DRUG RESPONSE PREDICTION: SHIFTING FROM CELL LINES TO PATIENTS THROUGH PATHWAY-STRUCTURED INTERACTION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Drug response models can achieve strong global correlations while remaining less accurate at ranking patients for the same drug. This discrepancy matters when transferring knowledge from cell-line screens to patient cohorts with limited response labels. We propose EviDR, an evidence-guided framework that combines large-scale PRISM supervision with limited BeatAML ex vivo response labels. EviDR maps both domains into a shared Reactome pathway space and integrates molecular drug representations with cross-fitted drug–pathway Evidence derived from source response associations. Reliability-weighted Evidence relevance guides a drug-conditioned pathway gate, while signed Evidence interacts with sample pathway states to construct patient-specific representations. Under patient-held-out five-fold evaluation, EviDR achieves the highest mean Macro Spearman of 0.329 among the evaluated methods, compared with 0.314 for TransDRP and 0.313 for MAC-CDR. It also achieves $1.790\times$ enrichment of experimentally sensitive samples relative to group-specific random selection. Comparisons with random patient–drug pair splitting highlight the importance of evaluating unseen patients, while pathway analyses identify recurring source associations and show prediction sensitivity to Evidence identity and direction. These findings support pathway-structured supervised transfer for within-drug response ranking and provide an explicit connection between source-derived associations and patient-specific pathway representations.

1 INTRODUCTION

Drug response prediction aims to estimate how a biological sample responds to a therapeutic compound from its molecular profile and the representation of the drug. Large pharmacogenomic screens such as PRISM provide extensive response measurements across cancer cell lines, offering abundant supervision for predictive modeling (Barretina et al., 2012; Iorio et al., 2016; Corsello et al., 2020). Patient-derived datasets remain considerably smaller because paired molecular profiles and ex vivo drug measurements are difficult to collect at scale. For example, BeatAML provides transcriptomic and drug response measurements from patients with acute myeloid leukemia, but contains far fewer observations than large cell-line screens (Tyner et al., 2018). Transferring knowledge from cell lines to patients is therefore a promising strategy for patient drug response prediction under limited target-domain labels.

A key challenge is that strong aggregate performance may not reflect generalization to unseen patients. Drug response distributions vary substantially across compounds, allowing a model to achieve high global correlation by capturing between-drug differences while remaining less accurate at ranking patients for the same drug. Randomly splitting patient–drug pairs can further place records from the same patient in both training and test sets (Bernett et al., 2026; Costello et al., 2014; Stransky et al., 2015). For drug-specific patient stratification, the relevant objective is to distinguish the responses of previously unseen patient-derived samples tested with the same compound. We therefore study supervised cell-line-to-patient transfer under patient-held-out evaluation, using per-drug Macro Spearman as the primary metric and Macro Pearson as an additional measure of within-drug predictive performance.

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.

Learning under this protocol requires addressing both label scarcity and the distributional differences between cell lines and patients (Sharifi-Noghabi et al., 2020; He et al., 2022; Chen et al., 2022). Although their gene-level expression profiles differ, both domains share functional processes such as cell-cycle regulation, DNA replication, apoptosis, and intracellular signaling. We map PRISM cell lines and BeatAML patients into a shared Reactome pathway space (Milacic et al., 2024; Tang et al., 2022; Li et al., 2023), enabling large-scale cell-line supervision and limited patient labels to be learned through a common functional representation. To organize drug-specific source knowledge, we derive drug–pathway Evidence from PRISM response data. For each drug, the Evidence records source-derived pathway associations, their direction, and their reliability across source-fold estimates, providing a structured representation that can be combined with the pathway state of each sample.

We introduce EviDR, an evidence-guided framework for cross-domain drug response prediction from cell lines to patients. EviDR constructs drug-specific pathway Evidence from PRISM through cross-fitting, assigning out-of-fold Evidence to PRISM training samples and full-fit Evidence to BeatAML samples. It decomposes Evidence into reliability-weighted relevance and signed response context: the former guides a drug-conditioned pathway gate, while the latter modulates the interaction between drug-associated pathways and patient pathway states. The resulting interaction is fused with sample and molecular drug representations to predict response AUC, providing a structured connection between source-domain associations and patient-specific predictions. We evaluate EviDR on PRISM-to-BeatAML transfer using patient-held-out outer five-fold evaluation. Random patient–drug splitting substantially inflates performance: Macro Spearman decreases from 0.577 to 0.329 for EviDR and from 0.609 to 0.313 for MAC-CDR when evaluation is changed to the patient-held-out setting. Under the patient-held-out setting, EviDR achieves a Macro Spearman of 0.329 and a Global Spearman of 0.695, outperforming MAC-CDR in three of five folds and for 60.6% of evaluable drugs. In the evaluated source-scaling fold, full PRISM supervision yields the strongest performance on unseen patients. Finally, highly ranked Evidence pathways are recovered across PRISM cross-fit subsets far more frequently than expected by chance, while perturbing Evidence identity or direction consistently changes patient-level predictions.

Our contributions are threefold. (1) We introduce EviDR, a pathway-structured transfer framework that combines shared Reactome representations, molecular drug features, and cross-fitted drug–pathway Evidence to model sample-specific drug responses. (2) We evaluate PRISM-to-BeatAML transfer under patient-held-out splits with within-drug ranking metrics, and compare this setting with patient–drug pair splitting to characterize the effect of evaluation protocols. (3) We assess predictive performance against representative baselines and examine source-supervision utility, sensitive-sample enrichment, and the incorporation of source-derived pathway associations into patient-specific representations.

2 Related Work

2.1 Drug Response Prediction with Biological Knowledge

Drug response prediction is commonly formulated as learning an interaction function between a biological sample and a compound. DeepCDR integrates multi-omics profiles with graph-based drug representations (Liu et al., 2020), GraphDRP examines graph neural networks for molecular encoding (Nguyen et al., 2021), and TransEDRP uses Transformer encoders to model genomic and chemical inputs (Li et al., 2022). Multimodal and contrastive objectives have further been introduced to improve drug–sample representation learning (Xia et al., 2024; Li et al., 2024; Lawrence et al., 2024; Wu et al., 2025). Biological knowledge can additionally organize these representations through pathways, molecular interaction networks, drug targets, or knowledge graphs (Tang et al., 2022; Shin et al., 2022; Kuenzi et al., 2020; Jin & Nam, 2021; Tang & Gottlieb, 2021).

Most existing methods are developed and evaluated on cancer cell-line benchmarks, where pooled metrics can be strongly influenced by between-drug response differences. High aggregate correlation therefore does not necessarily imply accurate ranking of patients receiving the same drug. We instead evaluate generalization to entirely unseen patients using per-drug correlations. EviDR also differs in how biological knowledge enters the predictor: rather than using pathways only as a fixed representation, it estimates reliability-weighted and signed drug–pathway associations from source responses and uses them to construct patient-specific drug interactions.

2.2 Cross-Domain Transfer from Cell Lines to Patients

Large pharmacogenomic screens provide abundant drug-response supervision, whereas matched molecular and response measurements remain limited in patient cohorts (Chen et al., 2022). This imbalance has motivated methods that transfer knowledge from preclinical systems to patient-derived samples. TRANSACT learns a nonlinear consensus representation shared by cell lines and tumors (Mourragui et al., 2021); few-shot approaches adapt predictors using limited target-domain observations (Ma et al., 2021); and pathway-based methods transfer models through functional representations of gene expression (Tang et al., 2022; Tang & Gottlieb, 2021).

More recent approaches explicitly model the discrepancy between source and target domains. TransDRP combines adversarial alignment with biological knowledge (Liu & Li, 2025), DeepSADR models transferable interactions between drug substructures and genomic functional subsequences (Zhang et al., 2026b), and MAC-CDR performs drug-conditioned adaptation using memory anchors and multimodal compound representations (Zhang et al., 2026a). These methods transfer predictive information through learned cross-domain representations. EviDR complements this direction with an explicit pathway-structured transfer mechanism: PRISM and BeatAML are mapped into a shared Reactome space, and cross-fitted drug–pathway Evidence derived from PRISM guides how each patient's pathway state interacts with a drug. This connects source-domain supervision to patient prediction through a drug-specific and biologically organized intermediate representation.

