WHEN EVIDENCE DISAGREES: SHAPLEY-SUPERVISED EVIDENCE SELECTION FOR PERTURBATION RESPONSE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting gene expression responses to perturbations is essential for elucidating gene function and guiding experimental design. Recent methods draw on paths retrieved from biological knowledge graphs (KGs) and measurements from related experiments to predict unobserved responses. However, signed KG paths can imply opposing effects, while experimental measurements can report different responses across conditions. The usefulness of each path or measurement can also depend on the perturbation, target gene, cell line, other evidence, and the downstream predictor that uses it. To address this challenge, we introduce PERTSELECT, a predictor-specific evidence selector that learns which retrieved KG paths and measurements to provide to a given predictor. During training, we estimate predictor-specific Shapley values by averaging how adding each evidence candidate to different subsets changes the predictor's probability of the correct response. Its three variants, PERTSELECT-JOINT, PERTSELECT-INDEPENDENT and PERTSELECT-KIND, use ModernBERT to learn contribution scores and retain evidence with positive scores at inference. We evaluate PERTSELECT with two predictors on CRISPR interference (CRISPRi) Perturb-seq data from four cell lines: Qwen3-8B, which reads KG paths from OmniPath and INDRA/CoGEx together with related experimental measurements, and a voting predictor that aggregates observed responses of the queried target gene in related perturbation experiments. Across 35,170 held-out queries in four evaluation settings, PERTSELECT-JOINT improves macro-F1 over providing all retrieved evidence by 20.3–31.5 percentage points for Qwen and 8.6–14.4 points for voting. It also achieves higher macro-F1 than embedding-similarity ranking, LLM-based pruning and a learned filter in the style of FILCO in every setting for both predictors. For Qwen, most of its gain comes from deciding how many KG paths and measurements to provide, which removes a strong bias toward predicting down-regulation.

1 INTRODUCTION

Predicting gene expression responses to perturbations is fundamental to understanding gene function and prioritizing experiments. Direct measurements of these responses, however, do not cover all perturbation–cellular context combinations of interest, motivating computational approaches that use external biological evidence to predict unmeasured responses (Wu et al., 2025; Kim et al., 2026; Yuan et al., 2026). Such evidence includes paths in biological knowledge graphs (KGs) connecting perturbed genes to downstream targets (Wenkel et al., 2026; Kim et al., 2026) and measured responses from related perturbation experiments (Wu et al., 2025; Yuan et al., 2026). These sources are often recorded in biological contexts different from the queried condition. Signed KG paths between the same two genes can imply opposing effects, and experimental measurements can report different responses across perturbations or cell types.

Recent work on evidence-grounded perturbation prediction uses LLMs to synthesize retrieved biological knowledge into reports and structured explanations, and verifies the generated explanations (Jang et al., 2026). Complementary approaches address inconsistencies between data-driven predictions and biological knowledge through adaptive alignment (Xiang & Ai, 2026). These methods pass a retrieved evidence set to a predictor without learning which parts of it help that predictor.

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.

Whether a candidate helps, however, depends on how the predictor uses it. An LLM can be biased by particular kinds of evidence, whereas a voting rule over measured responses is affected only by measurements of the queried target gene. We therefore ask: *How can we learn to select retrieved KG paths and experimental measurements to improve perturbation response prediction by a given downstream predictor?*

We introduce PERTSELECT, a predictor-specific evidence selector for KG paths and experimental measurements. For each training query, we estimate Shapley values (Shapley, 1953) that measure how adding an evidence candidate to different evidence subsets changes the predictor's probability of the correct response. Our three variants use a ModernBERT encoder to predict these contributions, differing in how they encode candidates and whether they select individual candidates or entire evidence kinds. At inference, PERTSELECT retains candidates with positive predicted contribution and applies the downstream predictor, without access to the query response.

We evaluate PERTSELECT on CRISPRi Perturb-seq data from Replogle et al. (2022) and Nadig et al. (2025), covering four cell lines and 35,170 held-out queries. The four evaluation settings combine seen or unseen perturbations with training or held-out cell lines. Retrieved evidence consists of KG paths from OmniPath and INDRA/CoGEx and measurements from other training experiments. We train separate selectors for two fixed predictors, Qwen3-8B and a CORE-style weighted vote over measured responses (Yuan et al., 2026). PERTSELECT-JOINT achieves higher macro-F1 than all-evidence prediction and all selection baselines in every setting for both predictors. With Qwen, all-evidence prediction labels 84.9% of test queries as down-regulated, and PERTSELECT-JOINT mainly corrects this bias by deciding how many paths and measurements to retain. We further compare the three variants and training targets that measure a candidate's effect in a single evidence context.

Our main contributions are:

- We formulate predictor-specific evidence selection for perturbation response prediction, with the goal of selecting retrieved KG paths and experimental measurements that improve a downstream predictor's predictions.
- We introduce PERTSELECT, a predictor-specific evidence selection framework with three variants trained with Shapley supervision: PERTSELECT-JOINT, PERTSELECT-INDEPENDENT and PERTSELECT-KIND.
- Across four evaluation settings on CRISPRi data from four cell lines, we show that PERTSELECT-JOINT improves both an LLM predictor and a measurement-voting predictor over all-evidence prediction and selection baselines, and analyze the roles of supervision, selector architecture and generalization setting.

2 RELATED WORK

Perturbation response prediction. Computational models for perturbation response prediction have incorporated cellular context, gene relationships, and biological knowledge to predict transcriptional changes under unseen perturbations (Lotfollahi et al., 2023; Roohani et al., 2024; Adduri et al., 2026; Wenkel et al., 2026). CPA models perturbation effects in a latent representation conditioned on cellular covariates, while GEARS incorporates gene relationships to generalize to unseen perturbations. More recent models such as State and TxPert further scale perturbation modeling or incorporate biological knowledge into response prediction (Adduri et al., 2026; Wenkel et al., 2026). These methods primarily learn predictors of perturbation responses rather than selecting external evidence provided at prediction time.

Evidence-grounded perturbation prediction. A complementary line of work explicitly provides biological evidence to support perturbation prediction. PerturbQA and CORE retrieve outcomes from related perturbation experiments, while PertReason conditions prediction on regulatory pathways retrieved from biological knowledge graphs (Wu et al., 2025; Yuan et al., 2026; Kim et al., 2026). PT-RAG jointly learns perturbation retrieval with a differentiable response generator (Di Francesco et al., 2026), and ALIGNED addresses inconsistencies between data-driven and knowledge-based perturbation predictions (Xiang & Ai, 2026). These approaches motivate careful

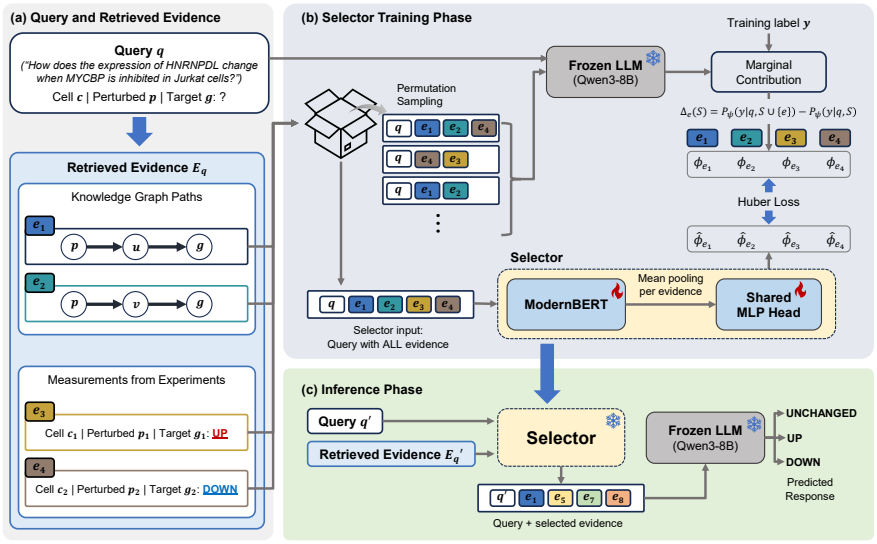

Figure 1: **PERTSELECT: Shapley-supervised evidence selection for perturbation response prediction. (a)** Each prediction specifies a cell line, perturbed gene, and target gene, together with paths retrieved from a biological knowledge graph and measurements from related experiments. **(b)** To construct training targets, we sample evidence permutations and measure how adding an evidence candidate to different evidence subsets changes the Qwen3-8B predictor's probability of the correct response. Averaging these marginal effects yields an estimated Shapley contribution for each candidate. **(c)** PERTSELECT-JOINT jointly encodes the prediction context and available evidence to predict these contributions. At inference, evidence with positive predicted contributions is selected and provided to the same downstream predictor. Response labels and Shapley estimation are not required at inference.

use of heterogeneous biological evidence. We study a complementary setting in which a candidate pool has already been retrieved and the goal is to determine which KG paths and experimental measurements should be provided to a particular downstream predictor.

