorafenib. Treated responses to all four compounds are excluded from fitting and retrieval; evaluation focuses on dabrafenib and vemurafenib. Other compounds retain their training-partition assignments (Appendix C). Separate predictors are fitted for the five cell lines under this common holdout, using molecular structure, target, pathway, and cell-line driver annotations as response-independent inputs.

Cell-line identities and driver annotations are obtained from Cellosaurus (Bairoch, 2018). C32 (RRID:CVCL_1097) is annotated with BRAF V600E. Independent culture experiments report reduced C32 viability following dabrafenib treatment (Tovar-Parra & Zammit-Mangion, 2024).

Figure 3(b) evaluates all DIR queries for each inhibitor in C32. Here, Prior vote is the unweighted up-regulation frequency among available same-gene DIR labels in the case-specific $\mathcal{D}_{\mathrm{mem}}$, with a decision threshold of 0.5. Paired accuracy gains are 0.123 (95% CI 0.084–0.166) for dabrafenib and 0.156 (95% CI 0.111–0.204) for vemurafenib, using the bootstrap procedure in Appendix D.

The RAS–RAF–MAPK response panel is the union of MSigDB Hallmark KRAS signaling up/down sets and Reactome RAF/MAP kinase cascade participants (Liberzon et al., 2015; Milacic et al., 2024). For each inhibitor and cell line, the panel includes all DIR queries whose genes belong to these sets, with DIR labels determined by the signs of the corresponding held-out benchmark log-fold changes. Table S10 reports directional agreement across cell lines using the definitions in Appendix D.

Figure 3(c) shows observed and predicted directions for the 21 dabrafenib and 23 vemurafenib queries in C32. Under both inhibitors, MMCA-CT recovers the observed down-regulation of DUSP6 and FYN and up-regulation of DLG2.

Table S10: Directional agreement on the RAS–RAF–MAPK response panel. Parentheses show correct/total drug–gene queries.

Cell line	Dabrafenib	Vemurafenib
C32	0.714 (15/21)	1.000 (23/23)
HOP62	0.500 (11/22)	0.906 (29/32)
HepG2/C3A	0.450 (9/20)	1.000 (10/10)
Hs 766T	0.650 (13/20)	0.960 (24/25)
PANC-1	0.400 (4/10)	0.857 (18/21)

F.5 Component ablations

The configurations in Section 3.6 share training partitions, the QWEN3.5-9B backbone, query context, retrieved KG facts, and observed target-gene response cases. Components are added successively:

A: Factorized response memory. All available memory-route response estimates are supplied as evidence to the frozen LLM, without transfer weighting or a fused prediction.

B: Conditional transfer estimation and fusion. Transfer weights and fusion combine the observed-response prediction with all available memory-route predictions. The corresponding evidence packets remain available to the LLM, and a supported fused prediction determines the output class (Appendix B.2).

C: Utility-guided evidence selection. Memory routes are selected according to their predicted utility when added to the observed-response route. The selected routes are then combined with the observed-response route for fusion, and their evidence packets are supplied to the LLM.

Relative to the factorized response memory configuration, adding conditional transfer estimation and fusion improves precision, recall, and both overall and macro-per-gene F1 on both tasks. Adding utility-guided evidence selection yields further gains on all four measures for both tasks (Table S11). In Tables 3 and S11, the Base evidence configuration corresponds to VCWORLD, whereas the complete configuration corresponds to MMCA-CT with QWEN3.5-9B in Section 3.2.

Table S11: Additional component-ablation metrics (QWEN3.5-9B, cross-perturbation).

		Overall			Macro-per-gene
Task	**Configuration**	**Precision ↑**	**Recall ↑**	**F1 ↑**	**F1 ↑**
DE	Base evidence	0.651±0.038	0.726±0.018	0.686±0.021	0.589±0.046
	+ A	0.884±0.012	0.911±0.023	0.897±0.016	0.562±0.089
	+ B	0.905±0.012	0.925±0.015	0.915±0.012	0.601±0.062
	+ C	**0.919±0.012**	**0.943±0.015**	**0.931±0.012**	**0.612±0.062**
DIR	Base evidence	0.496±0.093	0.479±0.020	0.486±0.045	0.401±0.035
	+ A	0.743±0.065	0.633±0.051	0.684±0.047	0.432±0.082
	+ B	0.761±0.052	0.658±0.049	0.706±0.041	0.447±0.089
	+ C	**0.773±0.052**	**0.684±0.049**	**0.726±0.041**	**0.462±0.089**

A: factorized response memory; B: conditional transfer estimation and fusion; C: utility-guided evidence selection.