3 Method

3.1 Problem Setup and Overview

We consider drug response prediction across a source cell-line domain and a target patient domain. The source and target datasets are

$$\mathcal{D}_s = \{(v_i^s, d_i, y_i^s, w_i^s)\}_{i=1}^{N_s}, \qquad \mathcal{D}_t = \{(v_j^t, d_j, y_j^t, w_j^t)\}_{j=1}^{N_t}, \tag{1}$$

where v is a gene-expression profile, d is a drug, y is drug-response AUC, and w denotes observation quality. Lower AUC indicates greater drug sensitivity. The target training set is substantially smaller than the source set, and the objective is to learn a predictor that generalizes to unseen target patients.

EviDR organizes cross-domain transfer around a shared biological coordinate system. Gene-expression profiles from both domains are mapped to K Reactome pathway activities,

$$x = \phi(v) \in \mathbb{R}^K, \tag{2}$$

and each drug is represented by a molecular fingerprint m_d, encoded as

$$z_d = f_{\text{drug}}(m_d). \tag{3}$$

As illustrated in Figure 1, EviDR first estimates drug–pathway response associations from PRISM, separates their relevance and direction, and uses them to construct a drug-conditioned interaction with the pathway state of each sample. The interaction representation is then combined with sample and drug embeddings to predict response AUC.

3.2 Cross-Fitted Drug–Pathway Evidence

For each drug d and pathway k, we estimate a signed response association from PRISM:

$$E_{d,k} = \text{Corr}_w\left(\widetilde{x}_k^s, \widetilde{y}_d^s\right), \tag{4}$$

where $\widetilde{x}_k^s$ and $\widetilde{y}_d^s$ are the pathway activity and drug response after lineage-wise mean centering. The weighted correlation accounts for PRISM measurement quality. The magnitude of $E_{d,k}$ measures association strength, while its sign indicates whether increased pathway activity is associated with higher or lower response AUC.

Directly estimating Evidence from all source observations would expose a PRISM training sample to an association partly constructed from its own response. We therefore divide PRISM samples into Q folds. A source sample in fold q receives Evidence estimated from the remaining folds:

$$E_d^{(-q)} = \text{Evidence}\left(\mathcal{D}_s \setminus \mathcal{D}_s^q\right). \tag{5}$$

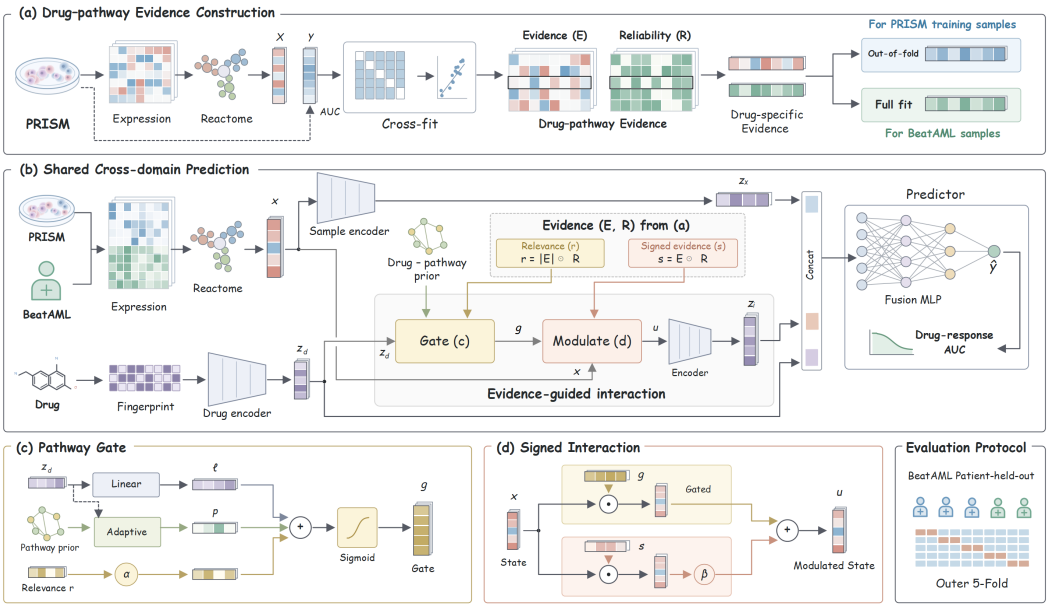

Figure 1: Overview of EviDR. (a) Cross-fitted, lineage-adjusted drug–pathway Evidence and its reliability are estimated from PRISM. PRISM training samples use out-of-fold Evidence, whereas BeatAML samples use Evidence fitted on the available PRISM training cohort. (b) PRISM and BeatAML expression profiles are mapped to a shared Reactome pathway space and combined with molecular drug representations. (c) Evidence relevance, a drug-derived gate, and a pathway prior determine which pathways are emphasized. (d) Signed Evidence modulates the selected pathway states before fusion and response prediction.

BeatAML samples receive E_d^{full}, estimated from the available PRISM training cohort without using BeatAML response labels.

We associate each drug–pathway estimate with a reliability score. For a PRISM sample in fold q, both Evidence and its analytical reliability are estimated from the remaining source folds. For BeatAML, the full-fit analytical reliability is additionally weighted by the sign consistency and variability of the Q cross-fit estimates:

$$R_{d,k}^{(-q)} = A\left(E_{d,k}^{(-q)}, n_d^{\text{eff},(-q)}\right), \tag{6a}$$

$$R_{d,k}^{\text{full}} = A\left(E_{d,k}^{\text{full}}, n_d^{\text{eff,full}}\right) \left|\frac{1}{Q}\sum_{q=1}^{Q}\text{sign}\left(E_{d,k}^{(-q)}\right)\right| \exp\left[-\frac{\text{Std}_q\left(E_{d,k}^{(-q)}\right)}{\tau}\right]. \tag{6b}$$

Here, $A(\cdot)$ is a sample-size-adjusted analytical confidence function and τ controls the cross-fit variability penalty. We remove low-reliability associations and retain at most M pathways per drug.

Evidence is then decomposed into relevance and signed components:

$$r_d = |E_d| \odot R_d, \qquad s_d = E_d \odot R_d. \tag{7}$$

The relevance vector r_d identifies pathways supported for drug d, whereas s_d additionally preserves the direction of their response associations.

3.3 Evidence-Guided Cross-Domain Prediction

For a sample–drug pair (x, d), EviDR first constructs a drug-specific pathway gate. Let h_d^{dir} and h_d^{nbr} denote pathway priors derived from known drug targets and their neighboring Reactome pathways. Their contributions are adaptively weighted using the molecular drug representation:

$$p_d = a_d^{\text{dir}} h_d^{\text{dir}} + a_d^{\text{nbr}} h_d^{\text{nbr}}, \qquad a_d^{\text{dir}}, a_d^{\text{nbr}} \geq 0. \tag{8}$$

The final pathway gate combines the learned drug signal, biological prior, and Evidence relevance:

$$g_d = \sigma\left(W_g z_d + b_g + p_d + \eta_r r_d\right), \tag{9}$$

where $\eta_r \geq 0$ is learned and $g_d \in [0,1]^K$. This gate emphasizes pathways relevant to the current drug while retaining information learned from molecular structure and known targets.

The signed component of Evidence is used to match the selected pathways with the state of the current sample:

$$u_{x,d} = x \odot g_d + \eta_s(x \odot s_d), \qquad \eta_s \geq 0. \tag{10}$$

The first term provides a drug-conditioned pathway representation, and the second captures whether the sample pathway state aligns with the signed response pattern observed in PRISM. Consequently, the same drug-level Evidence can produce different interaction representations for patients with different pathway states.

The original sample state and the Evidence-guided interaction are encoded separately:

$$z_x = f_{\text{sample}}(x), \qquad z_{x,d} = f_{\text{int}}(u_{x,d}). \tag{11}$$

The response predictor is

$$\widehat{y}_{x,d} = f_{\text{fusion}}\left(z_x \| z_d \| z_{x,d}\right), \tag{12}$$

where $\|$ denotes concatenation. The encoders and fusion MLP are shared across PRISM and BeatAML, allowing source supervision to shape the same interaction function used for patient prediction.