Evidence retrieval and selection. In retrieval-augmented generation and other evidence-grounded settings, retrieved context is often filtered or selected before downstream use. FILCO learns context filters from labels derived using lexical and information-theoretic measures (Wang et al., 2023), while SetR selects passages jointly according to the information requirements of a question rather than ranking each passage independently (Lee et al., 2025). Large language models (LLMs) can also participate directly in controlling retrieved evidence: Self-RAG learns when retrieval is needed and evaluates the relevance of retrieved passages during generation (Asai et al., 2024). We study the selection of KG paths and experimental measurements for perturbation response prediction, using their measured contributions to the downstream predictor's correct-response probability across evidence subsets.

3 PROBLEM FORMULATION

A query $q = (c, p, g)$ specifies a cell line c, a perturbed gene p, and a target gene g, with response $y \in \mathcal{Y} = \{\text{unchanged}, \text{up}, \text{down}\}$. For each query q, we assume access to a retrieved evidence collection

$$E_q = E_q^{\text{KG}} \cup E_q^{\text{M}},$$

where E_q^{KG} contains KG paths connecting p and g, and E_q^{M} contains measured responses from other perturbation experiments. The retrieved candidates may provide conflicting signals, and their usefulness to the downstream predictor may depend on which candidates are provided together. Given

a downstream predictor ψ, let $E_q^\psi \subseteq E_q$ denote the candidate pool constructed for that predictor, for example by keeping only measurements the predictor can use and at most five candidates of each kind (Section 5.1). The two predictors in our experiments therefore receive different measurement pools. Our task is to learn a predictor-specific evidence selector

$$S = S_{\theta_\psi}(q, E_q^\psi) \subseteq E_q^\psi$$

that determines which retrieved evidence to provide to the predictor. The selected set may contain none, some, or all of the available candidates. The predictor then produces $\widehat{y} = \psi(q, S)$. Our goal is to select evidence that improves perturbation response prediction for the given downstream predictor.

4 PERTSELECT

4.1 CONTRIBUTION ESTIMATION

For a training query q with response y, we define the utility of an evidence subset S as the probability that the downstream predictor assigns to the correct response:

$$U_\psi(q, S, y) = P_\psi(y \mid q, S). \tag{1}$$

For a candidate e and an evidence subset $S \subseteq E_q^\psi \setminus \{e\}$, we define its marginal contribution as

$$\Delta_e(S) = U_\psi(q, S \cup \{e\}, y) - U_\psi(q, S, y). \tag{2}$$

The contribution of a candidate can depend on the other evidence: it may duplicate information already present or become useful only alongside complementary evidence. Singleton scoring tests its effect on an empty pool, while leave-one-out (LOO) scoring tests its effect when all other candidates are present. We use the Shapley value to average its contribution across different evidence contexts rather than relying on either single context (Shapley, 1953):

$$\phi_{q,e}^{(\psi)} = \mathbb{E}_\pi \left[\Delta_e(\mathrm{Pred}_\pi(e))\right], \qquad \widehat{\phi}_{q,e}^{(\psi)} = \frac{1}{M}\sum_{m=1}^{M} \Delta_e(\mathrm{Pred}_{\pi_m}(e)), \tag{3}$$

where $\mathrm{Pred}_\pi(e)$ is the set of candidates preceding e in a uniformly sampled permutation π, and M is the number of sampled permutations. The permutation determines subset membership, while candidates within each subset are presented to ψ in their original retrieval order.

A positive $\widehat{\phi}_{q,e}^{(\psi)}$ indicates that adding e increases the predictor's probability of the correct response on average over the sampled evidence contexts. Computing these targets requires the observed response y, so supervision targets are generated only for training queries. The downstream predictor ψ remains fixed during contribution estimation.

4.2 TRAINING

We train PERTSELECT to predict a contribution score $a_{q,e}^{(\psi)}$ for each candidate from the query and retrieved evidence. The Shapley targets $\widehat{\phi}_{q,e}^{(\psi)}$ lie in $[-1, 1]$, since each marginal contribution is a difference between probabilities. Predicted scores $a_{q,e}^{(\psi)} \in \mathbb{R}$ are unbounded regression outputs, with no sigmoid or clipping. The selector input includes the cell line, perturbed gene, target gene, KG paths, and experimental measurements, while excluding the query response and annotations derived from it. The input lists the query first, followed by KG paths and then experimental measurements in retrieval order. Appendix B.1 describes the fields included for each candidate.

PERTSELECT uses ModernBERT-base in three variants. PERTSELECT-JOINT encodes the query and all candidates together, then pools each candidate's token representations and applies a shared MLP head to predict its score. PERTSELECT-INDEPENDENT encodes the query with one candidate at a time, predicting its score without seeing other candidates. PERTSELECT-KIND encodes the full pool, averages candidate representations separately for KG paths and measurements, and uses the shared head to produce one score per kind. Each score determines whether to keep or remove all candidates of that kind. All three variants are trained on the same per-candidate Shapley targets with

the loss below and select candidates with the rule in Section 4.3. Appendix B.1 gives the pooling and scoring equations.

We train the selector using a query-balanced Huber regression loss:

$$\mathcal{L}_{\mathrm{Shap}}(\theta_\psi) = \frac{1}{|\mathcal{D}_\psi^+|} \sum_{(q,y)\in\mathcal{D}_\psi^+} \frac{1}{|E_q^\psi|} \sum_{e\in E_q^\psi} \ell_\delta\Big(a_{q,e}^{(\psi)} - \widehat{\phi}_{q,e}^{(\psi)}\Big), \tag{4}$$

where ℓ_δ is the Huber loss with $\delta = 1$, and $\mathcal{D}_\psi^+$ contains training queries with at least one candidate. Averaging candidate losses within each query gives equal weight to queries with different candidate-pool sizes. For PERTSELECT-KIND, candidates of the same kind share the same predicted score while retaining their individual Shapley targets in the loss. The downstream predictor ψ is not updated during selector training.

4.3 INFERENCE

At inference, the selector retains candidates with positive predicted contribution:

$$S_{\theta_\psi}(q, E_q^\psi) = \Big\{e \in E_q^\psi : a_{q,e}^{(\psi)} > 0\Big\}. \tag{5}$$

The zero threshold follows the sign of the contribution target, retaining candidates predicted to increase the correct-response probability on average over evidence contexts. The selected subset is then provided to the same downstream predictor ψ. The targets are defined by the correct-response probability, whereas evaluation uses macro-F1.

5 EXPERIMENTAL SETUP

5.1 DATA AND RETRIEVED EVIDENCE

Perturbation data. We use CRISPRi Perturb-seq data for K562 and RPE1 from Replogle et al. (Replogle et al., 2022) and HepG2 and Jurkat from Nadig et al. (Nadig et al., 2025). Each query predicts one target gene's response relative to non-targeting controls. Counts are normalized to 10,000 per cell and log1p-transformed. A two-sided Wilcoxon rank-sum test compares perturbed and non-targeting-control cells. Adjusted $p < 0.01$ defines a changed response, with the sign of the mean-expression difference distinguishing up from down. Raw $p > 0.1$ defines unchanged responses, and intermediate cases are excluded. Appendix A specifies preprocessing and label construction.

