Identification of Heterogeneous Erlotinib Response Gene Sets Using Sample-Specific Counterfactual Causal Attribution

Hyekyoung Lee [1] **Seungjin Choi** [2] **Hyunjin Shin** [1]

Abstract

Identifying cell-line-specific gene sets associated with drug response is difficult because pharmacogenomic data are high-dimensional, continuous-valued, and rarely paired with validated sample-level mechanism labels. We present a mechanism-guided framework for model-based counterfactual attribution of erlotinib response from transcriptomic data. The framework first uses a BIRD-inspired evidence-integration step, motivated by Bayesian inference from abduction and deduction (BIRD), to combine expression-response association, mutation support, KEGG topology, and curated EGFR tyrosine kinase inhibitor biology in an unlabeled omics setting. This produces a BIRD-inspired panel, referred to as BIRD-20, designed to preserve erlotinib-relevant mechanism diversity. Using foldwise reconstructed BIRD-20 panels, predefined comparator panels, and mechanism-informed orderings, we approximate conditional counterfactual causal effect (CCCE) scores with causal normalizing flows (CNFs), enabling one-, two-, and three-gene counterfactual interventions in continuous expression and erlotinib logIC50 space. In five-fold held-out evaluation, BIRD-inspired panels, including foldwise train-selected panels, showed consistently higher hit rates than Random-KEGG-20 and KEGG-26 across one-, two-, and three-gene interventions. The resulting profiles provide interpretable, sample-level counterfactual hypotheses that prioritize experimentally testable gene combinations for downstream biological validation.

[1]MOGAM Institute for Biomedical Research, 15-19 Gangnam-daero 41-gil, Seocho-gu, Seoul, Republic of Korea [2]CROID Research and aSSIST University, Seoul, Korea. Correspondence to: Hyunjin Shin <hyunjin.shin@mogam.re.kr>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Large-scale pharmacogenomic resources such as the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC) have enabled systematic modeling of anticancer drug response from molecular features of cancer cell lines (Barretina et al., 2012; Yang et al., 2013; Iorio et al., 2016). A common task is to predict a continuous drug-response phenotype, such as the half-maximal inhibitory concentration (IC50), from genomic or transcriptomic measurements. Deep learning and multi-omics models have improved predictive performance by integrating gene expression, mutation, and copy number information (Sharifi-Noghabi et al., 2019). However, accurate prediction alone does not answer a key biological question: for a particular cell line, which genes or gene sets should be prioritized as plausible contributors to the observed drug response?

Prior work has explored more interpretable drug-response models. Pathway-guided neural networks constrain model structure using biological pathways (Deng et al., 2020), DrugCell uses a visible neural network organized by the hierarchy of human cell biology (Kuenzi et al., 2020), and DRPreter combines knowledge-guided graph neural networks with a transformer module to highlight drug-associated pathways (Shin et al., 2022). These approaches show that biological prior knowledge can improve interpretability, but they remain primarily predictive. Pathway saliency or feature importance does not necessarily correspond to counterfactual causal contribution, and most explanations are not defined at the individual-sample level.

Counterfactual modeling provides a structured route to this question, although it remains dependent on modeling assumptions. For biological perturbation data, CODEX uses counterfactual deep learning to predict unobserved perturbation outcomes and prioritize in silico experiments (Schrod et al., 2024). Related perturbation-response models such as GEARS and scGen also focus on predicting transcriptional responses under unseen genetic or cellular perturbations (Roohani et al., 2024; Lotfollahi et al., 2019). In causal inference, conditional counterfactual causal effect (CCCE) quantifies how much a cause or cause set contributes to an observed individual outcome (Zhao et al., 2023). However,

CCCE is most natural for discrete variables and exact enumeration, whereas transcriptomic drug-response analysis involves continuous gene expression and continuous IC50. Causal normalizing flows (CNFs) offer a useful alternative because they model continuous joint distributions and support interventional and counterfactual queries under causal ordering or graph assumptions (Javaloy et al., 2023). Our question is complementary to these perturbation-prediction studies: rather than predicting a new perturbation outcome alone, we rank gene sets that most reduce the observed high-IC50 state of each held-out cell line under a fixed mechanism-guided model.