3.4 Joint Learning Objective

EviDR is trained with separate PRISM and BeatAML mini-batches. For each domain $a \in \{s,t\}$, the regression objective combines a quality-weighted observation loss with a drug-balanced loss:

$$\mathcal{L}^a_{\text{reg}} = (1-\gamma)\mathcal{L}^a_{\text{row}} + \gamma\mathcal{L}^a_{\text{drug}}, \tag{13}$$

where both terms use the Huber loss. The drug-balanced term first averages over observations of the same drug and then assigns equal weight to each drug, preventing frequently measured compounds from dominating training.

Because the primary task is to rank patients receiving the same drug, we apply a drug-balanced pairwise objective to the standardized responses used during training. Let $\mathcal{D}^B_a$ denote the drugs with at least one valid pair in a mini-batch and let

$$\mathcal{P}_{a,d} = \{(i,j) : d_i = d_j = d,\ |\widetilde{y}_i - \widetilde{y}_j| > 0.05\}.$$

The ranking loss is

$$\begin{aligned} \mathcal{L}^{\text{rank}}_a &= \frac{1}{|\mathcal{D}^B_a|} \sum_{d \in \mathcal{D}^B_a} \frac{\sum_{(i,j)\in\mathcal{P}_{a,d}} \omega_{ij} \log\Big[1 + \exp\Big(-\operatorname{sign}(\widetilde{y}_i - \widetilde{y}_j)(\widehat{\widetilde{y}}_i - \widehat{\widetilde{y}}_j)\Big)\Big]}{\sum_{(i,j)\in\mathcal{P}_{a,d}} \omega_{ij}}, \\ \omega_{ij} &= \min\left(\frac{|\widetilde{y}_i - \widetilde{y}_j|}{s}, 1\right)\sqrt{w_i w_j}, \qquad s = 1. \end{aligned} \tag{14}$$

Here, $\widetilde{y}$ and $\widehat{\widetilde{y}}$ are the standardized response and prediction, respectively, and w_i denotes observation quality. If a mini-batch contains no valid pair, the corresponding ranking loss is set to zero.

The domain-specific loss is

$$\mathcal{L}^a = \mathcal{L}^a_{\text{reg}} + \lambda^a_{\text{rank}}\mathcal{L}^a_{\text{rank}} + \lambda_{\text{path}}\mathcal{L}^a_{\text{path}} + \lambda_{\text{moa}}\mathcal{L}^a_{\text{moa}} + \lambda_{\text{prior}}\mathcal{L}^a_{\text{prior}}. \tag{15}$$

The pathway and mechanism-of-action objectives regularize the drug representation using known biological annotations, while $\mathcal{L}_{\text{prior}}$ stabilizes the adaptive pathway-prior weights. The complete cross-domain objective is

$$\mathcal{L} = \lambda_s\mathcal{L}^s + \lambda_t\mathcal{L}^t. \tag{16}$$

PRISM mini-batches provide broad source supervision, whereas BeatAML mini-batches adapt the shared predictor to the patient domain. Detailed architectures, reliability parameters, and optimization settings are provided in Appendix A.1. Evidence estimation and reliability parameters are detailed in Appendix A.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets. We study cell-line-to-patient drug response transfer using PRISM as the source domain and BeatAML as the target domain. For PRISM, we retain secondary-screen measurements with valid AUC values, mapped expression profiles, and successful STR-profiling quality control. Replicate measurements are aggregated at the cell-line–drug level, with response-fit quality used as an observation weight. The processed source dataset contains 476 cancer cell lines and 617,776 response observations across 1,445 compounds (Corsello et al., 2020).

For BeatAML, we retain converged single-agent dose–response curves with available patient identifiers and matched RNA-seq samples. Drug names and molecular identities are reconciled between PRISM and BeatAML, resulting in 479 patients, 520 RNA samples, 35,748 response observations, and 99 shared drugs (Tyner et al., 2018). Lower AUC denotes greater drug sensitivity in both domains. Expression profiles are restricted to genes measured in both datasets, standardized using the training samples of each domain, and aggregated into 1,000 Reactome pathway activities (Milacic et al., 2024). Compounds are represented by 512-bit radius-2 Morgan fingerprints (Rogers & Hahn, 2010). Further preprocessing details are provided in Appendix A.3.

Evaluation Protocols. Our primary evaluation follows a patient-held-out outer five-fold protocol. All observations from the same BeatAML patient are assigned to a single fold, ensuring that test-patient molecular profiles are unseen during both training and model selection. Within each outer fold, models are selected using held-out validation patients, and final results are reported as the mean $\pm$ standard deviation across five test folds.

The primary metric is per-drug Macro Spearman. We compute the correlation between predicted and observed AUC across patients separately for each drug with at least five test observations and then average equally across drugs. This metric directly evaluates within-drug patient ranking. We additionally report Macro Pearson, Global Spearman, Global Pearson, RMSE, and MAE. Random patient–drug pair splitting is evaluated separately to measure the effect of patient overlap between training and test sets.

Baselines. We compare EviDR with Elastic Net and representative drug response models, including CaDRReS (Suphavilai et al., 2018), AITL (Sharifi-Noghabi et al., 2020), CODE-AE (He et al., 2022), WISER (Shubham et al., 2024), TransDRP (Liu & Li, 2025), and MAC-CDR (Zhang et al., 2026a). These baselines cover regularized regression, latent interaction modeling, adversarial transfer, representation disentanglement, weighted source transfer, and drug-conditioned domain adaptation. All baselines are evaluated under the same supervised-transfer protocol using PRISM and BeatAML training responses. BeatAML validation patients are used only for model selection, while held-out test patients are accessed only during final inference and evaluation. All methods use the same patient folds, input features, and model-selection criterion.

4.2 EXPERIMENTAL RESULTS

4.2.1 STRICT EVALUATION AND BASELINE PERFORMANCE

Table 1: Macro Spearman under different evaluation protocols. Results are mean $\pm$ standard deviation over five folds.

Protocol	EviDR	MAC-CDR
Pair-random	0.577 ± 0.014	0.609 ± 0.012
Patient-held-out	0.329 ± 0.040	0.313 ± 0.063

Evaluation Protocol Gap. We first examine how the data-splitting protocol affects the estimated ability to generalize to new patients. Table 1 compares random patient–drug pair splitting with patient-held-out evaluation. Under random pair splitting, EviDR and MAC-CDR achieve Macro Spearman correlations of 0.577 and 0.609, respectively. When patients are completely held out, their performance decreases to 0.329 and 0.313. Random pair splitting therefore increases the estimated Macro Spearman by 0.248 for EviDR and 0.296 for MAC-CDR.

Table 2: Drug response prediction under patient-held-out five-fold evaluation. Results are reported as mean ± standard deviation. Higher correlation is better, while lower RMSE and MAE are better. The best results are shown in bold, and the second-best results are underlined.

