GENERALIZING DRUG RESPONSE PREDICTION THROUGH EVIDENCE-GROUNDED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate drug response prediction for previously unseen drugs and cell lines is a central challenge in computational pharmacology. In this setting, a few measured responses provide useful context, but integrating these observations with molecular properties and cellular characteristics remains difficult. We propose a framework for generalizable drug response prediction through evidence-grounded reasoning. The framework uses a large language model to combine molecular and cellular representations with biological facts and a small set of observed responses. To connect these heterogeneous inputs to prediction, we construct chain-of-thought supervision from available drug and cell facts and support-response summaries, without using the query response in reasoning construction. A multi-stage training strategy first aligns biological inputs through response-free pretraining, then initializes evidence-grounded reasoning jointly with response prediction through a teacher-guided warmup. Subsequent numerical prediction updates alternate with reinforcement learning using REINFORCE, where the predictive loss relative to a reference reasoning trace guides the optimization of generated reasoning. A probability-weighted numerical token interface produces the response estimate. This links reasoning to its usefulness for response prediction. At inference, the framework uses the available biological evidence and measured responses to generate a reasoning trace and estimate the response of an unmeasured drug–cell pair. Across cell-blind and drug-blind GDSC2 evaluation with zero or five measurements, the framework records the lowest macro RMSE in the reported comparison, while contextual measurements improve both RMSE and PCC in each protocol.

1 INTRODUCTION

Drug response prediction (DRP) supports compound prioritization and the study of variation in treatment response. Pharmacogenomic screens connect molecular and cellular characteristics to drug sensitivity (Yang et al., 2013), but useful prediction must extend beyond the entities represented in training. Biological heterogeneity creates unfamiliar cellular contexts, while drug development introduces new compounds, making unseen-drug and unseen-cell generalization central requirements. Deep learning has advanced DRP by combining drug structure with molecular profiles: graph-based models such as DeepCDR and GraphDRP encode chemical connectivity, while DeepTTA uses a transformer-based drug representation (Liu et al., 2020; Nguyen et al., 2022; Jiang et al., 2022). An unseen entity requires transferring relationships learned from other compounds or cellular backgrounds, beyond predicting new pairs of familiar entities.

Comprehensive response profiles are costly to obtain, although a few initial assays may be feasible. This creates a valuable few-shot setting: combine available biological information with a small set of measured responses to predict unmeasured pairs (Ma et al., 2021). A practical framework should support both prediction before any measurements are available and adaptation as observations accumulate.

Few-shot learning makes this transition an explicit learning objective. Model-agnostic meta-learning trains an initialization that can adapt with a small number of examples (Finn et al., 2017); in DRP, TCRP studies transfer from high-throughput screens to patient-related settings (Ma et al., 2021), and metaDRP combines biological prior knowledge with fast adaptation across drug–tissue tasks (Guo

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.

et al., 2025). A few measurements can anchor sensitivity for a new entity. The modeling challenge is to combine this empirical context with biological information available before testing.

Large language models (LLMs) offer a common contextual interface for molecular representations, cellular profiles, and biological facts. Their pretraining and few-shot capabilities motivate transferring this heterogeneous knowledge to new drug–cell contexts (Brown et al., 2020). In-context learning (ICL) further allows measured responses to inform predictions without updating model weights. Unlike conventional fixed-input DRP regressors, this interface directly accepts a changing set of labeled examples. Related in-context modeling for drug synergy (Edwards et al., 2023) motivates its use for response prediction. Each measured drug–cell pair and its response enter the query context. Facts describe the entities, while measurements constrain their observed behavior, allowing a shared predictor to adapt as the assay budget changes.

Chain-of-thought (CoT) generation adds an analytical layer (Wei et al., 2022): a trace can organize biological facts, relate them to measurements, and propose conditional hypotheses for inspection or follow-up experiments. We ground these analytical leads in available inputs and train their generation using response-prediction feedback. The model learns to read biological representations, associate measurements with their entities, and return a numerical estimate. This motivates biological alignment, joint warmup, and subsequent prediction-guided generation.

We propose P2R-LLM, an LLM framework for generalizable DRP through evidence-grounded reasoning (Figure 1). Here, evidence-grounded means that generation is conditioned on the supplied biological facts and measured responses. Response-free biological alignment first trains input reading; teacher-guided warmup then initializes reasoning and quantitative prediction. Subsequent training combines numerical updates with prediction-guided REINFORCE. A probability-weighted numerical token interface converts the reasoning-conditioned state into a continuous response estimate. At inference, the model uses biological inputs and optional measurements to generate a trace and predict the query response.

Our contributions are threefold. First, we introduce a contextual few-shot DRP formulation that jointly represents biological information, measured responses, and an unseen query within an LLM, enabling adaptation through in-context observations with fixed inference weights. Second, we develop a prediction-guided reasoning framework: biological alignment and teacher warmup initialize the interface, and reference-relative REINFORCE links generated comparisons to quantitative prediction through the O100 expectation readout. Third, we evaluate unseen drugs and cells with zero or five measurements, reporting lower RMSE with contextual measurements in both protocols and the lowest macro RMSE in all four reported comparisons. Cell-blind ablations examine biological alignment, the trained reasoning pathway, and the shaped actor update, with the complete recipe producing the lowest errors at both support settings.