A second challenge is the dimensionality of the intervention space. Transcriptomic datasets contain tens of thousands of genes, making direct counterfactual modeling unstable, computationally expensive, and difficult to interpret. Purely pathway-based reduction can preserve canonical signaling structure but may overrepresent dense pathways, while marginal association or feature-importance ranking can select redundant genes and miss distinct resistance mechanisms. This issue is especially relevant for erlotinib, whose response and resistance involve multiple axes, including direct epidermal growth factor receptor (EGFR) signaling, bypass receptor tyrosine kinase (RTK) activation, mitogen-activated protein kinase (MAPK) escape, phosphoinositide 3-kinase/protein kinase B/phosphatase and tensin homolog (PI3K/AKT/PTEN) signaling, epithelial-mesenchymal transition (EMT)/AXL-associated state, and apoptotic priming. Thus, a useful counterfactual attribution pipeline must restrict the intervention space while preserving coverage of diverse response and resistance mechanisms.

We address these challenges with a mechanism-guided framework inspired by Bayesian inference from abduction and deduction (BIRD) (Feng et al., 2025). We adapt BIRD's factor-based evidence-integration principle to pharmacogenomic data and use conditional probability table (CPT)-derived relevance scores to construct compact gene panels. These panels define a tractable hypothesis space for a CNF-based model, which ranks sample-specific one-, two-, and three-gene counterfactual interventions under a mechanism-informed ordering. We assess the robustness of the panel-selection framework using foldwise reconstruction, an omics-independent variant, and text-, pathway-, and a size-matched random KEGG comparator.

Leakage-aware five-fold evaluation showed that BIRD-inspired panels had consistently higher hit rates than pathway-derived controls across one-, two-, and three-gene interventions. We further examined deterministic threshold and anchor sensitivity as robustness diagnostics and tested whether selected multi-gene profiles showed model-based non-additivity relative to additive single-gene benchmarks. The framework is designed to prioritize sample-specific, model-based counterfactual gene-set hypotheses for downstream biological validation.

Our contributions are threefold. First, we introduce a BIRD-inspired, mechanism-guided evidence-integration framework for constructing compact and mechanistically diverse gene panels without requiring causal-driver labels. Second, we combine these panels with CNFs to approximate CCCE-style sample-specific attribution over continuous transcriptomic and drug-response variables, including higher-order gene-set interventions. Third, we provide a leakage-aware and robustness-oriented evaluation using foldwise panel reconstruction, omics-independent selection, random and pathway-derived controls, and multiple counterfactual hit thresholds. Together, the proposed framework integrates prior biological knowledge, transparent evidence aggregation, and generative counterfactual modeling to prioritize cell-line-specific hypotheses for downstream experimental investigation.

2. Method

2.1. Problem setup and data

Our goal is to prioritize genes and gene sets that may contribute to cell-line-specific erlotinib response under the assumed model and to report them as counterfactual hypotheses. Let $Y \in \mathbb{R}^n$ denote erlotinib logIC50 and $X \in \mathbb{R}^{n \times p}$ the corresponding transcriptomic features. The dataset comprises transcripts per million (TPM) expression values for $p = 19{,}193$ genes across $n = 571$ cancer cell lines matched to GDSC erlotinib response measurements (Yang et al., 2013). Cell-line identifiers and gene annotations were standardized, and hotspot mutations, Kyoto Encyclopedia of Genes and Genomes (KEGG) topology, and curated EGFR-tyrosine kinase inhibitor (TKI) response and resistance mechanisms were used as auxiliary evidence for panel construction. We denote one-, two-, and three-gene interventions as CR1, CR2, and CR3, respectively. Because sample-specific causal-driver labels are unavailable, these profiles are interpreted as model-based hypotheses conditional on the selected panel, graph ordering, intervention anchors, and CNF model.