Method	Macro Spearman ↑	Macro Pearson ↑	Global Spearman ↑	Global Pearson ↑	RMSE ↓	MAE ↓
Elastic Net	0.220 ± 0.043	0.222 ± 0.044	0.668 ± 0.031	0.668 ± 0.034	48.113 ± 3.162	36.065 ± 1.909
CaDRReS	0.302 ± 0.049	0.295 ± 0.044	0.686 ± 0.019	0.693 ± 0.021	46.913 ± 2.406	34.928 ± 1.509
AITL	0.294 ± 0.057	0.288 ± 0.061	0.685 ± 0.022	0.691 ± 0.022	47.157 ± 2.307	35.073 ± 1.553
CODE-AE	0.308 ± 0.038	0.304 ± 0.042	0.692 ± 0.020	0.698 ± 0.021	46.599 ± 2.409	34.706 ± 1.470
WISER	0.304 ± 0.041	0.305 ± 0.036	0.689 ± 0.017	0.695 ± 0.018	46.962 ± 2.147	34.927 ± 1.434
TransDRP	0.314 ± 0.037	0.309 ± 0.043	0.691 ± 0.022	0.697 ± 0.023	46.782 ± 2.459	34.754 ± 1.556
MAC-CDR	0.313 ± 0.063	0.312 ± 0.066	0.693 ± 0.026	0.700 ± 0.027	46.522 ± 2.454	34.565 ± 1.737
EviDR	**0.329 ± 0.040**	**0.320 ± 0.045**	**0.695 ± 0.021**	**0.701 ± 0.024**	**46.422 ± 2.560**	**34.545 ± 1.677**

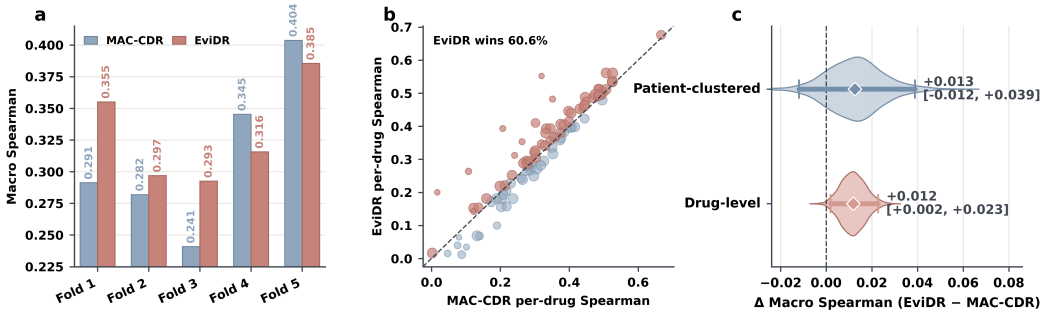

Figure 2: Cross-fold and drug-level consistency of EviDR relative to MAC-CDR. (a) Macro Spearman across five patient-held-out folds. (b) Per-drug Spearman correlations, where points above the diagonal favor EviDR and marker size indicates the number of test observations. (c) Patient-clustered and drug-level bootstrap distributions of the Macro Spearman difference. Diamonds indicate mean differences, horizontal intervals indicate 95% confidence intervals, and the dashed line denotes equal performance.

Pair-random splitting allows observations from the same patient to appear in both training and test sets. Consequently, the model has already observed the molecular profile of a test patient during training, making the prediction task substantially easier. Patient-held-out evaluation instead requires prediction for entirely unseen patients and directly measures the intended cross-domain generalization setting. We therefore use patient-held-out five-fold evaluation for all subsequent comparisons.

Baseline Performance. Table 2 compares EviDR with representative drug-response and cross-domain transfer methods under the same patient-held-out protocol. EviDR achieves the highest mean Macro Spearman of 0.329, compared with 0.314 for TransDRP and 0.313 for MAC-CDR. EviDR also obtains the highest mean Macro Pearson and Global Spearman among the evaluated methods, while its absolute prediction errors remain comparable to the strongest baselines. Descriptions of the baseline methods and their fold-level results are provided in Appendix A.4.

Notably, the Global Spearman correlations of all methods remain relatively high (0.668–0.695), whereas their Macro Spearman correlations are substantially lower (0.220–0.329). For EviDR, Global Spearman reaches 0.695, while Macro Spearman is 0.329. This discrepancy shows that pooled correlation is strongly influenced by response differences across drugs and can substantially overstate within-drug patient-ranking performance. We therefore treat per-drug Macro Spearman as the primary measure of cross-domain generalization throughout the paper.

Cross-Fold and Drug-Level Consistency. We further examine whether the improvement is distributed across evaluation folds and individual drugs. Figure 2 compares EviDR with MAC-CDR at the outer-fold and drug levels and quantifies the uncertainty of their performance difference.

EviDR outperforms MAC-CDR in three of the five outer folds. The largest improvements occur in Folds 1 and 3, where Macro Spearman increases by 0.064 and 0.052, respectively. Among 94 evaluable drugs, EviDR achieves a higher per-drug Spearman for 60.6% of them. Drug-level boot-

strap estimates an average improvement of 0.012 with a 95% confidence interval of $[0.002, 0.023]$, while patient-clustered bootstrap produces a similar mean improvement of 0.013 with an interval of $[-0.012, 0.039]$. Together, the fold-level, drug-level, and bootstrap results show that the overall improvement extends across multiple folds and a majority of drugs.

4.2.2 Transfer Utility and Patient Stratification

Effect of Source Supervision. We examine how the amount of PRISM supervision affects prediction on unseen patients. We fix the BeatAML training, validation, and test patients in Fold 1 and vary the proportion of PRISM observations used by the source-domain training loss from 0% to 100%. The model architecture, patient labels, drug–pathway Evidence, optimization budget, and random seed remain unchanged.

As shown in Figure 3a, performance remains nearly unchanged between 0% and 50% source supervision and then improves substantially at 75% and 100%. Macro Spearman increases from 0.310 at 0% supervision to 0.324 and 0.355 at 75% and 100%, respectively, while Macro Pearson reaches 0.357 with full supervision. Figure 3(b) shows the same pattern in absolute prediction accuracy: full PRISM supervision reduces RMSE from 46.511 to 45.123 and MAE from 35.271 to 34.167. In this evaluated fold, full PRISM supervision yields the strongest patient-side ranking performance and the lowest prediction errors, supporting the value of broad source-domain supervision for the shared predictor.

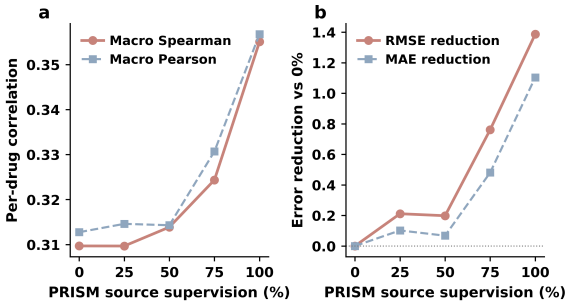

Figure 3: Effect of PRISM supervision on BeatAML Fold 1. Evidence is held fixed across all settings; 0% denotes no direct PRISM response-prediction loss, not the absence of PRISM-derived information. (a) Per-drug correlations. (b) RMSE and MAE reductions relative to 0% supervision.

Sensitive Patient Enrichment. We next test whether EviDR can prioritize sensitive samples for each drug. Within every outer test fold, we rank patient-derived samples by their predicted AUC separately for each drug and select approximately the lowest 20%. We then measure the fraction of selected samples that also belong to the corresponding lowest observed-AUC group. This analysis covers 85 drugs with sufficient test observations. The enrichment calculation is detailed in Appendix A.5.

Table 3: Top-20% sensitive-sample enrichment. For each drug–fold group, $k = \lceil 0.2n \rceil$ samples are selected. Random precision is k/n within each group and is reported as its mean across groups. Enrichment is computed within each group before aggregation.

Selection	Precision	Enrichment
Random	0.205	1.000×
EviDR	**0.366**	**1.790**×

As shown in Table 3, the samples selected by EviDR achieve a mean precision of 0.366, compared with a mean group-specific random expectation of 0.205. Computing enrichment within each drug–fold group before aggregation gives a mean enrichment of 1.790× (95% drug-level bootstrap CI: $[1.706, 1.872]$). These results show that EviDR's within-drug rankings preferentially retrieve previously unseen patient-derived samples with high measured drug sensitivity.

4.2.3 Evidence Analysis and Biological Interpretation

Evidence Stability and Prediction Sensitivity. We examine whether drug–pathway Evidence is stable across PRISM cross-fit estimates and whether the trained predictor responds to its identity and direction. Evidence is re-estimated after excluding each PRISM fold, and the highest-ranked pathways are compared for each drug. As shown in Figure 4a, the Top-5, Top-10, and Top-20 overlaps across 99 drugs are 28.9%, 33.1%, and 38.4%,

compared with random expectations of 0.5%, 1.0%, and 2.0%. Drug-associated pathways therefore recur across source-fold exclusion estimates substantially more often than expected by chance.