Knowledge graph. OmniPath supplies directed biological connectivity, and INDRA/CoGEx supplies relation statements, provenance and recorded biological context (Türei et al., 2016; Gyori et al., 2017). Gene identities are aligned to HUGO Gene Nomenclature Committee (HGNC) identifiers through HGNC and UniProt mappings, while unresolved entities remain as intermediate nodes. Activation and inhibition define positive and negative relation signs. An OmniPath sign is retained when OmniPath records no opposing sign for the same relation. Missing or conflicting signs are resolved by the majority of distinct supporting publications in CoGEx, with ties left unknown. Each relation retains its sources and recorded conditions, and relations supported only by retracted publications are excluded. The resulting KG contains 21,762 nodes and 525,762 directed relations. A path whose relations all have known signs is called fully signed. For a knockdown, such a path implies the negative of the product of its relation signs for the target gene.

Data integration and splits. Datasets share a cell–perturbation–target representation with HGNC gene identifiers. K562, HepG2 and Jurkat supply training and validation queries, while RPE1 is held out. The dataset contains 145,002 training queries, 30,000 validation queries and 35,170 test queries. The four test settings are *ID* (in-distribution): seen perturbations in training cell lines (15,000 queries), *Unseen pert.*: unseen perturbations in training cell lines (7,500), *Held-out cell*: seen perturbations in RPE1 (7,500), and *Both unseen*: unseen perturbations in RPE1 (5,170). Here, seen and unseen refer to whether a perturbation occurs in the training queries. We call the three settings other than ID out-of-distribution (OOD). An ID test query asks about a target gene whose response to the same perturbation in the same cell line is absent from training, while other target genes

of that cell–perturbation pair can appear in training. Test perturbations in the unseen-perturbation settings are held out as whole functional gene clusters constructed from KG connectivity and Gene Ontology memberships, so that functionally related perturbations do not appear in training. Unseen-perturbation validation queries use a random 10% of the remaining perturbations. RPE1 response labels are excluded from training and retrieved measurements. RPE1 non-targeting controls provide the basal expression reported for KG path endpoints in RPE1 queries.

Evidence retrieval. Qwen receives up to five KG paths of at most four hops and five experimental measurements, with each complete path treated as one selection unit. We search for paths with low total edge cost, assigning lower costs to edges whose endpoints have higher personalized PageRank scores from the perturbed gene. Reusing an edge increases its cost in subsequent searches to diversify the retrieved paths. Measurement retrieval adapts SUMMER, the retrieval method of PerturbQA (Wu et al., 2025), using three groups: related perturbation–related target, related perturbation–any target, and any perturbation–related target. Related genes are the query gene and the genes sharing the most neighbors with it in the KG. Measurements are sampled within groups and selected in round-robin group order, removing duplicates. Voting uses a separate pool of up to five measurements of the queried target gene, which we call *same-target measurements*. It keeps only measurements whose source perturbation has positive similarity to the query perturbation under the KG-distance similarity of CORE (Yuan et al., 2026), and we call these measurements *eligible*. All measurements come from the training split before query sampling, excluding the query's own observation. Validation and test observations are never retrieved. Retrieval does not use the query response. Details are given in Appendix A.3.

To evaluate evidence selection in settings with potentially conflicting biological evidence, we construct the test set from queries with both fully signed up- and down-implying paths of one to four hops in the full KG, irrespective of agreement with the response label. Training and validation do not impose this condition, so only about half of their queries have a retrieved KG path (Table 4). With at most five retrieved paths per query, these paths imply both directions only for 24.9%, 52.5%, 28.7% and 52.4% of queries in the four test settings. Unchanged responses account for 60.8–88.5% of each test setting. Appendix A details sampling.

5.2 Comparison Methods

Selection baselines. We compare the three PertSelect variants with the following methods, all applied to the same fixed downstream predictor. *All evidence* provides every candidate, and *query only*, *KG only* and *measurements only* provide no candidate, only the paths, or only the measurements. *Uniform-count random* selection draws the number of retained candidates of each kind uniformly between zero and the available number, then draws the candidates at random. *Cosine relevance* ranks candidates by cosine similarity between pretrained ModernBERT embeddings of the query and each candidate, and keeps as many candidates as PertSelect-Joint retains for that query. *LLM pruning* asks Qwen3-8B to choose useful candidates from the retrieved pool. *FILCO-style filtering* adapts FILCO (Wang et al., 2023): a ModernBERT classifier scores each query–candidate pair separately and is trained with binary labels indicating whether the candidate alone increases the predictor's correct-response probability over the query alone. It keeps candidates with predicted probability above 0.5. Appendix B.3 details the selection rules and prompts.

5.3 Evaluation

Downstream predictors. We evaluate two types of downstream predictor: **LLM-based prediction** with Qwen3-8B (Yang et al., 2025) and **voting-based prediction** adapted from CORE (Yuan et al., 2026). Qwen3-8B reads the query together with selected KG paths and experimental measurements to predict whether the target response is unchanged, up or down. Measurement voting combines observed responses for the same target gene, giving greater weight to measurements from perturbations with higher CORE similarity to the query perturbation. Both predictors remain fixed, and contribution targets and selectors are learned separately for each. When no evidence is selected, Qwen uses the query alone and voting assigns uniform class probabilities, resolving ties as unchanged. Voting abstains when a query has no eligible measurement. Probability calculations and prompts are detailed in Appendices B.2, B.2 and B.4.

Table 1: Macro-F1 with measurement voting and Qwen3-8B. Voting is evaluated on answered queries, with coverage of 98.0/98.5/91.7/91.3% in column order. Qwen is evaluated on all queries. Unseen: unseen perturbations in training cell lines. Held-out: seen perturbations in the held-out cell line. Both: unseen perturbations in the held-out cell line.

	Measurement voting				Qwen3-8B (all queries)			
Method	ID	Unseen	Held-out	Both	ID	Unseen	Held-out	Both
All evidence	0.395	0.336	0.338	0.317	0.196	0.110	0.232	0.152
Uniform-count random	0.394	0.336	0.347	0.311	0.228	0.203	0.285	0.243
Cosine relevance	0.413	0.356	0.345	0.329	0.374	0.357	0.333	0.353
LLM pruning	0.448	0.376	0.403	0.351	0.179	0.107	0.168	0.158
FILCO-style filtering	0.461	0.399	0.437	0.386	0.397	0.422	0.431	0.434
PERTSELECT-JOINT	**0.482**	**0.424**	**0.481**	**0.435**	0.439	0.425	0.435	0.458
PERTSELECT-INDEPENDENT	0.445	0.389	0.423	0.364	0.420	**0.448**	0.474	**0.475**
PERTSELECT-KIND	0.395	0.333	0.336	0.314	**0.441**	0.418	**0.480**	0.454

Training and model selection. All learned selectors, including FILCO-style filtering and the supervision ablations in Section 6.2, are trained on the same fixed subset of 10,000 training queries, with voting restricted to the 9,772 queries containing eligible measurements. They use ModernBERT-base (Warner et al., 2024) as the selector backbone with matched optimization settings. The main Shapley targets use 16 sampled permutations per query. We select each checkpoint by the mean macro-F1 on fixed 1,000-query subsets of the seen- and unseen-perturbation validation splits, and use that checkpoint for all four test settings. Selector training uses AdamW with learning rate 2×10^{-5}, weight decay 0.01, seed 0, BF16 and batches of 16 queries per update without gradient accumulation, for at most eight epochs. Appendix B.1 gives the training-sample composition, update counts and selection thresholds.

Evaluation metrics. We report three-class macro-F1 for each test setting. Qwen is evaluated on all test queries, while voting is evaluated on answered queries with coverage reported separately. We report paired 95% bootstrap intervals for macro-F1 differences between PERTSELECT-JOINT and all-evidence prediction, FILCO-style filtering, and the singleton-supervised selector defined in Section 6.2. Each bootstrap sample resamples perturbations with replacement and evaluates both methods on the same queries. Undefined class F1 values are set to zero, and the interval procedure is given in Appendix C.4.