We evaluated five panels: BIRD-20, NoOmics-BIRD-20, GPT-20, Random-KEGG-20, and KEGG-26. BIRD-20 denotes the leakage-aware panel reconstructed using training data within each cross-validation fold. NoOmics-BIRD-20 excludes cohort-derived expression, mutation, family-activity, and multi-source consistency factors to assess the utility of pathway/topology and curated mechanism structure alone. GPT-20 is a text-curated EGFR-TKI panel, KEGG-26 is a canonical pathway-derived panel, and Random-KEGG-20 is a size-matched KEGG-derived control. The five-fold evaluation therefore assesses the complete

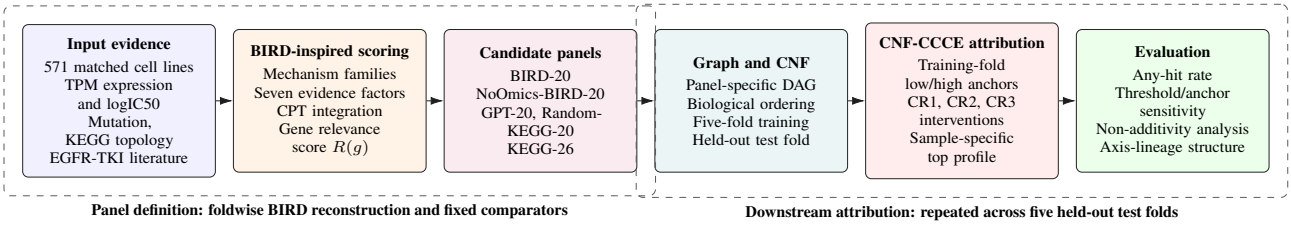

Figure 1. Overall analysis workflow. BIRD-20 is reconstructed within each training fold, while NoOmics-BIRD-20, GPT-20, Random-KEGG-20, and KEGG-26 are evaluated as predefined comparators under the same five-fold CNF-based counterfactual attribution protocol.

foldwise BIRD pipeline and the downstream CNF-CCCE procedure for the comparator panels. Additional implementation details for panel artifacts, foldwise BIRD-20 variability, comparator gene composition, and mechanism-axis coverage are provided in Appendices A.1–A.5.

2.2. Overview of the framework

The framework comprises two stages. First, BIRD-inspired evidence integration constructs a compact, mechanism-guided intervention space (Feng et al., 2025). A biologically informed variable ordering is then specified for CNF training. In this unlabeled omics setting, mechanism families are defined from prior knowledge, while gene-level evidence combines expression–response associations, mutation support, KEGG topology, and curated erlotinib/EGFR-TKI biology. Second, a CNF is trained on the selected variables to generate sample-specific counterfactuals and approximate CCCE-style scores for one-, two-, and three-gene interventions (Javaloy et al., 2023; Zhao et al., 2023).

Figure 1 summarizes the pipeline. Family-aware compression selects a 20-gene panel while preserving mechanistic diversity and reducing within-family redundancy. During five-fold evaluation, the BIRD-inspired panel is reconstructed using only the corresponding training fold and is referred to as BIRD-20. The resulting panel and ordering are then used to train the CNF and evaluate CR1–CR3 intervention sets. The same downstream counterfactual evaluation is then applied to the predefined comparator panels described above.

2.3. Gene selection by BIRD-inspired CPT evidence integration

2.3.1. Mechanism families and input evidence

Prior knowledge was organized into six erlotinib-relevant mechanism families: direct EGFR signaling, bypass RTK activation, MAPK-mediated escape, PI3K/AKT/PTEN survival signaling, EMT/AXL-associated resistance, and apoptotic priming, including the BIM/*BCL2L11* axis. These families served as coverage constraints to preserve mechanistic diversity in the final panel. An EGFR-centered expression-program list was used only as auxiliary evidence because it substantially overlapped with the direct EGFR family.

For each gene, seven evidence factors were calculated:

- **Expression association (F_1):** absolute Spearman correlation between lineage-residualized gene expression and erlotinib logIC50.
- **Mutation association (F_2):** absolute difference in mean logIC50 between mutated and wild-type cell lines.
- **KEGG centrality (F_3):** average of normalized PageRank, in-degree, and out-degree in the erlotinib-relevant KEGG union graph.
- **EGFR proximity (F_4):** inverse shortest-path distance to the nearest *EGFR*, *ERBB2*, *ERBB3*, or *ERBB4* seed, with unreachable genes assigned zero.
- **Literature membership (F_5):** normalized sum of the weights of curated mechanism families containing the gene.
- **Family activity support (F_6):** normalized support from response-associated expression activity of the gene's mechanism families, accounting for the expected sensitivity or resistance direction.
- **Multi-source consistency (F_7):** proportion of F_1–F_6 with scores of at least 0.67.