We further perturb Evidence while keeping the patient pathway state and molecular drug representation fixed. As shown in Figure 4b, replacing Evidence with that of another drug changes the predicted AUC by 0.503 on average, increasing to 0.731 for high-quality Evidence. High-quality Evidence denotes the upper third of drugs ranked by mean retained Evidence reliability within each outer fold (Appendix A.6). Reversing the Evidence sign produces mean prediction changes of 0.599 and 1.006 for all and high-quality Evidence, respectively. The larger changes for high-quality Evidence indicate that predictions are more sensitive to associations with higher estimated reliability. Together, these results show that drug–pathway Evidence is stable across source-fold exclusion estimates and that both its drug identity and direction affect patient-level predictions. Complementary component ablations are reported in Appendix A.7.

Case Study: Cytarabine. We use cytarabine to examine whether EviDR connects source-derived Evidence with biologically coherent, patient-specific response patterns. We select this drug because it has sufficient held-out observations and a well-characterized mechanism involving DNA synthesis, allowing the complete pathway from Evidence construction to patient-level prediction to be examined. As shown in Appendix Figure 5a, the leading PRISM-derived Evidence is concentrated in G1/S transcription, TP53-regulated cell-cycle arrest, E2F-mediated DNA replication, and DNA synthesis. This profile is consistent with the activity of cytarabine as a nucleoside analog that disrupts DNA synthesis in proliferating cells (Di Francia et al., 2021). In Appendix Figure 5b, P1–P6 denote the six pathways in panel a, ordered from top to bottom. Although the two samples share the same drug-level Evidence, their pathway states produce distinct signed interaction patterns: the sample with lower observed AUC exhibits predominantly negative values, whereas the sample with higher observed AUC exhibits predominantly positive values. These values represent the intermediate interaction supplied to the subsequent encoder, rather than a direct decomposition of the final prediction. As shown in Appendix Figure 5c, EviDR predicts AUC values of 169 and 233 for the two samples, preserving the ordering of their observed values of 156 and 240. The case illustrates how a shared drug-associated pathway profile is combined with sample-specific pathway states to produce distinct interaction representations and response estimates.

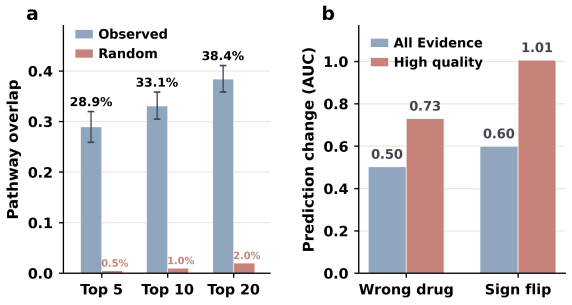

Figure 4: Source-fold stability and prediction sensitivity of drug–pathway Evidence. (a) Top-K pathway overlap across PRISM cross-fit estimates compared with random selection. (b) Mean absolute change in predicted AUC after wrong-drug replacement and sign reversal. Prediction changes are reported on the original BeatAML AUC scale.

5 Conclusion

We study cell-line-to-patient drug response prediction under strict patient-held-out evaluation, where models must generalize to entirely unseen patients and rank their responses to the same drug. Our results show that random patient–drug splitting and pooled global metrics can substantially overestimate this ability. We introduce EviDR, which transfers large-scale PRISM supervision through a shared Reactome pathway space and reliability-weighted drug–pathway Evidence. EviDR improves per-drug patient ranking over existing cross-domain baselines, benefits from increasing source supervision, and enriches experimentally sensitive patient-derived samples. Further analyses show that drug-associated Evidence is stable under PRISM fold exclusion and predictions are sensitive to Evidence identity and direction, while the cytarabine case connects shared drug Evidence with patient-specific pathway states and response estimates. Together, EviDR provides a biologically structured framework for generalizable and traceable patient drug response prediction.

REFERENCES

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

Judith Bernett, Pascal Iversen, Mario Picciani, Mathias Wilhelm, Katharina Baum, and Markus List. Critical evaluation of drug response prediction models with dreval. *Nature Communications*, 17 (1):4238, 2026.

Junyi Chen, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu, and Qin Ma. Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data. *Nature communications*, 13(1):6494, 2022.

Steven M Corsello, Rohith T Nagari, Ryan D Spangler, Jordan Rossen, Mustafa Kocak, Jordan G Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A Tang, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature cancer*, 1(2): 235–248, 2020.

James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-Ud-Din, Petteri Hintsanen, Suleiman A Khan, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202–1212, 2014.

Raffaele Di Francia, Stefania Crisci, Angela De Monaco, Concetta Cafiero, Agnese Re, Giancarla Iaccarino, Rosaria De Filippi, Ferdinando Frigeri, Gaetano Corazzelli, Alessandra Micera, et al. Response and toxicity to cytarabine therapy in leukemia and lymphoma: from dose puzzle to pharmacogenomic biomarkers. *Cancers*, 13(5):966, 2021.

Di He, Qiao Liu, You Wu, and Lei Xie. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, 4(10):879–892, 2022.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Iljung Jin and Hojung Nam. Hidra: hierarchical network for drug response prediction with attention. *Journal of Chemical Information and Modeling*, 61(8):3858–3867, 2021.

Brent M Kuenzi, Jisoo Park, Samson H Fong, Kyle S Sanchez, John Lee, Jason F Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer cell*, 38(5):672–684, 2020.

Patrick J Lawrence, Benjamin Burns, and Xia Ning. Enhancing drug and cell line representations via contrastive learning for improved anti-cancer drug prioritization. *NPJ Precision Oncology*, 8 (1):106, 2024.

Kun Li, Jia Wu, Bo Du, Sergey V Petoukhov, Huiting Xu, Zheman Xiao, and Wenbin Hu. Towards a better model with dual transformer for drug response prediction. *arXiv preprint arXiv:2210.17401*, 2022.

Kun Li, Xiuwen Gong, Jia Wu, and Wenbin Hu. Contrastive learning drug response models from natural language supervision. In *33rd International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence, 2024.

Yihui Li, David Earl Hostallero, and Amin Emad. Interpretable deep learning architectures for improving drug response prediction performance: myth or reality? *Bioinformatics*, 39(6):btad390, 2023.

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 2020.

Xuan Liu and Menglu Li. Knowledge-guided domain adaptation model for transferring drug response prediction from cell lines to patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 523–531, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jianzhu Ma, Samson H Fong, Yunan Luo, Christopher J Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk FA Wessels, Marc Hafner, Roded Sharan, Jian Peng, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021.

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.

Soufiane MC Mourragui, Marco Loog, Daniel J Vis, Kat Moore, Anna G Manjon, Mark A van de Wiel, Marcel JT Reinders, and Lodewyk FA Wessels. Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proceedings of the National Academy of Sciences*, 118(49):e2106682118, 2021.

Tuan Nguyen, Giang TT Nguyen, Thin Nguyen, and Duc-Hau Le. Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(1):146–154, 2021.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Hossein Sharifi-Noghabi, Shuman Peng, Olga Zolotareva, Colin C Collins, and Martin Ester. Aitl: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*, 36(Supplement_1):i380–i388, 2020.

Jihye Shin, Yinhua Piao, Dongmin Bang, Sun Kim, and Kyuri Jo. Drpreter: interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. *International Journal of Molecular Sciences*, 23(22):13919, 2022.

Kumar Shubham, Aishwarya Jayagopal, Syed Mohammed Danish, Prathosh Ap, and Vaibhav Rajan. Wiser: Weak supervision and supervised representation learning to improve drug response prediction in cancer. *arXiv preprint arXiv:2405.04078*, 2024.

Nicolas Stransky, Mahmoud Ghandi, Gregory V Kryukov, Levi A Garraway, Joseph Lehar, Manway Liu, Dmitriy Sonkin, Audrey Kauffmann, Kavitha Venkatesan, Elena J Edelman, et al. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84, 2015.

Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. Predicting cancer drug response using a recommender system. *Bioinformatics*, 34(22):3907–3914, 2018.

Yi-Ching Tang and Assaf Gottlieb. Explainable drug sensitivity prediction through cancer pathway enrichment. *Scientific reports*, 11(1):3128, 2021.