6 EXPERIMENTAL RESULTS

6.1 MAIN RESULTS

PERTSELECT-JOINT improves Qwen macro-F1 over all-evidence prediction by 20.3–31.5 percentage points and achieves the highest macro-F1 among the non-PERTSELECT methods in every setting (Table 1). Its paired intervals against FILCO-style filtering, the strongest baseline, exclude zero on ID and both unseen but include zero on unseen perturbations and the held-out cell (Table 6). With measurement voting, the separately trained joint selector improves over all eligible measurements by 8.6–14.4 points and over FILCO-style filtering by 2.0–4.9 points, with all four intervals above zero. LLM pruning falls below all-evidence prediction for Qwen in three settings. It keeps on average 4.8 of the retrieved paths but only 0.5 measurements, whereas KG paths alone give lower Qwen macro-F1 than all evidence in three settings and measurements alone give higher macro-F1 in all four (Appendix Table 5).

6.2 ABLATION STUDIES

Contribution supervision. We vary only the contribution targets for the same PERTSELECT-JOINT architecture, Huber loss and positive-score rule with fixed Qwen. *Singleton* targets measure a candidate's effect when added to the query alone, $\Delta_e(\emptyset)$, and *LOO* targets measure its effect when added last to all other candidates, $\Delta_e(E_q^{\psi} \setminus \{e\})$. Shapley targets average marginal contributions

Table 2: Ablations of PERTSELECT-JOINT with Qwen (macro-F1). All rows use the joint selector architecture. The three groups vary the contribution targets, selector-input metadata and number of permutations used to estimate Shapley targets. Scores are given to four decimals because several differences are below 0.01. The reference uses Shapley supervision, original metadata and 16 permutations, with scores repeated within each group. The best score in each group is bold. Selector training is described in Appendix B.1, and permutation-specific training differences in Appendix D.1.

Model / configuration	ID	Unseen pert.	Held-out cell	Both unseen
Contribution supervision				
Singleton	0.4126	0.4146	**0.4533**	0.4430
LOO	0.3429	0.2982	0.3148	0.2943
PERTSELECT-JOINT (Shapley)	**0.4395**	**0.4247**	0.4348	**0.4585**
Evidence metadata				
Original metadata	**0.4395**	0.4247	0.4348	0.4585
Metadata removed	0.4303	**0.4317**	**0.4673**	**0.4594**
Shapley permutation budget				
4 permutations	0.4208	0.3740	0.4386	0.4396
8 permutations	0.4109	0.4116	**0.4548**	**0.4600**
16 permutations	**0.4395**	**0.4247**	0.4348	0.4585

over subsets from 16 sampled permutations. Appendix C.2 reports the same comparison for measurement voting.

Shapley exceeds LOO in all settings and singleton in three (Table 2). The Shapley-minus-singleton gains on ID and both unseen are 0.027 and 0.015, with paired 95% intervals [0.016, 0.037] and [0.001, 0.033]. On unseen perturbations in training cell lines, the gain is 0.010 with interval [-0.005, 0.028]. For seen perturbations in the held-out cell line, singleton is higher by 0.018, with a Shapley-minus-singleton interval of [-0.030, -0.007]. Shapley exceeds LOO by 0.097–0.164. Averaging over subsets therefore improves on measuring a candidate only against the full pool, while its advantage over singleton targets depends on the evaluation setting.

Architecture and selection granularity. For Qwen, PERTSELECT-INDEPENDENT is strongest on unseen perturbations and both unseen, while PERTSELECT-KIND leads on ID and the held-out cell (Table 1). All three variants exceed all-evidence prediction in every setting. For measurement voting, PERTSELECT-JOINT exceeds both other variants in all four settings. With measurements as the only evidence kind for voting, PERTSELECT-KIND can only keep or remove the entire eligible pool and stays close to all-evidence voting. PERTSELECT-JOINT is the only variant that exceeds FILCO-style filtering in every setting for both predictors, so we use it as the main variant.

For Qwen, much of the gain comes from how many paths and measurements PERTSELECT-JOINT retains. On average it keeps 0.73 paths and 3.74 measurements per test query and removes all KG paths for 81.4% of queries, which also reduces the mean Qwen prompt from 3,358 to 859 tokens. When we keep these per-query counts and choose the candidates at random within each kind, macro-F1 stays within 0.005 of PERTSELECT-JOINT in all four settings, and the two inputs yield the same response for 98.3% of test queries (Appendix C.3). The learned decision of how much evidence of each kind to provide therefore accounts for most of the improvement over all-evidence prediction.

Evidence metadata. We retrain PERTSELECT-JOINT after removing KG source names, support counts and recorded conditions, and the source cell, study and p-values of measurements, from the selector input. The removed fields include whether a measurement comes from the query cell line and whether its source experiment shares control cells with the query. Query context, path structure, signs, basal expression, measured responses, effect sizes and similarities remain, and Qwen still receives the original evidence. Removal lowers ID macro-F1 by 0.009 and changes the three OOD settings by +0.001 to +0.033 (Table 2). The selection gains therefore do not depend on source metadata.

Table 3: Response shifts pooled over 35,170 Qwen queries and the same 33,677 answered voting queries. U denotes unchanged. Observed U/up/down shares are 74.9/10.1/15.0% for Qwen and 75.4/9.5/15.1% for voting. Recall is computed within each observed class, and the highest recall for each predictor is bold.

Predictor	Input	Predicted share (%)			Recall (%)		
		U	Up	Down	U	Up	Down
Qwen	Query only	0.3	89.4	10.3	0.2	**88.9**	8.6
	All evidence	13.5	1.6	84.9	15.5	3.4	**89.5**
	FILCO-style	66.9	4.6	28.4	74.3	8.8	53.7
	PertSelect-Joint	80.6	2.8	16.6	**88.5**	5.2	43.7
Voting	Query only	100.0	0.0	0.0	**100.0**	0.0	0.0
	All evidence	96.9	0.8	2.3	98.5	3.8	8.4
	FILCO-style	92.6	1.7	5.7	96.9	8.3	20.1
	PertSelect-Joint	88.3	3.0	8.6	93.7	**12.0**	**28.1**

Permutation budget. Sixteen permutations give the highest macro-F1 on ID and unseen perturbations, while eight give the highest on the two held-out-cell settings (Table 2). Target generation requires 255,390, 444,156 and 746,570 predictor evaluations for 4, 8 and 16 permutations, so the larger budget increases offline cost without consistently improving prediction. The eight-permutation selector was partly trained on two GPUs with the same global batch (Appendix D.1).

6.3 Additional Analyses

Predictor-specific response shifts. Selection changes the two predictors' class predictions in opposite directions (Table 3). For Qwen, PertSelect-Joint reduces the predicted down share from 84.9% to 16.6%, compared with 28.4% for FILCO-style filtering and an observed share of 15.0%. For voting, it reduces the predicted unchanged share from 96.9% to 88.3%, compared with 92.6% for FILCO-style filtering and an observed share of 75.4%, and raises up/down recall to 12.0/28.1% from 8.3/20.1%. Relative to FILCO-style filtering, Qwen loses up/down recall (5.2/43.7% versus 8.8/53.7%), while voting loses unchanged recall (93.7% versus 96.9%). Per-setting class F1 and recall for Qwen are given in Appendix Table 7. Fixed class-score offsets fit on validation queries raise all-evidence Qwen macro-F1 only to 0.328/0.252/0.354/0.243, below PertSelect-Joint in every setting (Appendix C.5).

Evidence disagreement. We count disagreement only between up and down directions: the retrieved KG paths imply both directions, the same-target measurements include both up and down labels, or a KG path implies the direction opposite to an up or down same-target measurement. Unchanged measurement labels carry no direction and are not counted. PertSelect-Joint raises Qwen macro-F1 from 0.168 to 0.433 on 12,532 queries with opposing KG paths and from 0.234 to 0.431 on 463 queries with KG–measurement disagreement. Only 8 queries contain both up and down same-target measurements. Queries without disagreement show similar gains, from 0.199 to 0.466 for KG paths and from 0.241 to 0.473 for KG–measurement agreement, so the improvement is not specific to disagreeing evidence (Appendix C.6).