2 Related Work

Drug response models and few-shot adaptation. DeepCDR integrates drug graphs and cellular information for response prediction (Liu et al., 2020). Related architectures in our comparison pool include GraphDRP and PASO (Nguyen et al., 2022; Wu et al., 2025). More recent models study substructure–pathway attention, pathway-guided graph convolution, and higher-order topology (Han et al., 2026; Yoon & Lee, 2026; Shen et al., 2026). TCRP establishes the relevance of few-shot adaptation across drug-response domains (Ma et al., 2021). SynerGPT studies in-context learning for drug synergy and drug design (Edwards et al., 2023). Our setting concerns single-agent continuous response, fixed same-entity support panels, and generated comparisons whose utility is evaluated within the prediction computation. This formulation couples the support-conditioned prediction with a learned reasoning pathway for quantitative DRP.

Language models for biological response prediction. DRPLLM extracts last-layer Llama representations from biological and chemical prompts and trains numerical regressors, including an MLP and a DNN (Gnanaolivu et al., 2025). SensitiveCancerGPT also studies generative language models on structured omics for drug sensitivity (Chowdhury et al., 2025). P2R-LLM develops a representation conditioned on the current support observations and on an explicitly generated comparison, with response-stage adapter training and an outcome-based generation objective. The response is

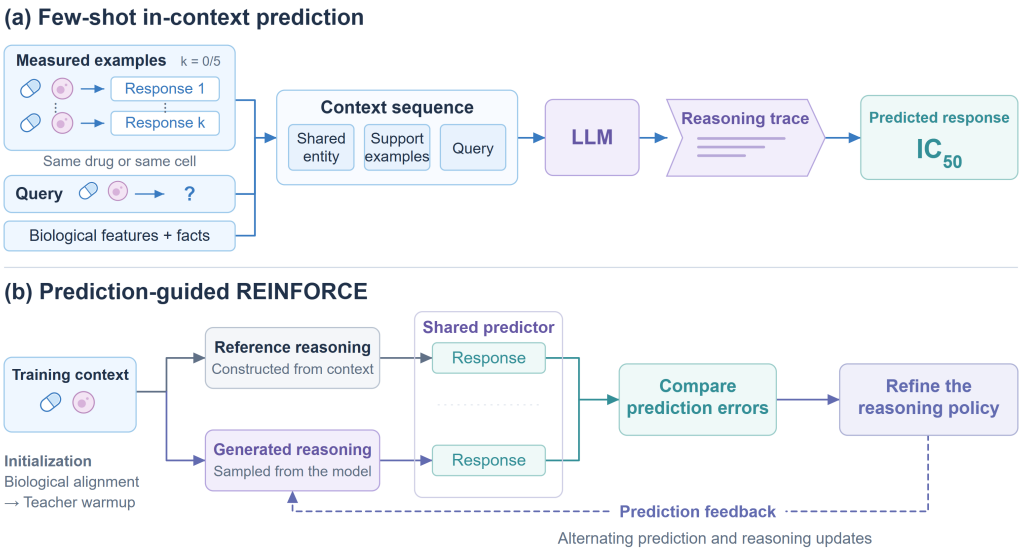

Figure 1: **Overview of P2R-LLM.** Biological inputs and optional measured responses form the context for predicting an unseen drug–cell pair. A generated reasoning trace precedes the response estimate. During training, reference and sampled traces are evaluated by the shared predictor; their relative prediction errors guide the REINFORCE update. The numerical output uses LN_IC50, and support values are training-fitted bin representatives.

read as an expectation over numerical tokens from the reasoning-conditioned state, connecting text generation to a quantitative objective.

Intermediate text and parameter-efficient adaptation. Chain-of-thought prompting uses intermediate text to support downstream answers (Wei et al., 2022). We use generated text as a prefix to continuous regression, and train its policy with a score-function objective rather than backpropagating through sampled token identities. LoRA supplies the parameter-efficient adaptation mechanism (Hu et al., 2022); KPGT supplies the offline molecular representation (Li et al., 2023). Our framework brings these input and adaptation mechanisms into a single contextual prediction process, using response-error feedback to optimize the generated intermediate text for quantitative DRP.

3 METHOD

3.1 DRUG RESPONSE PREDICTION WITH CONTEXTUAL MEASUREMENTS

Let $q = (d, c)$ denote a drug–cell query and y_q its response in LN_IC50 units, where lower values indicate greater sensitivity. We consider drug-blind and cell-blind prediction, holding out the target drug or cell from response training. A separate model is trained for each protocol. For an unseen entity e, a fixed panel $S_k(e)$ contains either no measurements ($k = 0$) or five measured pairs ($k = 5$). Supports share the target drug in drug-blind prediction and the target cell in cell-blind prediction. The five support pairs and their structure–cell aliases are excluded from the scored queries in both settings. Panel construction does not use response values.

The model receives biological evidence E_q and the available measurements as context $x = \text{Serialize}(E_q, S_k(e))$. It generates a reasoning trace $r \sim \pi_\theta(\cdot \mid x)$ and predicts $\widehat{y} = F_\theta(x, r)$. This interface addresses two complementary needs: biological evidence supports prediction before measurements are available, and contextual examples allow adaptation to new measurements with fixed model weights.

Instruction (excerpt)

```
Predict the query response from supplied drug actions, cell observations and optional supports. Write a
short Evidence / Limitation / Conclusion mechanism hypothesis.
[...]
```

[Shared drug]

```
SMILES: {structure} Structure: {description}
Nominal targets: {targets} Nominal pathway: {pathway}
Drug features: {projected features}
```

Support 1

```
Cell: {cell annotations} Cell features: {projected features}
Facts: {action and expression facts}
Observed response: percentile bin <IC50_{bin}>;
representative LN_IC50={ln_ic50}; TRAIN z={z}.
[... Support 2 through Support 5 ...]
```

Query

```
Cell: {query cell annotations} Cell features: {projected features}
Facts: {query action and expression facts}
```

Output:

```
<MECH>
```

Figure 2: **Prompt structure for drug-blind five-shot prediction.** This abridged template places the shared drug before support cells and the query. Braces denote data or projected-feature placeholders; generation begins at <MECH>. At $k = 0$, support blocks are omitted; cell-blind prediction instead uses a shared cell and drug counterparts.