Yi-Ching Tang, Reid T Powell, and Assaf Gottlieb. Molecular pathways enhance drug response prediction using transfer learning from cell lines to tumors and patient-derived xenografts. *Scientific Reports*, 12(1):16109, 2022.

Jeffrey W Tyner, Cristina E Tognon, Daniel Bottomly, Beth Wilmot, Stephen E Kurtz, Samantha L Savage, Nicola Long, Anna Reister Schultz, Elie Traer, Melissa Abel, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*, 562(7728):526–531, 2018.

Yang Wu, Ming Chen, and Yufang Qin. Anticancer drug response prediction integrating multi-omics pathway-based difference features and multiple deep learning techniques. *PLOS Computational Biology*, 21(3):e1012905, 2025.

Xiaoqiong Xia, Chaoyu Zhu, Fan Zhong, and Lei Liu. Transcdr: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. *BMC biology*, 22(1):227, 2024.

Licai Zhang, Xiao Kang, Xinxing Yang, Lin Wang, Genke Yang, and Jian Chu. Memory-anchored multimodal cross-domain adaptation for drug response prediction from cell lines to patients. *Pattern Recognition*, pp. 114142, 2026a.

Yuanpeng Zhang, Zhijian Huang, Ziyu Fan, Siyuan Shen, Yahan Li, Shangqian Wu, Min Wu, and Lei Deng. Deepsadr: Deep transfer learning with subsequence interaction and adaptive readout for cancer drug response prediction. In *International Conference on Learning Representations*, volume 2026, pp. 116485–116513, 2026b.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and language polishing, including sentence rephrasing, grammar correction, and improvements to readability, clarity, and textual flow. These tools were not used to generate research ideas, develop the methodology, design experiments, generate or analyze research data, or interpret experimental results. All scientific contributions and research decisions were made by the authors.

The authors reviewed and revised all AI-assisted text to ensure technical accuracy and consistency with the methods, results, and conclusions of this work. We take full responsibility for the final content of the manuscript, including any text produced or revised with the assistance of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No new human-subject or animal experiments were conducted in this study. All datasets were obtained and used in accordance with the relevant licenses, access conditions, and usage guidelines. No personally identifiable information was used, and no experiments raised privacy or security concerns. We have taken care to avoid potential bias or discriminatory outcomes and are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including the training procedure, model configurations, data partitions, and evaluation protocol, is described in detail in the paper. The code, experiment configurations, and processed data required to reproduce the reported results are included in the supplementary materials. We believe these measures will enable other researchers to reproduce our work and further advance research in this field.

A APPENDIX

A.1 ARCHITECTURE AND OPTIMIZATION DETAILS

A.1.1 MODEL ARCHITECTURE

EviDR uses the same encoders and prediction head for PRISM and BeatAML. The final model contains 766,989 trainable parameters and operates on 1,000-dimensional Reactome pathway activities and 512-bit Morgan fingerprints. All hidden representations have dimension 128. Table 4 provides the complete architecture.

The sample encoder maps the pathway-activity vector to a 128-dimensional representation. The drug encoder uses two linear layers followed by output LayerNorm. The adaptive prior-weight network takes the drug embedding together with four prior statistics: the fractions of positive entries and the

Table 4: Architecture of the final EviDR model.

Component	Input	Architecture	Output
Sample encoder	1,000	Linear(1000,128)–GELU–Dropout(0.3)	128
Drug encoder	512	Linear(512,128)–GELU–Dropout(0.3) Linear(128,128)–LayerNorm	128
Prior-weight network	132	Linear(132,32)–GELU–Linear(32,2)	2
Learned pathway gate	128	Linear(128,1000)	1,000
Interaction encoder	1,000	Linear(1000,128)–GELU–Dropout(0.3) Linear(128,128)	128
Fusion MLP	384	Linear(384,128)–GELU–Dropout(0.3) Linear(128,128)–GELU	128
Prediction head	128	Linear(128,1)	1
Pathway auxiliary head	128	Linear(128,1000)	1,000
MOA auxiliary head	128	Linear(128,N_{MOA})	N_{MOA}

mean scores of the direct and neighboring prior vectors. It produces drug-specific weights for the two prior sources, bounded above by 4 and 2 and initialized at 2 and 1, respectively.

As defined in Section 3.3, reliability-weighted Evidence relevance is added to the drug-conditioned pathway gate, while signed Evidence is combined with the gated sample state. The resulting interaction is encoded into 128 dimensions. The sample, drug, and interaction representations are then concatenated and processed by the Fusion MLP and a shared linear prediction head.

A.1.2 Joint Training and Optimization

Each optimization step contains one PRISM mini-batch and one BeatAML mini-batch, with batch sizes of 64 and 32, respectively. Half of each PRISM mini-batch is sampled from drugs also observed in BeatAML, while the remaining half is sampled from the broader source collection. This sampling scheme retains broad source supervision while increasing the frequency of drugs relevant to the target task.

Response labels are standardized using statistics estimated only from the corresponding training split. For each domain–drug group with at least five valid training observations, the normalized response is defined as $\widetilde{y}_{i,d} = (y_{i,d} - \mu_d^{(\mathrm{domain})})/\sigma_d^{(\mathrm{domain})}$. Groups with fewer than five observations or invalid variance use the corresponding domain-level mean and standard deviation. Model outputs are transformed back to the original AUC scale before evaluation.

For each domain, the regression objective assigns equal weight to the quality-weighted row-level Huber loss and the drug-balanced Huber loss. The within-drug ranking-loss weight is 0.25 for BeatAML and 0.0625 for PRISM. Only pairs receiving the same drug and satisfying

$$|\widetilde{y}_{i,d} - \widetilde{y}_{j,d}| > 0.05$$

are included in the ranking objective. Thus, 0.05 is a pairwise normalized-AUC difference threshold rather than a ranking margin. The pathway and mechanism-of-action auxiliary losses have weights of 0.05 and 0.02, respectively, and the adaptive-prior regularization weight is 0.001. The PRISM and BeatAML domain objectives are combined with weights of 0.5 and 1.0.

We optimize EviDR using AdamW for at most 60 epochs (Loshchilov & Hutter, 2017), with 500 optimization steps per epoch. The learning rate is reduced by a factor of 0.5 after three validation epochs without improvement. Gradients are clipped to a maximum norm of 5.0. Checkpoint selection is based exclusively on BeatAML validation Macro Spearman, and training stops after 18 consecutive epochs without improvement. The selected checkpoint is evaluated once on the corresponding patient-held-out test fold. All five folds use random seed 2026 and the same configuration. The complete training and optimization settings are summarized in Table 5.

The pathway and mechanism auxiliary heads use positive-weighted binary cross-entropy, with targets derived from the direct drug–pathway prior and retained mechanism annotations, respectively. The adaptive prior weights in equation 8 are regularized toward their initial values of 2 and 1. BeatAML observations are randomly sampled in mini-batches, and ranking pairs are formed only

Table 5: Training and optimization settings used for EviDR.

Setting	Value	Setting	Value
Trainable parameters	766,989	Hidden dimension	128
Optimizer	AdamW	Dropout	0.3
Learning rate	1×10^{-4}	Weight decay	1×10^{-4}
PRISM batch size	64	BeatAML batch size	32
Steps per epoch	500	Maximum epochs	60
Scheduler factor	0.5	Scheduler patience	3
Early-stopping patience	18	Gradient clipping	5.0
PRISM loss weight	0.5	BeatAML loss weight	1.0
Row/drug Huber mixture	0.5 / 0.5	Pairwise AUC-gap threshold	0.05
PRISM ranking weight	0.0625	BeatAML ranking weight	0.25
Pathway auxiliary weight	0.05	MOA auxiliary weight	0.02
Prior regularization	0.001	Random seed	2026

between samples receiving the same drug. Pairs with a standardized AUC difference no greater than 0.05 are excluded; if no valid pair is available, the ranking loss is set to zero.