Multi-source consistency was included as a corroboration factor rather than as an independent evidence source. It upweighted genes supported by multiple evidence channels. Continuous factors were min–max normalized to $[0, 1]$, and cohort-derived quantities were estimated using only the corresponding training fold. The evidence factors quantified gene-level support, whereas the mechanism-family constraints promoted coverage and reduced within-family redundancy.

2.3.2. Genome-wide BIRD-inspired CPT integration

Table 1 illustrates the normalized evidence profiles of representative genes. Within each training fold, continuous

Table 1. Example evidence-factor matrix for BIRD-inspired gene scoring.

Evidence factor	*EGFR*	*GRB2*	*AXL*	*GAS6*	*ERBB4*	*AKT1*
Expression association	0.28	0.20	0.34	0.32	0.10	0.17
Mutation association	0.72	–	–	–	–	–
KEGG centrality	0.65	0.82	0.13	0.01	0.55	0.78
EGFR proximity	1.00	0.50	0.33	0.25	1.00	0.33
Literature membership	0.67	0.67	0.63	0.63	0.40	0.33
Family activity support	0.09	0.09	1.00	1.00	0.06	0.38
Multi-source consistency	0.33	0.17	0.17	0.17	0.17	0.17

factors were min–max normalized across eligible genes and discretized into low, medium, and high states using tertile cutoffs. Missing association evidence was assigned to the neutral medium state, whereas absence of graph-based or curated support was assigned to the low state. The relationship between this adaptation and the original BIRD framework is clarified in Appendix A.3 (Feng et al., 2025).

For each factor F_j, a prespecified CPT related its three evidence states to three relevance classes, $O \in \{h, m, w\}$, denoting high, moderate, and weak relevance. CPT entries were converted to factor-specific relevance distributions using fixed class priors. For a gene g with factor state $s_j(g)$, these distributions were aggregated as

$$P_{\mathrm{CPT}}(O = o \mid g) = \sum_j \alpha_j P(O = o \mid F_j = s_j(g)),$$

where $\sum_j \alpha_j = 1$. The integrated relevance score was

$$R(g) = P_{\mathrm{CPT}}(O = h \mid g) + 0.5 P_{\mathrm{CPT}}(O = m \mid g).$$

This procedure assigned relevance scores to all 19,193 genes. Panel selection was subsequently restricted to a predefined erlotinib/EGFR-TKI mechanism shortlist. We therefore use $R(g)$ as a CPT-derived evidence score rather than as a calibrated probability of causal relevance. The panel-construction summary and shortlist are reported in Appendix A.2.

2.3.3. Family-aware panel compression

The six mechanism families contained 69 unique curated genes. From this set, a 25-gene shortlist was defined to represent the major erlotinib response and resistance mechanisms. Within each training fold, the shortlisted genes were reranked using their CPT-derived relevance scores and compressed into a 20-gene panel.

For a candidate panel subset P, the selection objective was

$$J(P) = \sum_{g \in P} R(g) + \lambda_q \sum_{g \in P} Q(g) - \lambda_{nf} \sum_f \big(n_f(P) - 1\big)^2,$$
$$\text{subject to } |P| = 20 \quad \text{and} \quad n_f(P) \geq 1 \text{ for all } f,$$

where $Q(g)$ denotes family representativeness and $n_f(P)$ is the number of selected genes assigned to family f. The objective balances gene-level evidence, family representation, and within-family redundancy.

All cohort-derived evidence scores and panel-selection steps were recomputed using only the training samples in each fold. Consequently, the five-fold analysis used a fold-specific BIRD-20 panel rather than a single full-cohort panel. The resulting fold-specific BIRD-20 panels and their variability are reported in Appendix A.4. The 25-to-20 compression weights, family-representativeness score, and exhaustive enumeration algorithm are summarized in Appendix A.2.