3.2 Connecting biological inputs to the language model

We use Qwen3-14B-Base with trainable LoRA adapters. Molecular and cellular features enter through separate projectors. An offline KPGT drug vector $x_d \in \mathbb{R}^{2304}$ and a normalized expression vector $x_c \in \mathbb{R}^{978}$ are each mapped to four hidden-width tokens:

$$P_m(x_m) = \mathrm{reshape}_{4\times h}\left(W_{m,2}\,\mathrm{GELU}(W_{m,1}\,\mathrm{LN}(x_m) + b_{m,1}) + b_{m,2}\right), \quad m \in \{d, c\}. \tag{1}$$

The intermediate width is 256. Offline representations and native backbone weights remain fixed; the projectors and adapters are learned.

The context serializes a shared-entity header followed by support counterparts and the query counterpart. Textual evidence includes molecular structure, nominal targets and pathways, cell lineage and disease annotations, and bounded action-gene and expression facts. At most two action and two expression facts are exposed by the corresponding fact fields. These facts ground the generated analysis in a compact, inspectable input. They complement continuous features without requiring unrestricted biological descriptions. Support measurements use the numerical representation described below. The query response never appears in the prediction context or in reasoning construction.

Figure 2 shows how these inputs form the generation prefix.

3.3 A probability-weighted numerical answer interface

We extend the model with 100 atomic response tokens, <IC50_00>, ..., <IC50_99>, denoted O100. A fixed transformation is fitted to legal training-query and training-support responses: standardization gives (μ, σ), and quantile boundaries define 100 bins. Let v_j be the mean standardized training response in bin j; empty bins use their boundary midpoint. Each support displays its atomic token, representative standardized value v_j, and representative $\mu + \sigma v_j$ in original units. These are bin representatives, rather than the exact observed measurement.

After the reasoning block closes, the controller supplies <RESP>. Its final hidden state $h_\theta(x, r)$ produces 100 logits through trainable output rows U:

$$p_j(x, r) = \frac{\exp(U_j h_\theta(x, r))}{\sum_{i=0}^{99} \exp(U_i h_\theta(x, r))}, \tag{2}$$

$$\widehat{z}_\theta(x, r) = \sum_{j=0}^{99} p_j(x, r) v_j, \qquad \widehat{y}_\theta(x, r) = \mu + \sigma \widehat{z}_\theta(x, r). \tag{3}$$

The prediction is continuous within the representative range $[\mu + \sigma \min_j v_j,\ \mu + \sigma \max_j v_j]$. The added input and output rows are initialized from the backbone's native numerical spellings and

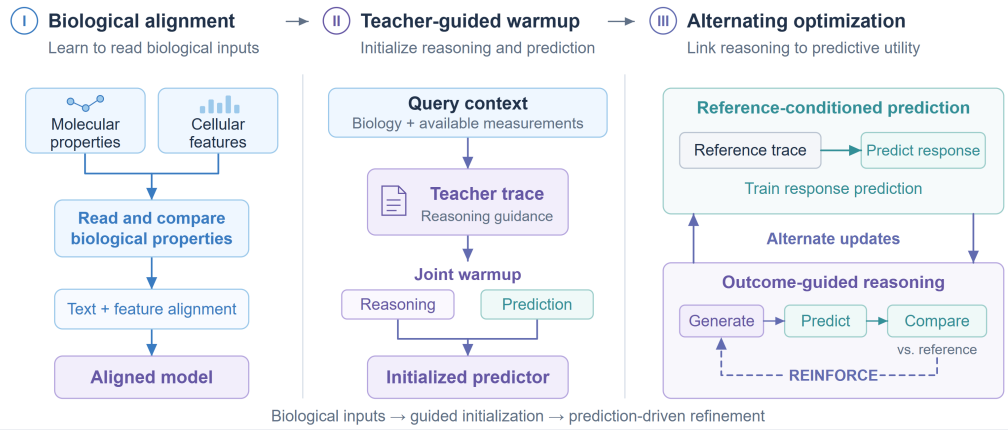

Figure 3: **Three-stage learning architecture.** Biological alignment trains input reading, and teacher-guided warmup initializes reasoning and prediction. The final stage alternates reference-conditioned numerical training and outcome-guided reasoning updates. Prediction error relative to the context-derived reference supplies the REINFORCE signal, while generated-prefix numerical loss also trains prediction. Learned parameters carry forward across stages.

trained during response learning. Reasoning generation uses the native text vocabulary; the answer distribution is computed in one readout, rather than by generating and parsing a number.

For training query bin b_q, the numerical objective is

$$\ell_\theta(x, r) = \left(\widehat{z}_\theta(x, r) - v_{b_q}\right)^2. \tag{4}$$

The regression target is the representative v_{b_q}, while evaluation uses the original continuous response y_q. Thus O100 specifies both the output parameterization and a quantized supervision target. We add categorical bin cross entropy to supervise the distribution as well as its expectation:

$$\mathcal{L}_{\text{resp}}(x, r) = \ell_\theta(x, r) - 0.1 \log p_{b_q}(x, r). \tag{5}$$

The MSE term optimizes the numerical estimate, while the auxiliary term encourages probability mass on the target bin. Exact held-out responses are used only for scoring, never for fitting bins. This interface connects the LLM's contextual state directly to the response scale.