A.2 Evidence Estimation Details

Drug–pathway Evidence is estimated exclusively from PRISM training data. For each drug d and Reactome pathway k, we remove cancer-lineage means from pathway activities and quality-weighted response AUCs, and compute

$$E_{d,k} = \mathrm{Corr}_w\left(\widetilde{x}_k, \widetilde{y}_d\right). \tag{17}$$

Missing responses receive zero weight. Associations are estimated only for drugs with an effective sample size of at least 20, where

$$n_d^{\mathrm{eff}} = \frac{\left(\sum_i w_{i,d}\right)^2}{\sum_i w_{i,d}^2}. \tag{18}$$

The analytical reliability of each association is

$$R_{d,k}^{\mathrm{analytic}} = \left[1 - \exp\left(-\frac{|\operatorname{arctanh}(E_{d,k})|\sqrt{\max(n_d^{\mathrm{eff}} - 3, 0)}}{3}\right)\right] \min\left(\frac{n_d^{\mathrm{eff}}}{20}, 1\right). \tag{19}$$

PRISM cell lines are assigned to five deterministic cross-fitting folds using random seed 2026. A source observation in fold q receives Evidence estimated from the remaining folds, preventing its response from contributing to its own Evidence. BeatAML observations receive full-fit Evidence estimated from all available PRISM training samples, without using BeatAML response labels.

For full-fit Evidence, analytical reliability is further weighted by the stability of the five cross-fit estimates:

$$R_{d,k}^{\mathrm{full}} = R_{d,k}^{\mathrm{analytic,full}} \left|\frac{1}{5}\sum_{q=1}^{5} \mathrm{sign}\left(E_{d,k}^{(-q)}\right)\right| \exp\left[-\frac{\mathrm{Std}_q\left(E_{d,k}^{(-q)}\right)}{0.25}\right]. \tag{20}$$

Associations with reliability below 0.5 are removed. For each drug, at most 64 pathways with the largest $|E_{d,k}|R_{d,k}$ are retained. The resulting relevance and signed components are $r_d = |E_d| \odot R_d$ and $s_d = E_d \odot R_d$, respectively. Table 6 summarizes the fixed parameters used to construct cross-fitted Evidence in all experiments.

A.3 Data Processing and Cross-Domain Matching

Drug-response processing. For PRISM, we retain secondary-screen records with valid AUC values, mapped expression profiles, and successful STR-profiling quality control. When a cell-line–drug pair appears in multiple screens, we prioritize MTS010, followed by MTS006, MTS005,

Table 6: Parameters used for Evidence estimation.

Setting	Value	Setting	Value
Association	Weighted Pearson	Lineage adjustment	Enabled
Cross-fitting folds	5	Random seed	2026
Minimum effective size	20	Reliability threshold	0.5
Maximum pathways per drug	64	Variability scale	0.25

and HTS002, and average replicates within the selected screen. Each observation is weighted by $0.5 + 0.5\ \mathrm{clip}(R^2, 0, 1)$, where R^2 measures dose–response fit quality. For BeatAML, we retain converged, paper-included single-agent curves with valid patient and RNA-sample identifiers. Replicates for the same patient-sample–drug combination are averaged; decreasing curves receive weight 1.0 and retained increasing curves receive weight 0.25.

Cross-domain drug matching. Drug names are converted to lowercase, parenthetical development codes are removed, and non-alphanumeric characters are discarded. Candidate PRISM–BeatAML matches are then checked using the drug-identity audit table, and only accepted mappings are retained. This procedure yields 99 drugs shared between the two domains. Cleaned SMILES strings are encoded as 512-bit radius-2 Morgan fingerprints using RDKit; all drugs in the processed dataset were successfully encoded without fallback representations.

Expression and pathway representation. Gene symbols are standardized to uppercase, duplicate entries are averaged, and genes are restricted to those measured in both expression datasets. PRISM and BeatAML expression profiles are standardized separately using gene-wise means and standard deviations estimated from the corresponding training samples. For domain $a \in \{s, t\}$, the standardized expression of gene g is

$$z_{i,g}^{(a)} = \frac{v_{i,g}^{(a)} - \mu_g^{(a)}}{\sigma_g^{(a)}}. \tag{21}$$

Human Reactome pathways containing 5–300 shared genes are retained, with target-associated and more specific pathways prioritized. The initial activity of pathway k is computed as

$$\bar{x}_{i,k}^{(a)} = \frac{1}{|G_k|} \sum_{g \in G_k} z_{i,g}^{(a)}, \tag{22}$$

where G_k denotes the genes assigned to pathway k. The resulting pathway activities are standardized again using pathway-level means and standard deviations estimated from the training samples of each domain:

$$x_{i,k}^{(a)} = \frac{\bar{x}_{i,k}^{(a)} - \mu_{k,\mathrm{path}}^{(a)}}{\sigma_{k,\mathrm{path}}^{(a)}}. \tag{23}$$

Validation and test samples are transformed using the corresponding training-fitted parameters. The final representation contains 1,000 pathways covering 7,627 genes.

Drug–pathway priors. Known drug targets are mapped to the selected Reactome pathways to construct a direct prior h_d^{dir}. One-hop neighbors in the Reactome relation graph define a separate neighboring prior h_d^{nbr}, with each neighboring pathway inheriting the maximum score from its adjacent direct pathways. The two prior vectors remain separate and are adaptively combined using the drug-specific weights defined in equation 8. The direct and neighboring weights are initialized at 2 and 1, respectively, and are updated during training.

A.4 Baseline Methods and Fold-Level Results

A.4.1 Baseline Methods

We compare EviDR with representative statistical, drug–sample interaction, and cross-domain transfer methods. The selected baselines cover the main modeling strategies used in drug response prediction.

Table 7: Fold-level Macro Spearman under patient-held-out evaluation. The final column reports mean ± standard deviation across the five folds.

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± Std.
Elastic Net	0.229	0.183	0.203	0.194	0.290	0.220 ± 0.043
CaDRReS	0.331	0.280	0.269	0.256	0.373	0.302 ± 0.049
AITL	0.284	0.270	0.246	0.277	0.393	0.294 ± 0.057
CODE-AE	0.326	0.293	0.265	0.292	0.364	0.308 ± 0.038
WISER	0.296	0.292	0.257	0.307	0.368	0.304 ± 0.041
TransDRP	0.342	0.294	0.270	0.302	0.362	0.314 ± 0.037
MAC-CDR	0.291	0.282	0.241	**0.345**	**0.404**	0.313 ± 0.063
EviDR	**0.355**	**0.297**	**0.293**	0.316	0.385	**0.329 ± 0.040**

Table 8: Reproducibility of drug–pathway Evidence across PRISM cross-fit subsets. Confidence intervals are obtained by bootstrapping drugs for 5,000 iterations.

Pathway set	Observed overlap	95% CI	Random overlap
Top-5	28.9%	[25.9%, 32.0%]	0.5%
Top-10	33.1%	[30.5%, 35.8%]	1.0%
Top-20	38.4%	[35.8%, 41.1%]	2.0%

- **Elastic Net** is a regularized linear regression model combining ℓ_1 and ℓ_2 penalties. It provides a conventional statistical baseline for predicting drug response from concatenated sample and drug features.
- **CaDRReS** learns low-dimensional representations of biological samples and drugs in a shared latent space, where their interaction is used to predict drug response.
- **AITL** uses adversarial inductive transfer learning to extract representations shared between cell-line and patient domains, while adapting the response predictor to limited target-domain labels.
- **CODE-AE** separates common biological signals from domain-specific variation through shared and private representation components, supporting transfer from cell lines to patient-derived samples.
- **WISER** combines supervised domain-invariant representation learning with weak supervision and subset selection to incorporate pseudo-labeled patient profiles into drug-response prediction.
- **TransDRP** combines biological knowledge with adversarial domain adaptation to learn drug response representations transferable from cancer cell lines to patient samples.
- **MAC-CDR** performs drug-conditioned cross-domain adaptation using memory anchors and multimodal compound representations to connect cell-line and patient response patterns.

DeepSADR is not included in Table 2 because its published formulation targets binary clinical response classification, whereas adapting it to continuous BeatAML AUC regression and within-drug ranking would substantially alter its original objective and training pipeline. Together, these baselines cover regularized regression, latent drug–sample interaction, adversarial transfer, shared–private representation learning, weak supervision and subset selection, and drug-conditioned domain adaptation.