2.4. DAG construction and comparator panels

For each panel, a mechanism-informed acyclic graph was constructed to derive the variable ordering used for CNF training. The graphs were not learned from the data. Direct KEGG- or literature-supported edges were retained, and pathway shortcut edges were added when intermediate signaling components were absent from a compact panel. Feedback edges were excluded to preserve acyclicity.

A separate graph and topological ordering were constructed for each fold-specific BIRD-20 panel using the same rules. The graph was used to determine the CNF ordering rather than to impose an additional edge-wise regularization constraint. Panel and graph artifacts are summarized in Appendix A.1.

All five panels described in Appendix A.1 were evaluated using the same downstream CNF-CCCE procedure.

2.5. CNF-based continuous counterfactual modeling

Let $X = (X_1, \ldots, X_m)$ denote the selected gene-expression variables and Y denote erlotinib logIC50. For each panel and fold, a causal normalizing flow (CNF) was trained on (X, Y) to model their joint observational distribution and generate continuous counterfactual outcomes (Javaloy et al., 2023). A mechanism-informed variable ordering was supplied to the model, with upstream signaling components preceding downstream components and Y placed last. No edge-wise graph regularization was applied.

All variables were standardized using training-fold statistics. BIRD-20 was reconstructed within each training fold, whereas comparator panels followed their predefined construction protocols. CNF architecture, optimizer settings, checkpoint selection, random seed, and inverse-transformation details are summarized in Appendix A.6.

For an observed sample vector x, we first obtained its factual latent representation,

$$z^{\mathrm{fact}} = f_\theta(x).$$

For an intervention on gene set $S \subseteq P$ with anchor as-

signment a, we constructed $x^{\text{int}}(S, a)$ by replacing the corresponding gene-expression values and encoded it as $z^{\text{int}} = f_\theta(x^{\text{int}})$. A hybrid latent vector was then formed by taking the coordinates of the intervened genes from z^{int} and retaining all remaining coordinates from z^{fact}. Applying the inverse CNF transform produced the counterfactual vector, whose response coordinate was used as the counterfactual erlotinib logIC50. This latent-coordinate replacement is our implementation of the intervention operator under the fitted triangular CNF ordering. No new latent noise was sampled. Each anchor assignment produced one deterministic counterfactual outcome per anchor-perturbation round.

2.6. Sample-specific counterfactual scoring

We used the CNF to rank gene sets predicted to reduce erlotinib logIC50 for individual cell lines, following the motivation of CCCE while accommodating continuous expression and response variables (Zhao et al., 2023). The analysis was restricted to held-out cell lines whose observed logIC50 exceeded the corresponding training-fold mean.

For each gene, low- and high-expression anchors were defined by its 15th and 85th percentiles in the training fold. For a gene set S, all mixed low/high assignments were evaluated:

$$\mathcal{A}_S^{(r)} = \prod_{g \in S} \{a_{g,r}^{\text{low}}, a_{g,r}^{\text{high}}\}.$$

Accordingly, CR1, CR2, and CR3 evaluated 2, 4, and 8 intervention directions, respectively.

We used 10 anchor-perturbation rounds. In each round, the low and high anchors were perturbed by Gaussian noise with standard deviation equal to 5% of the gene-specific 15th–85th percentile span and clipped to the training-fold expression range. For each candidate set and round, the direction yielding the largest normalized reduction was retained:

$$\Delta_r(S; x) = \max_{a \in \mathcal{A}_S^{(r)}} \frac{Y_{\text{obs}}(x) - \widehat{Y}_{\text{cf}}(S, x, a)}{\max\{|Y_{\text{obs}}(x)|, \epsilon\}}.$$

Candidate sets were ranked by their mean reduction across the 10 rounds,

$$\overline{\Delta}(S; x) = \frac{1}{10} \sum_{r=1}^{10} \Delta_r(S; x).$$

The primary threshold was 0.30, with sensitivity analyses at 0.10, 0.20, 0.30, 0.40, and 0.50. These thresholds refer to normalized reductions on the logIC50 scale, not percentage reductions on the original IC50 scale. We additionally recorded the hit fraction, defined as the proportion of rounds in which the selected gene set exceeded the corresponding threshold.