3.4 Stage I: Response-free biological alignment

Before response learning, we align biological inputs with the language model using property-reading tasks with single-digit textual answers. A text-only calibration phase updates LoRA for 256 optimizer steps. A joint phase then updates LoRA and the projectors for 2,048 steps using molecular and expression-reading tasks, masked cell tasks, and drug/cell pair comparisons. Cross entropy supervises the property answer. This stage develops an input-reading interface through response-free supervision. Data exclusions and expression normalization follow the blind protocol. The O100 rows remain frozen during this stage. The resulting weights initialize response learning directly. Figure 3 summarizes how the three training stages connect.

3.5 Stage II: Teacher-guided reasoning warmup

A small teacher corpus establishes how to organize biological evidence and contextual measurements into a concise trace. For each protocol, 128 training queries from distinct target entities are selected independently of their response values, and each receives both a $k = 0$ and a $k = 5$ context. A language-model teacher writes an *Evidence / Limitation / Conclusion* trace from the visible fact packet. The query response and query bin are excluded. Traces express conditional hypotheses and acknowledge limits on transferring support measurements across different compounds or cellular backgrounds.

For teacher trace r^*, warmup jointly trains its token-mean cross entropy, including the closing delimiter, and the response estimate:

$$\mathcal{L}_{\text{warm}} = \mathcal{L}_{\text{resp}}(x, r^*) + 0.25\,\mathcal{L}_{\text{text}}(r^* \mid x). \tag{6}$$

We use 100 updates with effective batch size 64, repeatedly cycling through the shuffled set of 256 teacher contexts. This supplies an initial reasoning behavior and aligns the numerical readout to evidence-conditioned states. Teacher text supervision ends after warmup.

3.6 Stage III: alternating prediction and reasoning optimization

The final stage connects generated reasoning to predictive utility. It alternates two update types over one complete pass through the prepared training pool. Each query uses its stored $k = 0$ or $k = 5$ context. The ratio of numerical to reasoning updates progresses through 8:2, 7:3, 6:4, and 5:5 over exposure quarters, increasing the frequency of sampled-reasoning optimization.

Reference-conditioned prediction. A deterministic reference trace $r_0(x)$ summarizes visible action/expression matches, measurement availability, and, when present, the support-bin median and range. It states transfer limitations and is constructed without the query response. This reference is distinct from the teacher-authored warmup traces. The prediction update minimizes $\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{resp}}(x, r_0)$, allowing gradients through the contextual computation to the adapters, projectors, and O100 rows.

Outcome-guided reasoning. For the next batch, the current model evaluates the reference-conditioned error $e_0 = \ell_\theta(x, r_0)$ without gradients and samples one trace $r \sim \pi_\theta(\cdot \mid x)$ per context. Let $e_r = \ell_\theta(x, r)$. The sampled trace is trained with

$$A(x, r) = \text{sg}[e_0 - e_r + 0.02q(r) - 0.05g(L)], \tag{7}$$

$$\mathcal{L}_{\text{reason}} = \mathcal{L}_{\text{resp}}(x, r) - 0.05\,A(x, r) \sum_{t \in \mathcal{A}(r)} \log \pi_\theta(r_t \mid x, r_{<t}). \tag{8}$$

Here L is the number of sampled tokens, $g(L) = \text{clip}((L - 112)/16, 0, 1)$, and $q(r) = 1$ for a nonempty trace that naturally closes with </MECH> before 128 tokens. The format bonus rewards completed traces, and the length penalty discourages using the entire generation budget. The prediction component of the advantage compares MSEs, excluding bin CE. The detached reference error is independent of the sampled trace and acts as an action-independent baseline for the actor term. It centers the sampled feedback; the reference-conditioned prediction update additionally trains the readout on structured evidence summaries. Appendix D gives the policy-gradient interpretation. The first loss term trains prediction with sampled token identities fixed; the second is a REINFORCE update to generation. Only actually sampled actions receive policy credit. The reference is evaluated by the current model, with no separately learned critic or frozen reference model.

Both alternating update types use FP32 and effective batch 64, retaining the final partial batch; precision and replay checks are detailed in Appendix A. Teacher text supervision ends after warmup. At inference, greedy reasoning generation is followed by the O100 expectation, using either zero or five measurements without parameter updates. Generated traces organize biological facts and support comparisons into analytical leads for inspection and follow-up experimentation.

4 Experimental Protocol

We evaluate entity-held-out generalization within GDSC2, separately for drugs and cells, with zero and five contextual measurements. All P2R-LLM results, including ablations, use Qwen3-14B-Base.

Q-PRED: Generalization to unseen drugs and cells. We use GDSC2 with entity-disjoint source folds: folds 2–9 for training, fold 1 for validation, and fold 0 for testing, with no duplicated drug–cell observations across the three partitions. Entities are held out by ID. Each protocol has a fixed DEV200 screen with ten queries from each of 20 target entities. After support and structure-alias exclusions, drug-blind testing contains 14,323 queries from 24 drugs, and cell-blind testing contains 13,695 queries from 69 cells. Zero-measurement prediction is the primary selection setting.

The primary metric is entity-macro RMSE in original LN_IC50 units:

$$\mathrm{MacroRMSE} = \frac{1}{|E|} \sum_{e \in E} \sqrt{\frac{1}{|Q_e|} \sum_{q \in Q_e} (\widehat{y}_q - y_q)^2}. \tag{9}$$

We also report entity-macro PCC over defined correlations. For P2R-LLM, the checkpoint with lowest generated $k = 0$ DEV macro RMSE under O100 expectation readout is evaluated at $k = 0$ and $k = 5$, with fixed weights and greedy text decoding. Validation frequency, tie handling, and runtime settings appear in Appendix A.