A.4.2 Fold-Level Results

Table 7 reports Macro Spearman separately for the five patient-held-out test folds. EviDR achieves the highest score in the first three folds, while MAC-CDR performs best in Folds 4 and 5. Averaged across all folds, EviDR obtains the highest Macro Spearman of 0.329. These results also show that performance varies across patient folds, motivating the use of five-fold evaluation rather than a single train–test split.

Table 9: Mean absolute prediction changes under Evidence perturbations. All values are reported on the original BeatAML AUC scale.

Perturbation	All retained Evidence	High-quality Evidence
Wrong-drug replacement	0.503	0.731
Sign reversal	0.599	1.006

For Figure 2, we pool the patient-held-out test predictions from all five outer folds. In the patient-clustered bootstrap, 479 patients are sampled with replacement and all patient–drug observations belonging to each sampled patient are retained; Macro Spearman is recomputed for both methods and their difference is recorded. In the drug-level bootstrap, paired per-drug Spearman differences are computed for the 94 evaluable drugs, after which drugs are sampled with replacement and the mean difference is recorded. Both analyses use 5,000 iterations with random seed 2026.

A.5 Sensitive-Sample Enrichment

Sensitive-sample enrichment is calculated independently for each drug within every outer test fold. Fold–drug groups with fewer than 25 observations are excluded. For a group containing n observations, we define

$$k = \lceil 0.2n \rceil . \tag{24}$$

The k samples with the lowest predicted AUC are selected as predicted responders, while the k samples with the lowest observed AUC form the experimentally measured responder set. Precision and enrichment are computed as

$$\text{Precision} = \frac{|\mathcal{S}_{\text{pred}} \cap \mathcal{S}_{\text{obs}}|}{k}, \qquad \text{Enrichment} = \frac{\text{Precision}}{k/n}. \tag{25}$$

Here, k/n is the expected precision under random selection. Enrichment greater than one therefore indicates that samples prioritized by the model contain experimentally sensitive responders more frequently than expected by chance.

For drugs appearing in multiple outer test folds, fold-level precision and enrichment are first averaged within each drug. The final analysis contains 85 drugs. The 95% confidence interval is estimated by resampling drugs with replacement for 5,000 iterations using random seed 2026.

A.6 Additional Evidence Analysis Details

A.6.1 Reproducibility across PRISM Cross-Fit Subsets

We evaluate whether the drug–pathway associations recovered from PRISM remain consistent when different source samples are excluded. For each of the 99 drugs shared with BeatAML, we construct five Evidence profiles using the PRISM cross-fitting procedure. Each profile is represented by the reliability-weighted association $E_d^{(-q)} \odot R_d^{(-q)}$. For every pair of cross-fit profiles, we select the pathways with the largest absolute values and calculate the fraction shared between their Top-K sets. Pairwise overlaps are first averaged within each drug and then across drugs.

Under random pathway selection, the expected overlap is $K/1000$ because the shared representation contains 1,000 Reactome pathways. As shown in Table 8, the observed Top-5, Top-10, and Top-20 overlaps are 28.9%, 33.1%, and 38.4%, respectively, compared with random expectations of 0.5%, 1.0%, and 2.0%. The consistently higher observed overlaps show that source-fold exclusion estimates repeatedly prioritize related pathways for the same drug.

A.6.2 Prediction Response to Evidence Perturbations

We next test whether the trained predictor responds to the Evidence assigned to each drug. These experiments are performed at inference time without retraining the model. For every BeatAML test patient–drug pair, the patient pathway state and molecular drug representation are fixed, while only

Table 10: Ablation results under patient-held-out five-fold evaluation. Evidence-guided target uses PRISM-derived Evidence but omits direct training on PRISM response observations. Pathway-shuffled Evidence permutes pathway identities within each drug while preserving its Evidence values and reliability. Results are reported as mean ± standard deviation. The best results are shown in bold.

a. Source-domain utilization

Variant	Macro Spearman ↑	Macro Pearson ↑
Target-only	0.319 ± 0.037	0.312 ± 0.043
Evidence-guided target	0.321 ± 0.041	0.312 ± 0.045
Full EviDR	$\mathbf{0.329 \pm 0.040}$	$\mathbf{0.320 \pm 0.045}$

b. Evidence component diagnostics

Variant	Macro Spearman ↑	Macro Pearson ↑
No Reliability	0.327 ± 0.038	0.320 ± 0.043
No Relevance	0.328 ± 0.039	0.320 ± 0.044
No Signed Evidence	0.328 ± 0.039	0.320 ± 0.044
Pathway-shuffled Evidence	0.328 ± 0.040	0.319 ± 0.044
Full EviDR	$\mathbf{0.329 \pm 0.040}$	$\mathbf{0.320 \pm 0.045}$

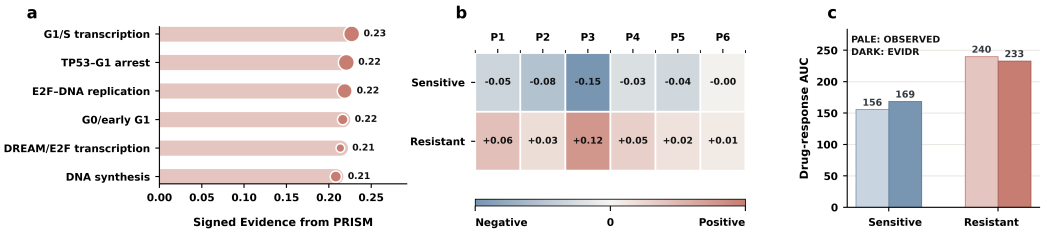

Figure 5: Cytarabine case illustration. (a) Leading source-derived drug–pathway associations, with marker size indicating Evidence reliability. (b) Sample-specific signed interaction values, where P1–P6 correspond to the pathways in panel (a); blue and red indicate negative and positive values, respectively. (c) Observed and EviDR-predicted AUC values for the same samples. Lower AUC indicates greater experimental drug sensitivity.

the Evidence input is modified. Wrong-drug replacement assigns each test drug the complete Evidence and reliability profile of another evaluated drug. Sign reversal multiplies the signed Evidence by -1 while retaining its reliability. We then measure the absolute difference between the original and perturbed predicted AUC; the experimentally observed AUC is not changed or used in this calculation.

We report results for all drugs with retained Evidence and for the high-quality subset. Within each outer fold, high-quality drugs are defined as the upper third according to their mean retained Evidence reliability. As shown in Table 9, wrong-drug replacement changes predicted AUC by 0.503 on average, increasing to 0.731 for high-quality Evidence. Sign reversal produces mean changes of 0.599 and 1.006, respectively. The larger response in the high-quality subset is consistent with the reliability mechanism: Evidence receiving stronger support has a greater effect on the learned patient–drug interaction.

A.7 Source-Utilization Ablations and Evidence Diagnostics

Table 10a separates three levels of source-domain utilization. Target-only training achieves a Macro Spearman of 0.319, while introducing PRISM-derived Evidence without direct optimization on PRISM responses gives 0.321. Full EviDR combines Evidence with joint source–target training and achieves the highest mean performance, with a Macro Spearman of 0.329 and Macro Pearson

of 0.320. The largest improvement therefore comes from direct PRISM response supervision, with Evidence providing pathway-structured, drug-conditioned context within the full model.

Table 10b evaluates the Evidence design while retaining direct PRISM supervision. All component variants achieve similar performance, with Macro Spearman ranging from 0.327 to 0.328; the full formulation obtains the highest mean. The perturbation analysis in Figure 4 provides a complementary view: prediction changes are larger for high-quality Evidence when its drug identity or direction is altered, indicating that its influence is drug-dependent and is not fully summarized by dataset-level ablation averages. Together, the analyses identify source supervision as the main contributor to overall accuracy and Evidence as the structured interface through which drug-specific pathway associations enter patient-level interaction representations.

A.8 Cytarabine Case Visualization

Figure 5 provides the complete visualization of the cytarabine case study discussed in the main text, tracing the prediction from PRISM-derived pathway Evidence to patient-specific pathway interactions and response estimates.