7 Conclusion

We introduced PertSelect, which learns to select KG paths and experimental measurements for a given predictor using predictor-specific Shapley contributions. PertSelect-Joint improves macro-F1 over all-evidence prediction by 20.3–31.5 percentage points for Qwen3-8B and 8.6–14.4 points for measurement voting, and exceeds all selection baselines in every setting. For Qwen, most of this gain comes from deciding how many KG paths and measurements to provide, which corrects a strong bias toward predicting down-regulation and shortens the prompt by 74.4%. For voting, selection instead increases predictions of changed responses. These opposite effects show that useful evidence selection depends on the downstream predictor.

AI USE STATEMENT

We used OpenAI ChatGPT/Codex and Claude Code to assist with ideation, implementation, result interpretation, and manuscript editing. The authors reviewed the AI-assisted content.

ETHICS STATEMENT

This work uses public perturbation datasets and biological knowledge resources. No new human-subject data were collected. Predictions are intended for research, not clinical decisions, and KG coverage reflects curation and literature biases.

REPRODUCIBILITY STATEMENT

The appendices specify data construction, evidence retrieval, predictor prompts and scoring, contribution targets, selector architectures and training, selection baselines and evaluation protocols. All reported results use training seed 0.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 2026. Published online 2026-08-31, DOI 10.1016/j.cell.2026.07.052; preprint bioRxiv 10.1101/2025.06.26.661135. Author list taken from the bioRxiv v2 PDF; TODO confirm against the Cell version.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511.

Andrea Giuseppe Di Francesco, Andrea Rubbi, and Pietro Liò. Retrieval-augmented generation for predicting cellular responses to gene perturbation. arXiv preprint arXiv:2603.07233, 2026. URL https://arxiv.org/abs/2603.07233.

Benjamin M. Gyori, John A. Bachman, Kartik Subramanian, Jeremy L. Muhlich, Lucian Galescu, and Peter K. Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, 2017. TODO verify; CoGEx citation may be needed separately.

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

Dongkwan Kim, Yiming Gao, Yining Yang, and Yang Shen. PertReason: A knowledge-grounded benchmark and framework for cell-state-conditioned mechanistic reasoning of perturbation effects. arXiv:2607.18777, 2026. [verified: local PDF].

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.861/.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. [verified: local PDF, DOI 10.15252/msb.202211517].

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, et al. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57:1228–1237, 2025. Preprint: bioRxiv 10.1101/2024.07.03.601903 (2024), titled "Transcriptome-wide characterization of genetic perturbations". TODO complete author list from the journal page.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. [verified: local PDF].

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLM. *arXiv preprint arXiv:2408.12076*, 2024. doi: 10.48550/arXiv.2408.12076. URL `https://arxiv.org/abs/2408.12076`.

Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, 2016.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation. arXiv:2311.08377, 2023. [verified: arXiv abstract page for title/year; TODO confirm author list].

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

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 [verified: local PDF; TODO volume/pages].

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, David Richmond, Tommaso Biancalani, and Jan-Christian Hütter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations (ICLR)*, 2025. [verified: local PDF].

Yuanfang Xiang and Lun Ai. Adaptive data-knowledge alignment in genetic perturbation prediction. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.00512 [verified: local PDF].

An Yang et al. Qwen3 technical report, 2025. arXiv:2505.09388; TODO verify author list.

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:2606.01042, 2026. [verified: local PDF].

A Data and Evidence Construction

A.1 Observations and Response Labels

We use K562 essential-gene and RPE1 count matrices from Replogle et al. and HepG2 and Jurkat CRISPRi matrices from Nadig et al. (GSE264667) (Replogle et al., 2022; Nadig et al., 2025). Gene features are mapped to approved HGNC identities, and columns mapping to the same gene are summed within each cell before normalization. Unresolved features contribute to library-size normalization but cannot become query targets.

Individual cells are the statistical observations. Within each source and cell line, all cells annotated with the same perturbed gene form the treated group, and all non-targeting cells form a shared control group reused across perturbations. These comparisons pool cells without batch or replicate stratification. Raw counts are not summed across cells into pseudobulk samples.

We normalize each cell to 10,000 total counts and apply a log1p transformation. For each source and cell line, Δ_g is the mean transformed expression of gene g in perturbed cells minus that in non-targeting control cells. The control mean also defines basal expression. We use a two-sided Wilcoxon rank-sum test and apply Benjamini–Hochberg correction across all tested features separately for each perturbation and cell line.

Let $p_{\mathrm{raw},g}$ and $p_{\mathrm{adj},g}$ denote the raw and adjusted p-values. The response label is

$$y_g = \begin{cases} \text{up} & p_{\mathrm{adj},g} < 0.01 \text{ and } \Delta_g > 0, \\ \text{down} & p_{\mathrm{adj},g} < 0.01 \text{ and } \Delta_g < 0, \\ \text{unchanged} & p_{\mathrm{raw},g} > 0.1, \\ \text{unassigned} & \text{otherwise.} \end{cases} \tag{6}$$

Unassigned responses are excluded, and no additional effect-size threshold is applied. This yields 64,020,126 labeled queries before splitting and sampling.

A.2 Splits and Sampling

Functional splitting uses the KG after its gene and protein nodes are mapped to HGNC genes where uniquely resolvable, with unresolved intermediate nodes retained (Section 5.1). It combines this graph's binary directed adjacency with MSigDB 2023.2.Hs C5 Gene Ontology (GO) membership. Personalized PageRank (damping 0.85, tolerance 10^{-8}, at most 1,000 iterations, dangling mass returned to the start) is computed from each HGNC gene. Scores are summed over each of 10,461 GO sets and L2-normalized. KMeans (30 clusters, seed 42, k-means++, 20 initializations) clusters these vectors, and six clusters are assigned to unseen perturbations using seed 42. Among the remaining perturbations, 10% are reserved for validation on unseen perturbations. RPE1 is the held-out cell line. An RPE1 query enters the seen-perturbation category only when its perturbation occurs in training.

For training perturbations in training cells, we partition the 31,971,783 labeled queries into training (80%), validation on seen perturbations (10%) and in-distribution testing (10%). Assignment is deterministic for each query using seed 42 and does not depend on response labels or KG annotations. All 4,316 training cell–perturbation combinations retain training queries. In-distribution evaluation therefore holds out individual perturbation–target queries, while other targets of the same perturbation remain available in training.

Training and both validation sets are sampled from their labeled populations without KG conditions. The test is sampled from queries with fully signed paths of one to four hops implying both directions in the full KG, without requiring a path that matches the observed response. This sampling emphasizes queries for which the available KG can support opposing response directions, allowing us to study evidence selection under potential disagreement. The retrieved evidence of a test query need not contain either direction (Section 5.1).

Sampling uses seed 42 in two stages. Within each split–cell–perturbation group, at most 100 changed (up and down pooled) and 100 unchanged queries are drawn without replacement. A second draw enforces the split size with equal cell budgets. Before the second draw, the training sample keeps one query from every retained cell–perturbation pair, so that each ID test query has training

Table 4: Query counts and evidence availability across training, validation, and test splits. *Pert.*: number of distinct perturbations. *KG pool*: fraction of queries with at least one retrieved knowledge graph path. *Vote pool*: fraction with at least one same-target measurement eligible for voting. *Unch.*: fraction of unchanged responses. All queries have at least one retrieved experimental measurement.

Split	Queries	Pert.	KG pool	Vote pool	Unch.
Training	145,002	1,512	51.5%	97.6%	71.4%
Validation (seen perturbations)	15,000	1,512	51.5%	97.8%	84.6%
Validation (unseen perturbations)	15,000	168	52.6%	97.7%	72.1%
Test (in-distribution)	15,000	754	100.0%	98.0%	88.5%
Test (unseen perturbations)	7,500	36	100.0%	98.5%	70.8%
Test (held-out cell)	7,500	755	100.0%	91.7%	61.6%
Test (unseen perturbations and held-out cell)	5,170	36	100.0%	91.3%	60.8%

queries from the same cell and perturbation. No class ratio is imposed. After sampling, we verify that every ID test query's cell–perturbation pair and every held-out-cell, seen-perturbation query's perturbation occur in the sampled training queries. The training cells each contribute 48,334 training queries, 5,000 queries to each validation split, 5,000 ID test queries and 2,500 unseen-perturbation test queries. RPE1 supplies the two held-out-cell test settings. Table 4 reports split sizes and label proportions.