Baselines. Comparisons use the same test queries and support panels. Baselines predict through the O100 expectation readout and use residual kernel ridge regression (KRR) for five-measurement adaptation. Model-specific input qualifications are summarized in Appendix C.

MLP learns nonlinear mappings from drug and cell features through stacked fully connected layers, providing a general-purpose neural prediction baseline. **XGBoost** (2016) is a regularized gradient-boosted decision-tree method that captures nonlinear feature interactions through an additive ensemble (Chen & Guestrin, 2016).

DeepCDR (2020) combines molecular graph representations with cellular multi-omics information in a hybrid graph convolutional framework (Liu et al., 2020). **GraphDRP** (2022) represents compounds as molecular graphs and combines graph-derived drug features with cellular genomic representations for response prediction (Nguyen et al., 2022).

PASO (2025) integrates pathway-based multi-omics differences with drug SMILES representations. It combines transformers, multi-scale convolutions, and attention to model drug–cell relationships (Wu et al., 2025). **DiSPA** (2026) models chemical substructure–pathway interactions through bidirectional conditioning and differential cross-attention, emphasizing context-dependent drug action (Han et al., 2026).

GCNPath (2026) combines pathway-level gene-expression representations with graph convolution to model drug response across cellular contexts (Yoon & Lee, 2026). **TopDr** (2026) represents drug and cell relationships with multiscale simplicial complexes, learning higher-order interactions beyond pairwise graph connections (Shen et al., 2026).

The cell-blind component controls and their results are presented in Section 5.3.

5 Results

5.1 Generalization to unseen cells and drugs

Tables 1 and 2 compare P2R-LLM with eight baselines under cell-blind and drug-blind evaluation. P2R-LLM records the lowest RMSE in all four combinations of blind protocol and measurement availability, and the highest PCC in three of the four settings in the reported comparison.

Cell-blind prediction. At $k = 0$, P2R-LLM achieves an RMSE of 1.2337 and PCC of 0.9011, compared with 1.3261 and 0.8977 for MLP, the strongest baseline on both metrics in this setting. This corresponds to a 6.97% reduction in RMSE. At $k = 5$, P2R-LLM reaches an RMSE of 1.2106 and PCC of 0.9110, improving over the strongest baseline values of 1.2471 and 0.8973, again from MLP. The corresponding RMSE reduction is 2.93%. The model performs well both from biological inputs alone and when a small response panel becomes available.

Drug-blind prediction. At $k = 0$, P2R-LLM's RMSE of 1.6109 improves on PASO's 1.7116 by 5.88%. Its PCC of 0.5776 is close to TopDr's 0.5806 and exceeds the other baselines. With five measurements, P2R-LLM achieves an RMSE of 1.2632, a 0.95% reduction relative to DeepCDR's 1.2753, and the highest PCC of 0.6009, compared with the strongest baseline value of 0.5755. The model thus combines low prediction error with the strongest few-shot correlation among the compared methods.

Table 1: Cell-blind prediction (macro averages). Baselines use residual KRR at $k = 5$. Best values in each column are bold.

Method	$k = 0$		$k = 5$	
	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
MLP	1.3261	0.8977	1.2471	0.8973
XGBoost	1.9240	0.8650	1.7833	0.8593
DeepCDR	1.4646	0.8820	1.3399	0.8820
GraphDRP	1.5706	0.8826	1.3513	0.8808
PASO	1.4445	0.8799	1.3329	0.8794
DiSPA (adapted)	2.0938	0.6751	2.0822	0.6709
GCNPath (adapted)	1.3778	0.8891	1.2819	0.8876
TopDr (transductive)	1.3376	0.8948	1.2599	0.8936
P2R-LLM (ours)	**1.2337**	**0.9011**	**1.2106**	**0.9110**

Table 2: Drug-blind prediction (macro averages). Baselines use residual KRR at $k = 5$. Best values in each column are bold.

Method	$k = 0$		$k = 5$	
	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
MLP	1.7520	0.5675	1.3350	0.5624
XGBoost	1.9298	0.3932	1.4673	0.3540
DeepCDR	1.9159	0.5690	1.2753	0.5698
GraphDRP	2.2485	0.5618	1.3202	0.5617
PASO	1.7116	0.5369	1.2974	0.5390
DiSPA (adapted)	1.9914	0.4338	1.3818	0.4344
GCNPath (adapted)	1.9032	0.5624	1.3725	0.5584
TopDr (transductive)	1.9122	**0.5806**	1.2811	0.5755
P2R-LLM (ours)	**1.6109**	0.5776	**1.2632**	**0.6009**

5.2 Few-shot adaptation through contextual measurements

Providing five measured responses improves both RMSE and PCC in both blind protocols without updating the model's weights. For unseen cells, RMSE decreases from 1.2337 to 1.2106, a 1.87% reduction, while PCC increases from 0.9011 to 0.9110. For unseen drugs, RMSE decreases from 1.6109 to 1.2632, a 21.58% reduction, while PCC rises from 0.5776 to 0.6009. The PCC increases also show that contextual adaptation changes within-entity association: adding an entity-specific constant to every zero-shot prediction would leave that entity's PCC unchanged.

The baselines also benefit from support-based KRR adaptation in RMSE. Most show small cell-blind PCC decreases; DeepCDR is unchanged at the displayed precision. P2R-LLM instead integrates the measurements into its biological context and improves both metrics. This illustrates the practical value of a shared interface that accepts new measurements without parameter updates. The ablations below examine the trained components within that interface.