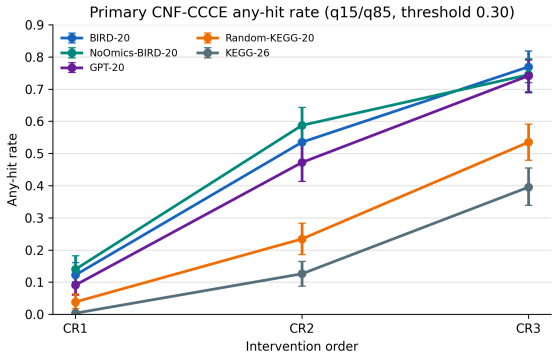

Figure 2. CNF-CCCE any-hit rates for CR1, CR2, and CR3 using q15/q85 base anchors and a normalized logIC50 reduction threshold of 0.30. Error bars denote 95% bootstrap confidence intervals over held-out high-logIC50 cell lines.

3. Results

3.1. Five-fold held-out evaluation of intervention panels

We evaluated BIRD-20, NoOmics-BIRD-20, GPT-20, Random-KEGG-20, and KEGG-26 using the same five-fold CNF-CCCE protocol. BIRD-20 was reconstructed using only the training data within each fold, whereas the comparator panels followed their predefined construction protocols. Evaluation was restricted to held-out cell lines whose observed erlotinib logIC50 exceeded the corresponding training-fold mean, yielding 286 high-logIC50 cell lines.

The primary endpoint was the any-hit rate using q15/q85 base anchors and a normalized logIC50 reduction threshold of 0.30. A sample was counted as a hit when its top-ranked gene set exceeded this threshold in at least one of 10 anchor-perturbation rounds. The threshold represents a normalized reduction on the logIC50 scale.

Figure 2 summarizes the CR1–CR3 any-hit rates, and Table 2 reports the corresponding estimates and 95% bootstrap confidence intervals. As shown in Figure 2 and Table 2, both BIRD-inspired panels had higher hit rates than the KEGG-derived controls across all intervention orders. NoOmics-BIRD-20 had the highest CR1 and CR2 rates, whereas BIRD-20 had the highest CR3 rate. The BIRD-inspired panels were numerically higher than GPT-20 for CR1 and CR2, while BIRD-20, NoOmics-BIRD-20, and

Table 2. Pooled five-fold CNF-CCCE any-hit rates using q15/q85 base anchors and a normalized logIC50 reduction threshold of 0.30. Values are hit rates with 95% bootstrap confidence intervals.

Panel	CR1	CR2	CR3
BIRD-20	0.122 [0.087, 0.161]	0.535 [0.476, 0.591]	0.769 [0.720, 0.818]
NoOmics-BIRD-20	0.140 [0.101, 0.182]	0.587 [0.531, 0.643]	0.745 [0.692, 0.794]
GPT-20	0.091 [0.059, 0.126]	0.472 [0.413, 0.528]	0.741 [0.689, 0.790]
Random-KEGG-20	0.038 [0.018, 0.063]	0.234 [0.185, 0.283]	0.535 [0.479, 0.591]
KEGG-26	0.004 [0.000, 0.011]	0.126 [0.087, 0.164]	0.395 [0.339, 0.455]

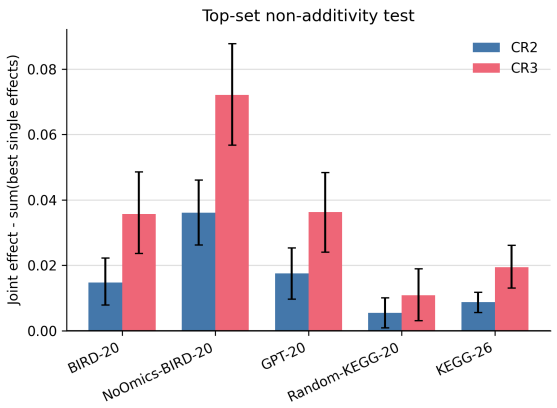

Figure 3. Model-based non-additivity of selected CR2 and CR3 profiles. Bars show the mean difference between the selected joint-intervention effect and the additive single-gene benchmark. Error bars denote 95% bootstrap confidence intervals over held-out high-logIC50 cell lines.

GPT-20 showed similar CR3 rates.