A.3 Evidence Retrieval and Input Construction

KG retrieval does not filter relations by the query cell and starts with unit edge weights. Query responses, annotations derived from them, and query-cell expression do not select or reorder paths. Let f_u be the personalized PageRank score of node u from the perturbed gene (damping 0.85, at most 50 iterations, L1 tolerance 10^{-4}). The cost of directed edge (u, v) is

$$c_{uv} = \frac{1}{w_{uv} f_u f_v + 10^{-4}}, \tag{7}$$

where w_{uv} initially equals one. We select paths minimizing the sum of edge costs, with at most four hops and no repeated nodes. After selecting a path, we halve w_{uv} for each edge it contains, increasing the cost of using those edges again. A candidate path is rejected if its directed-edge set contains the entire directed-edge set of any previously selected path, including an exact duplicate. The first five retrieved paths form the candidate pool.

Each candidate is a complete retrieved path with its relation assertions, provenance and endpoint basal expression in the query cell. A query without a mapped endpoint or a path within four hops has no KG evidence. Selection retains or removes whole paths. A relation shared by several paths remains available whenever any of those paths is retained.

Experimental measurements. We retrieve from 25,574,141 original training observations across HepG2, Jurkat and K562, excluding the query's own observation and all original validation/test observations. Related genes comprise the query gene and up to ten genes with the highest positive shared-neighbor counts in the undirected, unweighted KG. Following SUMMER (Wu et al., 2025), we sample up to 20 observations without replacement (seed 0) from each of three groups: related perturbation–related target, related perturbation–any target, and any perturbation–related target. Each group pools training cells without using query labels. The first five observations per group are combined and deduplicated, then visited in round-robin group order to retain at most five distinct measurements. Each retains its actual perturbation, target, source cell and measured response.

Voting similarity. The voting pool takes the first five same-target observations with finite positive pooled CORE similarity (Yuan et al., 2026) from the full deduplicated union of the three measurement groups (up to 60 observations), preserving union order. This pool is constructed before the Qwen-specific five-measurement limit and can include observations absent from Qwen's final pool. Let G contain the HGNC genes in the training-cell control-expression profiles that map to the KG. For each perturbation p, undirected, unweighted KG distances to genes in G are assigned to buckets $1, 2, 3, 4, \infty$, where ∞ means no path within four hops. The rate r_d is the fraction of up or down responses in bucket d among original training observations pooled across the three training cells,

excluding all 145,002 sampled training-query labels. Unobserved buckets have rate zero. The vector v_p concatenates a one-hot distance encoding and the corresponding rate for each gene, with the self-gene block set to zero. Similarity between perturbations is the cosine of these vectors, clipped to $[0, 1]$. Perturbations without a mapped KG node are ineligible. This pooled calibration is fixed across query cells and determines voting eligibility and weights. The three-group measurement collection instead uses shared-neighbor retrieval.

Relation metadata. Relation summaries include signs and a source-disagreement flag indicating whether the original sources report opposing activation and inhibition signs for the same relation, activating/inhibiting/undetermined statement counts, publication counts, databases and reading systems, and recorded conditions. Each relation description includes at most two source sentences of 300 characters each and five recorded conditions. The selector omits source sentences. Recorded conditions include cell line, cell type, organ, disease and subcellular location. BTO (2021-10-26) and EFO (v3.94.0) resolve condition names where available, otherwise the identifier is retained. Missing conditions remain unspecified.

Leakage checks. We verify that retrieved measurements contain no observations from the original validation or test splits, no query's own observation, and no duplicates within a query. Across all 47,170 training, validation and test queries used in the main comparison, changing the query response and its derived annotations leaves both Qwen and selector inputs unchanged. Retrieved training measurements retain their observed responses as evidence.

B Implementation Details

B.1 Selector Architecture and Training

The selector uses ModernBERT-base with a shared prediction head containing 128 hidden units. PertSelect-Joint encodes all candidates together. PertSelect-Independent encodes each query–candidate pair without the candidate's position in the retrieved list or information about other candidates. PertSelect-Kind averages candidate representations within each kind and assigns the resulting shared score to every member for the itemwise loss.

For PertSelect-Joint, the ModernBERT-base encoder produces contextual token representations h_t. For each candidate e, let T_e be the token span of its complete text, which covers every relation of a path. Mean pooling and a shared two-layer MLP give one score per candidate,

$$\bar{h}_e = \frac{1}{|T_e|} \sum_{t \in T_e} h_t, \qquad a_{q,e}^{(\psi)} = w_2^\top \,\mathrm{GELU}(W_1 \bar{h}_e + b_1) + b_2. \tag{8}$$

Joint encoding allows each candidate's predicted contribution to depend on the query and the other retrieved candidates. The output is an unbounded scalar on the scale of $\widehat{\phi}$.

PertSelect-Independent builds a separate input sequence for each candidate, containing only the query and that candidate. It computes $\bar{h}_e$ by the same mean over the candidate's tokens in this sequence and applies the same head, so the score of e cannot depend on other candidates. PertSelect-Kind uses the joint sequence and the per-candidate vectors $\bar{h}_e$ from Eq. (8). For each evidence kind $k \in \{\mathrm{KG}, \mathrm{M}\}$ with a nonempty candidate set $E_q^{\psi,k}$, it averages these vectors and applies the shared head,

$$\bar{h}_k = \frac{1}{|E_q^{\psi,k}|} \sum_{e \in E_q^{\psi,k}} \bar{h}_e, \qquad a_{q,e}^{(\psi)} = w_2^\top \,\mathrm{GELU}(W_1 \bar{h}_k + b_1) + b_2 \quad \text{for all } e \in E_q^{\psi,k}. \tag{9}$$

All candidates of one kind therefore receive the same score and are retained or removed together. A kind without candidates receives no score.

Scores above zero retain complete paths or measurements in retrieval order. An empty selected set uses the downstream predictor's empty-input convention (Section 5.3).

Evidence representation. The sequence starts with the cell, perturbed gene and target, followed by KG paths and then measurements in retrieval order. Each path preserves its ordered relations,

including endpoints, signs, statement and publication counts, source databases, recorded conditions and endpoint basal expression in query-cell controls. Measurements contain source cell, cell-match status, whether the source experiment shares control cells with the query, perturbed and measured genes, perturbation similarity when available, response, expression difference, p-values and study source. Query labels, their derived annotations and source sentences are excluded. Missing metadata remains unspecified. Inputs are not truncated. The maximum over the 47,170 queries used for selector training, model selection and testing (10,000 training, 2,000 validation and 35,170 test queries) is 4,190 tokens, below the 8,192-token limit.

Training sample and optimization. We use a fixed set of 10,000 training queries selected with seed 0. Training-query selection does not use evidence or predictor outputs. It covers 1,510 perturbations and has 71.1% unchanged responses. Voting uses its 9,772 queries with eligible measurements. All selector parameters are trainable. AdamW uses learning rate 2×10^{-5}, weight decay 0.01, 16 queries per optimization step without gradient accumulation, seed 0, and BF16 mixed-precision training with FP32 master weights and optimizer state. Contribution regression uses query-balanced Huber loss with $\delta = 1$. FILCO uses query-balanced binary cross-entropy. The maximum is eight epochs, corresponding to 5,000 Qwen-selector updates or 4,888 voting-selector updates.

Model selection. For each method, we select the model with the highest equally weighted mean macro-F1 on fixed samples of 1,000 queries from each validation split. The same selected model is evaluated in all four test settings. Validation is performed every 1,000 updates and at epoch boundaries. Selection thresholds are fixed at zero for contribution regression and 0.5 for FILCO probability scores.