5.3 Ablation study

We examine biological input alignment (BIO), explicit chain-of-thought reasoning (CoT), and prediction-guided reinforcement learning (RL) under cell-blind prediction. Table 3 reports macro RMSE and PCC at $k = 0$ and $k = 5$. The full model achieves the best value in every column, and removing any of the three components increases prediction error in both measurement settings.

Biological input alignment. The w/o BIO variant omits response-free Stage I and starts directly with response learning. Its RMSE increases from 1.2337 to 1.2377 at $k = 0$ and from 1.2106 to 1.2169 at $k = 5$, with corresponding PCC decreases of 0.0001 and 0.0021. The observed changes are small and directionally consistent, with a larger RMSE difference when measurements are available.

Table 3: Cell-blind component ablations. Best values in each column are bold.

Variant	$k = 0$		$k = 5$	
	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
Full model	**1.2337**	**0.9011**	**1.2106**	**0.9110**
w/o BIO	1.2377	0.9010	1.2169	0.9089
w/o CoT	1.2461	0.8984	1.2310	0.9000
w/o RL	1.2417	0.8977	1.2352	0.8988

Explicit reasoning. The w/o CoT variant predicts directly from the same biological and support inputs, replacing reasoning-dependent training with numerical prediction updates. At $k = 0$, this direct predictor already improves RMSE from MLP's 1.3261 to 1.2461; the full model makes a further 0.0124 reduction to 1.2337. Thus, most of the observed gap to MLP is present without generated reasoning. At $k = 5$, the full model reduces RMSE from 1.2310 to 1.2106 (1.66%) and increases PCC from 0.9000 to 0.9110. This tests the complete trained pathway, including intermediate autoregressive computation, rather than isolating the semantic content of the text.

Prediction-guided reinforcement learning. The w/o RL variant removes the complete shaped actor update, including prediction-error, format, and length feedback, while retaining teacher warmup, sampled traces, and the numerical prediction objective. Its RMSE increases to 1.2417 at $k = 0$ and 1.2352 at $k = 5$, while PCC decreases to 0.8977 and 0.8988. The five-shot degradation is the largest among the three ablations: the full model improves RMSE by 1.99% and PCC by 0.0122 relative to w/o RL. This comparison measures the combined contribution of the shaped actor objective alongside teacher initialization and numerical training.

Interaction with contextual measurements. The full model's RMSE decreases by 0.0231 when moving from $k = 0$ to $k = 5$. The corresponding decreases are 0.0208 without BIO, 0.0151 without CoT, and 0.0065 without RL. The full model also exhibits the largest PCC increase, 0.0099, compared with 0.0079, 0.0016, and 0.0011, respectively.

6 Conclusion

We introduce P2R-LLM, an LLM framework that combines biological evidence, in-context few-shot adaptation, and prediction-guided reasoning for drug response prediction. Biological alignment, teacher warmup, and reference-relative reinforcement learning connect heterogeneous inputs and generated comparisons to a continuous response estimate. Contextual observations improve RMSE and PCC in both entity-held-out GDSC2 protocols, and the complete recipe has the lowest cell-blind ablation errors. The framework also records the lowest macro RMSE in all four main comparisons. These findings support contextual measurement integration for unseen-entity prediction and motivate further study of generated analytical traces as a bridge between biological inputs and quantitative DRP.

AI USE STATEMENT

Generative AI assisted manuscript organization, and prose drafting. A language-model teacher generated warmup reasoning traces from visible biological facts and support summaries, as described in Section 3.

ETHICS STATEMENT

This work concerns retrospective prediction in cancer cell lines. Predictions and generated comparisons are research outputs and are not validated treatment recommendations. Biological annotations may be incomplete and unevenly distributed across cell lineages and compounds. Any release must respect the licenses of the underlying datasets, model weights, and annotation sources.

REPRODUCIBILITY STATEMENT

Appendix A specifies the numerical interface, supervision, training schedule, and implementation settings. The source supplement documents required data assets, training and evaluation entry points, and the software and hardware environment. We report entity-macro RMSE and PCC for both blind protocols using the evaluation procedure in Section 4.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Shaika Chowdhury, Sivaraman Rajaganapathy, Lichao Sun, Liewei Wang, Ping Yang, James R Cerhan, and Nansu Zong. Sensitivecancergpt: Leveraging generative large language model on structured omics data to optimize drug sensitivity prediction. *bioRxiv*, 2025.

Carl Edwards, Aakanksha Naik, Tushar Khot, Martin D. Burke, Heng Ji, and Tom Hope. Synergpt: In-context learning for personalized drug synergy prediction and drug design. *CoRR*, abs/2307.11694, 2023. doi: 10.48550/ARXIV.2307.11694. URL https://doi.org/10.48550/arXiv.2307.11694.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Rohan Gnanaolivu, Priyanka Vasanthakumari, Brenna C Novotny, Jamie C Overbeek, Yuanhang Liu, Alexander Partin, Nicholas Chia, Yitan Zhu, and Chen Wang. Drpllm: A large language model-based framework for predicting drug response in cancer using multi-omics data. In *International Conference on Bioinformatics and Computational Biology*, pp. 189–203. Springer, 2025.

Hui Guo, Xiang Lv, Shenghao Li, Daichuan Ma, Yizhou Li, and Menglong Li. Fast-adapting graph neural network with prior knowledge for drug response prediction across preclinical and clinical data. *Journal of Pharmaceutical Analysis*, 15(10):101386, October 2025. ISSN 2095-1779. doi: 10.1016/j.jpha.2025.101386. URL http://dx.doi.org/10.1016/j.jpha.2025.101386.

Yewon Han, Sunghyun Kim, Eunyi Jeong, Sungkyung Lee, Seokwoo Yun, and Sangsoo Lim. Dispa: Differential substructure-pathway attention for drug response prediction. *CoRR*, abs/2601.14346, 2026. doi: 10.48550/ARXIV.2601.14346. URL https://doi.org/10.48550/arXiv.2601.14346.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Likun Jiang, Changzhi Jiang, Xinyu Yu, Rao Fu, Shuting Jin, and Xiangrong Liu. DeepTTA: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics*, 23(3), March 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac100. URL http://dx.doi.org/10.1093/bib/bbac100.

Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1), November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43214-1. URL http://dx.doi.org/10.1038/s41467-023-43214-1.

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinform.*, 36(Supplement-2):i911–i918, 2020. doi: 10.1093/BIOINFORMATICS/BTAA822. URL https://doi.org/10.1093/bioinformatics/btaa822.

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.

Tuan Nguyen, Giang T. T. Nguyen, Thin Nguyen, and Duc-Hau Le. Graph convolutional networks for drug response prediction. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 19(1):146–154, 2022. doi: 10.1109/TCBB.2021.3060430. URL https://doi.org/10.1109/TCBB.2021.3060430.

Cong Shen, Guancen Lin, Chuan-Shen Hu, and Yong Wang. Topology-Aware Deep Learning on Higher-Order Structures for Drug Response Prediction. *Advanced Science*, 13(46), May 2026. ISSN 2198-3844. doi: 10.1002/advs.75816. URL http://dx.doi.org/10.1002/advs.75816.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Yang Wu, Ming Chen, and Yufang Qin. Anticancer drug response prediction integrating multi-omics pathway-based difference features and multiple deep learning techniques. *PLoS Comput. Biol.*, 21(3):1012905, 2025. doi: 10.1371/JOURNAL.PCBI.1012905. URL https://doi.org/10.1371/journal.pcbi.1012905.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon A. Forbes, Nidhi Bindal, David Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and

Mathew Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(Database-Issue):955–961, 2013. doi: 10.1093/NAR/GKS1111. URL https://doi.org/10.1093/nar/gks1111.

Hyeon Jun Yoon and Minho Lee. GCNPath: introspecting drug response prediction with pathway-guided graph convolution networks. *Communications Biology*, 9(1), April 2026. ISSN 2399-3642. doi: 10.1038/s42003-026-09957-5. URL http://dx.doi.org/10.1038/s42003-026-09957-5.

A Implementation and Information Boundaries

Stages and model variants. The supplement labels teacher warmup A, reference-conditioned prediction B, and sampled-reasoning reinforcement C. These correspond to manuscript Stage II and the two update types within Stage III; Stage I is BIO. All main and ablation experiments use Qwen3-14B-Base (base revision `0b0bd373`). Native base weights remain frozen, with rank-16, alpha-32 LoRA, separate four-token entity projectors, and 100 added response input/output rows. Maximum sequence length is 4,096; reasoning generation is capped at 128 tokens. Checkpoints store compact trainable parameters, not the frozen backbone.

Teacher and factual inputs. Teacher traces are authored by `gpt-6-astra` with medium reasoning effort. For each protocol, 128 TRAIN queries receive separate zero- and five-support contexts, producing 256 teacher texts. Each text is authored from its individual factual packet without access to the query response label; the frozen texts are reused during warmup. The current prepared data package fixes these teacher assignments and the factual inputs. Drug packets contain canonical SMILES, structure descriptions, original target/pathway annotations, and selected visible action records. Cell packets contain lineage, disease, subtype and origin metadata together with selected expression/context records. Frozen KPGT and expression banks supply the numerical features. These stored inputs are shared across training and inference; inference requires no live retrieval. The package documents the consumed fields and query-label exclusion, but does not provide a complete upstream source/version and filtering audit for every textual record. Freezing these inputs establishes repeatability, not their biological correctness or response independence throughout upstream collection.

Biological alignment. BIO performs 256 text-only calibration updates and 2,048 joint updates, using effective batch 32 and physical microbatch four. The added response rows remain frozen. Calibration updates LoRA at 10^{-4}; joint alignment additionally updates projectors at 5×10^{-5}. Each phase uses a fresh AdamW optimizer, weight decay 0.01, clipping at one, ten warmup updates and linear decay to 10% of the peak rate. BIO uses seed 42 and BF16 AMP. Tasks include cell reading, masked expression, cell pairs, drug reading, and drug pairs, with weights $(0.25, 0, 0.25, 0.25, 0.25)$ in calibration and $(0.125, 0.5, 0.125, 0.125, 0.125)$ in joint training.

Response optimization and exposure. Stage II repeats the frozen 256 teacher contexts for 100 updates of 64 contexts, giving 6,400 exposures rather than 6,400 distinct teacher generations. Stage III traverses 111,491 drug-blind queries in 1,743 updates or 110,576 cell-blind queries in 1,728 updates, retaining the final partial batch. The B:C ratios over exposure quarters are 8:2, 7:3, 6:4, and 5:5, giving 1,133/610 B/C updates for drug-blind and 1,123/605 for cell-blind. Each query uses its stored support setting; batches are not required to have an exact half-zero/half-five composition.