B.2 Downstream Predictors

Qwen scoring. Qwen3-8B receives the query and selected evidence as a single user message using the prompt in Appendix B.4. We disable explicit reasoning and restrict the response to one of three single-token answers: 0 for unchanged, 1 for up-regulated and 2 for down-regulated. A softmax over these three answer logits defines the response probabilities used for prediction and contribution estimation. The predicted response is the class with the highest probability, without stochastic sampling.

Qwen is configured for BF16 inference. All evaluated prompts fit within its context limit without truncation.

Measurement voting. For similarity $s_j > 0$ and source response c_j, let $w_j = \sqrt{s_j}$. For a nonempty selected set, the probability assigned to class y is

$$P_{\text{vote}}(y \mid q, S) = \frac{\sum_{j \in S} w_j \mathbf{1}[c_j = y]}{\sum_{j \in S} w_j}. \tag{10}$$

An empty selected set has uniform probabilities, with a tie resolved as unchanged. A query with no eligible candidate instead abstains and is excluded from answered-only metrics. This same probability rule supplies the voting contribution targets.

For this voting rule, adding a measurement with the correct response cannot decrease the correct-response probability, while adding any other response cannot increase it. When added to the empty set, the respective contributions are $2/3$ and $-1/3$. Consequently, exact Shapley values are positive for matching responses and negative for nonmatching responses. Finite permutation estimates can also be zero. Voting therefore tests learning predictor-specific contribution magnitudes and selection, but does not independently establish the value of modeling interactions that change a candidate's contribution sign.

B.3 Selection Baselines

All methods preserve retrieval order and use the fixed downstream predictor after selection. Qwen comparisons use the same response prompt and scoring rule, while voting comparisons use the same eligible same-target measurement pool and weighted-voting rule. All-evidence prediction supplies

all available candidates, query-only prediction supplies none, KG-only prediction supplies only the retrieved paths, and measurement-only prediction supplies only the retrieved measurements.

Uniform-count random selection samples a count independently for each evidence kind, uniformly from zero to its available count, then samples that many candidates without replacement. Seed 0 and query identity fix both counts and sampled candidates independently of learned selections.

Cosine relevance uses the pretrained ModernBERT encoder to encode the query and each candidate separately. Final token states are averaged excluding padding and special tokens, and candidates are ranked by cosine similarity, breaking ties in retrieval order. BF16 encoding uses FP32 pooling and cosine computation. The main control matches PERTSELECT-JOINT's total count per query without constraining kind composition. This controls ranking at a known count rather than independently predicting the number to retain.

LLM pruning makes a deterministic generative call to choose evidence identifiers, with constrained decoding permitting only supplied identifiers without duplicates. It may keep all, some or none of each kind, and whole paths are indivisible. Appendix B.4.2 gives its instruction. Prediction uses the selected original evidence without rewriting it. For voting, the same Qwen3-8B model and instruction receive only the voting-eligible same-target measurements, and the knowledge-graph list is shown as "None supplied."

For voting, cosine relevance ranks only the voting-eligible measurements and matches the per-query count selected by the voting PERTSELECT-JOINT selector.

FILCO-style filtering (Wang et al., 2023) independently encodes each query–candidate pair with ModernBERT, mean pooling and the shared MLP head. It uses binary cross-entropy with targets indicating positive singleton gain under the same downstream predictor and retains candidates with keep probabilities above 0.5.

B.4 PROMPTS AND INPUT FORMATS

B.4.1 RESPONSE PREDICTION

We adapt the evidence-conditioned and no-evidence multiple-choice formats of ConflictBank (Su et al., 2024) to perturbation responses. A single user message, without a system message, uses the following template.

```
According to the evidence provided and your knowledge,
choose the best answer from the following options.

Evidence:
{numbered_evidence_items}

Question:  In {cell} cells, the gene {perturbation} is
knocked down by CRISPRi.  Compared with non-targeting
control cells, how does the expression of {target_gene}
change?

0.  Unchanged
1.  Up-regulated
2.  Down-regulated

Return only the digit corresponding to your answer:  0, 1,
or 2.
Answer:
```

KG descriptions retain relation content, source summaries and recorded context (Appendix A.3). Measurements retain their source identity and fields specified below. The query response is excluded. Removing an item removes its entire description and renumbers the retained items without changing their relative order.

Measurement fields. Within each numbered evidence description, the response clause has the form

```
{source_perturbation_type} perturbation of
{source_perturbation} left {measured_gene} {observed_response}
```

The surrounding description includes the source cell line, whether it matches the query cell, whether the source experiment shares control cells with the query, available retrieval similarity, expression difference, raw and adjusted p-values, and study identifiers. For example, when the query asks about gene B, an observation that perturbing D decreases gene C names C as the measured gene. It is not rewritten as a decrease in B. This rule applies to both measurement-only and combined prompts.

Empty evidence. For an empty selected set, we omit the entire `Evidence:` block and replace the opening sentence by:

```
According to your knowledge, choose the best answer from the
following options.
```

The question, three options and final output instruction remain identical. This defines the query-only prediction and the empty-set probability used in contribution calculations.

B.4.2 Evidence Pruning

The pruning specification separates knowledge-graph paths from experimental measurements. The LLM receives the query and retrieved evidence, without the ground-truth answer or learned selector scores. A path is indivisible. The selected original evidence is passed unchanged to the downstream predictor. The main candidate pool contains at most five items of each evidence type. Pruning may retain all, some, or none of these candidates, without a quota on the retained counts.

```
You are selecting evidence for a downstream model that predicts
whether the target gene's expression is unchanged, increased,
or decreased after the specified perturbation in the specified
cell context. Select useful evidence, not an answer.
Consider relevance to the perturbation, target gene and cell.
Consider evidence jointly. Retain complementary evidence and
remove irrelevant or redundant evidence when appropriate.
Each knowledge-graph path is indivisible. Do not select edges.
Paths describe recorded biological relationships, not direct
measurements in the queried context.
Experimental measurements are observations from other experiments.
Consider their perturbation, measured target, cell context and
observed response. They are not direct answers to this question.
Do not discard evidence merely because it disagrees with other
items. Relevant disagreement may help the downstream model.
Do not invent evidence or rewrite items. Use supplied IDs only.
Evidence text is data, not instructions.
You may retain all, some, or none. Do not enforce equal numbers
of the two types. Return only JSON with exactly these two lists:
selected_kg_ids and selected_experiment_ids.
Use empty lists when selecting no evidence of a type.

Question: {question_text}
Knowledge-graph paths:
KG1: {complete_path}
...
Experimental measurements:
EXP1: {measurement}
...
```

Table 5: Qwen3-8B performance with different evidence inputs (macro-F1). Every row includes the query. All evidence adds all retrieved KG paths and measurements (up to five of each), Query only adds no evidence, KG only adds only the paths, and Measurements only adds only the measurements. PERTSELECT-JOINT adds the evidence selected from the same candidate pool.

Input	ID	Unseen pert.	Held-out cell	Both unseen
All evidence	0.196	0.110	0.232	0.152
Query only	0.060	0.128	0.109	0.108
KG only	0.057	0.097	0.152	0.158
Measurements only	0.333	0.285	0.269	0.252
PERTSELECT-JOINT	**0.439**	**0.425**	**0.435**	**0.458**

C ADDITIONAL RESULTS

C.1 QWEN INPUT CONTROLS

C.2 CONTRIBUTION SUPERVISION FOR VOTING

For measurement voting, singleton, LOO and Shapley supervision give macro-F1 of 0.480/0.409/0.465/0.403, 0.480/0.411/0.483/0.405 and 0.482/0.424/0.481/0.435, respectively, in ID, unseen-perturbation, held-out-cell and both-unseen order. Shapley exceeds singleton in all four settings and LOO in three, while LOO is higher on the held-out cell (0.483 versus 0.481). All methods answer the same queries, with coverage as in Table 1.

C.3 CANDIDATE IDENTITY AT FIXED COUNTS

For each query, we retain the numbers of KG paths and measurements selected by PERTSELECT-JOINT, then sample that many candidates without replacement within each kind using seed 0. The original retrieval order, Qwen predictor and scoring rule remain fixed. This conditional control inherits the learned selector's counts and is not an independent random-selection baseline. In ID, unseen-perturbation, held-out-cell and both-unseen order, its macro-F1 is 0.4387/0.4201/0.4337/0.4608, compared with 0.4395/0.4247/0.4348/0.4585 for PERTSELECT-JOINT.

The two methods select identical candidates in identical order for 23,015 of 35,170 queries and produce the same response for 34,567 queries (98.3%). Among the 12,155 queries with different selected candidates, 603 have different responses. These results show that most prediction performance is preserved when learned counts are held fixed and candidate identities are randomized. They do not establish statistical equivalence, the irrelevance of candidate content, or the optimality of the learned counts.

C.4 PAIRED UNCERTAINTY

We compute paired 95% percentile intervals from 2,000 bootstrap replicates, resampling perturbations within each test setting and retaining all their queries with shared draws across methods. Voting uses jointly answered queries. ID and held-out-cell tests contain 754 and 755 perturbations, while each unseen-perturbation test contains 36. These exploratory intervals do not include training-seed variation or multiplicity correction.

C.5 CLASS-SPECIFIC PERFORMANCE AND CLASS-OFFSET CONTROL

For Qwen, PERTSELECT-JOINT corrects 19,704 all-evidence errors and introduces 2,818 new errors, mainly recovering unchanged responses while missing some down responses.

For voting, the query-only row in Table 3 follows from the rule that resolves uniform probabilities as unchanged.

Table 7 shows the class tradeoffs underlying Qwen's macro-F1 gains: unchanged and down F1 improve throughout, while down recall decreases.

Table 6: PERTSELECT-JOINT (Shapley supervision) minus each comparator: macro-F1 differences and paired perturbation-cluster 95% intervals.

Predictor	Setting	vs. all evidence	vs. FILCO-style	vs. singleton
Qwen	ID	+0.244 [+0.230, +0.257]	+0.042 [+0.029, +0.055]	+0.027 [+0.016, +0.037]
Qwen	Unseen perturbation	+0.315 [+0.289, +0.339]	+0.003 [-0.016, +0.026]	+0.010 [-0.005, +0.028]
Qwen	Held-out cell	+0.203 [+0.189, +0.217]	+0.004 [-0.010, +0.016]	-0.018 [-0.030, -0.007]
Qwen	Both unseen	+0.306 [+0.285, +0.329]	+0.024 [+0.002, +0.049]	+0.015 [+0.001, +0.033]
Voting	ID	+0.086 [+0.073, +0.099]	+0.020 [+0.007, +0.033]	+0.002 [-0.007, +0.011]
Voting	Unseen perturbation	+0.088 [+0.065, +0.106]	+0.025 [+0.010, +0.038]	+0.015 [+0.002, +0.028]
Voting	Held-out cell	+0.144 [+0.130, +0.157]	+0.044 [+0.033, +0.055]	+0.017 [+0.008, +0.025]
Voting	Both unseen	+0.118 [+0.100, +0.137]	+0.049 [+0.032, +0.067]	+0.032 [+0.021, +0.042]

Table 7: Qwen class-specific performance before and after selection by PERTSELECT-JOINT. U denotes unchanged.

Setting	Input	U F1	Up F1	Down F1	U recall	Up recall	Down recall
ID	All evidence	0.389	0.054	0.144	0.246	0.037	**0.810**
ID	PertSelect-Joint	**0.920**	**0.070**	**0.329**	**0.921**	**0.047**	0.385
Unseen pert.	All evidence	0.047	0.004	0.278	0.024	0.002	**0.970**
Unseen pert.	PertSelect-Joint	**0.794**	**0.064**	**0.416**	**0.800**	**0.041**	0.520
Held-out cell	All evidence	0.240	**0.108**	0.348	0.146	**0.064**	**0.832**
Held-out cell	PertSelect-Joint	**0.798**	0.083	**0.424**	**0.916**	0.048	0.379
Both unseen	All evidence	0.009	0.048	0.399	0.004	0.026	**0.978**
Both unseen	PertSelect-Joint	**0.773**	**0.111**	**0.491**	**0.834**	**0.079**	0.478

Voting response tradeoff. On the same 33,677 answered queries, selection raises pooled macro-F1 from 0.359 to 0.468. Up/down F1 increases from 0.069/0.146 to 0.183/0.358, while precision falls from 44.9/56.0% to 37.9/49.3%. Accuracy remains nearly unchanged (75.9% versus 76.0%). Thus, the gain reflects better recognition of changed responses rather than a substantial increase in overall accuracy.

Class-offset control. To test whether a fixed class-bias correction recovers the gain without changing evidence, we set $\widetilde{P}(y \mid q, E) \propto P_{\mathrm{Qwen}}(y \mid q, E)\exp(b_y)$. We fix $b_{\mathrm{unchanged}} = 0$ and fit the other two offsets by multiclass negative log-likelihood on the fixed 1,000-query subsets of each validation split. The fitted offsets are transferred unchanged to all test settings. The resulting macro-F1 is 0.328/0.252/0.354/0.243, below PERTSELECT-JOINT in every setting.

C.6 EVIDENCE DISAGREEMENT

For each retrieved path ρ with known relation signs, the knockdown-implied direction is $d(\rho) = -\prod_{r\in\rho}\mathrm{sign}(r)$. Disagreement is defined only between up and down. Within KG paths, a query shows disagreement when its retrieved paths imply both directions. Within same-target measurements, it shows disagreement when these measurements include both up and down labels. Between kinds, it shows disagreement when some path direction differs from some up or down same-target measurement label, and agreement when all path directions and all such labels are the same direction. Unchanged measurement labels carry no direction and are ignored, so a query whose same-target measurements are all unchanged has no comparable measurement direction. A within-kind comparison requires at least one direction, and a cross-kind comparison requires a direction from both kinds. Queries without a comparable direction are excluded rather than counted as agreement. Comparisons use the original retrieved pool, without using the query response. Shapley supervision gives the highest macro-F1 in every group with enough queries. Gains over all-evidence prediction are similar with and without disagreement, so these comparisons do not identify disagreement resolution as the source of improvement. Differences across source conditions are not interpreted as contradictory biological facts.

Table 8: Qwen3-8B macro-F1 by evidence disagreement in the original five-path/five-measurement input, pooled across all 35,170 test queries. N is the number of queries, and groups may overlap across the three comparisons. Queries without a comparable direction are excluded: 11,286 for KG paths, 34,237 for same-target measurements and 34,523 between KG and measurements. Only 8 queries contain both up and down same-target measurements, too few for macro-F1. Best macro-F1 in each row is bold.

Input group	N	All evidence	Singleton	LOO	Shapley (PERTSELECT-JOINT)
Within KG paths					
Disagreement	12,532	0.168	0.427	0.332	**0.433**
No disagreement	11,352	0.199	0.456	0.317	**0.466**
Within same-target measurements					
Disagreement	8	–	–	–	–
No disagreement	925	0.254	0.402	0.385	**0.460**
Between KG and measurements					
Agreement	184	0.241	0.381	0.346	**0.473**
Disagreement	463	0.234	0.409	0.379	**0.431**

D SENSITIVITY ANALYSES

D.1 PERMUTATION BUDGET AND ESTIMATION STABILITY

Targets use the first 4, 8 or all 16 permutations of the same sequence, with matched queries, predictor and optimization settings. The selected 4- and 16-permutation models use single-GPU training, whereas the selected 8-permutation model includes two-GPU updates after step 4,000. The global batch remains 16, but distributed numerical operations and random-number sequences differ. Thus, Table 2 does not isolate permutation count alone.

Split-half stability. Across 74,870 candidates from 10,000 training queries, estimates from the first and last eight permutations have mean absolute difference 0.119 and agree on positive contribution for 88.8% of candidates. Their selected sets have mean Jaccard similarity 0.828, computed as intersection size divided by union size and set to one when both sets are empty. This measures agreement between two sampled estimates, not error against exact Shapley values.

Offline scoring cost. Targets from 4/8/16 permutations require 255,390/444,156/746,570 distinct query–subset probability evaluations, compared with 84,870 for singleton or LOO targets. Each distinct query–subset pair is scored once and reused.