Stages II and III share AdamW with weight decay 0.01, clipping at one, 20 warmup updates, peak learning rate 10^{-4}, and final rate 10^{-5}. The schedule spans warmup plus the full alternating pass. Response seeds are 85 for drug-blind and 86 for cell-blind. Stage II uses BF16 AMP; Stage III uses strict FP32 with TF32 disabled. Physical response-training microbatches contain one example, accumulated to effective batch 64. Numeric projection remains FP32. Replay tolerances are 0.05 per sampled token and 0.5 for summed log probabilities.

Format and length shaping. The qualification bonus requires a nonempty body, a naturally generated closing mechanism delimiter, and fewer than 128 sampled tokens. Empty text, EOS without closure, and controller-forced completion receive no bonus. The length penalty starts beyond 112 sampled tokens and saturates at 128. These additions shape completion behavior; they do not label a trace as biologically valid. Teacher CE appears only during warmup, and reference MSE is detached in C.

Data and inference. Frozen prepared assets contain query/support memberships, drug and cell features, teacher texts, and the TRAIN-derived response transform. Query labels enter only losses or metrics; the query is excluded from its support panel. Neither response normalization nor O100 representatives are refitted on held-out data. Evaluation uses greedy native-text generation, BF16 AMP over FP32-resident weights, FP32 numerical projection, and SDPA. The fixed test batch size

is eight. Both support settings cover the complete test populations. Macro PCC averages defined entity correlations with at least 20 observations.

Hardware and runtime accounting. The hardware configuration uses a single NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB memory. The software environment uses Linux, Python 3.12, PyTorch 2.9.1 with CUDA 12.8, Transformers 4.57.3, and PEFT 0.18.0. The clean entry has no deployment time cutoff or batch-size benchmark. Training time starts after model/data loading and includes BIO, response training, validation and checkpoint writes. Test timing includes model/data loading and both support settings. Offline features, data preparation, and teacher generation are outside this timing scope. Table 4 summarizes the estimated full-schedule compute budget supplied with the code package. These estimates do not establish measured throughput, mean generated length, or peak memory for the reported model, nor its cost relative to direct prediction.

Table 4: Estimated wall-clock budget in hours on one RTX PRO 6000 96 GB.

Stage	Drug-blind	Cell-blind
BIO alignment	≈ 3	≈ 3
Response training	98.7	118.8
Validation	0.54	0.57
Full test ($k = 0, 5$)	15.84	15.64
Other overhead	1–2	1–2
Total	119–120	139–140

B Evaluation scope and interpretation

The experiments evaluate retrospective single-agent GDSC2 response with one training seed per protocol, fixed entity-disjoint response folds, and fixed support panels. Given the compute cost of 14B training, available computational resources limit the experiments to one training seed per protocol; multi-seed replication is left for future evaluation. These partitions were also used during prior project development. The 8:1:1 partition was used without special test-based selection: fold 0 was not a selection set, and checkpoint selection uses DEV as described below. An unseen entity is defined by its ID's exclusion from response training. This does not certify that identical canonical structures under different drug IDs are absent across all folds. Structure-alias filtering between supports and scored queries is a separate exclusion, not a global structure or scaffold split. The cell-blind ablations evaluate the predictive contributions of BIO, the complete CoT pathway, and the actor update. Generated traces serve as conditional analytical hypotheses for follow-up investigation. The O100 prediction is an expectation over training-derived bin representatives, and its numerical range follows from those representatives.

Evaluation and reporting details. Validation follows teacher warmup, alternating-stage update 100, every 500 subsequent updates, and the endpoint. Strictly lower DEV macro RMSE replaces the best checkpoint; ties retain the earlier checkpoint. Both test settings use that checkpoint with greedy generation capped at 128 tokens. The compute budget includes both support settings in the full test population. The reported scores are single-run results, without averaging across training seeds or support-panel draws. The reported point differences do not establish stability across training seeds, support panels, or test-entity resampling. Comparisons retain model-specific training and adaptation procedures. Generated-text benefits over a deterministic evidence summary and the share of RMSE improvement explained by simple support calibration remain open questions.

C Baseline comparison inputs

Baselines retain their model-specific representations and use O100 expectation predictions, with residual KRR adaptation from the same five support measurements. DiSPA and GCNPath are adapted to the available gene features; GCNPath's expression references use training cells. TopDr uses all entities' unlabeled features to construct its graph, with response optimization restricted to

training labels, and is therefore identified as transductive. P2R-LLM additionally integrates textual facts and a pretrained language backbone. The w/o CoT control retains these biological and support inputs for direct prediction.

D NUMERICAL SUPERVISION AND POLICY-GRADIENT INTERPRETATION

The bin loss $-\log p_{b_q}$ is categorical cross entropy. Numerical ordering enters through the representative values and the squared-error term, rather than through the categorical loss itself. Training targets v_{b_q} and exact standardized observations $z_q = (y_q - \mu)/\sigma$ are distinct; test RMSE is computed against the original observed responses.

For fixed x, the reference error $e_0(x)$ does not depend on sampled r. Under on-policy sampling,

$$\mathbb{E}_{r\sim\pi_\theta}\left[\mathrm{sg}(e_0(x))\nabla_\theta \log \pi_\theta(r \mid x)\right] = 0. \tag{10}$$

It therefore centers the actor estimator without changing its expected direction. The baseline choice affects estimator variance and finite-sample updates; the separate reference-conditioned numerical update also determines which hidden states train the prediction interface. The framework combines these training roles with a shared predictor and the O100 readout.

Reasoning updates sample from the full native-text softmax at temperature one, without top-k or top-p truncation. Test generation uses greedy decoding. The w/o RL control removes the entire actor term, including its prediction-error, qualification, and length components. The w/o CoT control changes both training and intermediate computation. These controls address the trained pathways; semantic-content interventions are a distinct comparison.