3.2. Model-based non-additivity of selected multi-gene profiles

For a 20-gene panel, CR1 evaluates 20 gene candidates with two anchor directions, whereas CR3 evaluates 1,140 gene triplets with eight mixed low/high assignments. Higher CR3 hit rates may therefore arise partly from the expanded search space.

To assess model-based non-additivity, we compared the mean normalized reduction of each top-ranked CR2 or CR3 profile with the sum of the best single-gene reductions for its constituent genes. For a selected gene set S, we defined

$$\Delta_{\text{nonadd}}(S;x) = \overline{\Delta}(S;x) - \sum_{g\in S} \max_{a\in\{\text{low},\text{high}\}} \overline{\Delta}(\{g\},a;x),$$

where $\overline{\Delta}$ denotes the mean normalized reduction across anchor-perturbation rounds. Positive values indicate that the selected joint intervention produced a larger model-predicted reduction than the additive single-gene benchmark.

Table 3. Model-based non-additivity of selected CR2 and CR3 profiles. The reported q values are Benjamini–Hochberg-adjusted p values from one-sided Wilcoxon signed-rank tests against zero.

Panel	CR	n	Mean	Median	95% CI	q_{BH}
BIRD-20	CR2	286	0.0147	0.0072	[0.0079, 0.0222]	1.53×10^{-4}
BIRD-20	CR3	286	0.0357	0.0211	[0.0236, 0.0486]	5.16×10^{-8}
NoOmics-BIRD-20	CR2	286	0.0361	0.0210	[0.0262, 0.0460]	8.18×10^{-11}
NoOmics-BIRD-20	CR3	286	0.0721	0.0468	[0.0567, 0.0878]	4.91×10^{-17}
GPT-20	CR2	286	0.0175	0.0029	[0.0097, 0.0253]	4.44×10^{-3}
GPT-20	CR3	286	0.0363	0.0168	[0.0240, 0.0484]	6.58×10^{-7}
Random-KEGG-20	CR2	286	0.0055	0.0062	[0.0009, 0.0101]	6.95×10^{-3}
Random-KEGG-20	CR3	286	0.0108	0.0096	[0.0031, 0.0190]	6.89×10^{-3}
KEGG-26	CR2	286	0.0087	0.0048	[0.0056, 0.0118]	6.39×10^{-7}
KEGG-26	CR3	286	0.0194	0.0114	[0.0130, 0.0261]	3.81×10^{-8}

mark.

Figure 3 summarizes the score distributions, and Table 3 reports the corresponding estimates and statistical tests for all five evaluated panels with matched round-level outputs. All reported CR2 and CR3 comparisons had positive mean non-additivity scores, with confidence intervals above zero (Figure 3 and Table 3). NoOmics-BIRD-20 showed the largest mean scores for both CR2 and CR3, followed by GPT-20 and BIRD-20, whereas Random-KEGG-20 and KEGG-26 had smaller positive effects. These results indicate non-additive behavior among the selected profiles under the fitted CNF model. Because the profiles were selected according to their joint counterfactual effects, this analysis does not constitute an unbiased test of biological gene–gene synergy.

3.3. Sensitivity to reduction thresholds and anchor definitions

We assessed whether the panel comparisons were sensitive to the normalized reduction threshold or anchor definition using the same noise-perturbed any-hit rule as the primary evaluation. Figure 4 compares hit rates at normalized reduction thresholds from 0.10 to 0.50 using q15/q85 anchors. As shown in Figure 4, hit rates decreased as the required reduction became more stringent. The BIRD-inspired and GPT-20 panels nevertheless retained higher CR2 and CR3 rates than the KEGG-derived controls across the evaluated thresholds. Differences among BIRD-20, NoOmics-BIRD-20, and GPT-20 were smaller and varied with intervention order and threshold. Figure 5 compares q10/q90, q15/q85, q20/q80, and q25/q75 anchors at a fixed threshold of 0.30.

More extreme anchors generally produced larger counterfactual reductions, whereas anchors closer to the center of the training distribution yielded lower hit rates (Figure 5). The broad separation between the mechanism- or text-guided panels and the KEGG-derived controls was maintained across anchor definitions.

3.4. Lineage structure of CR3 mechanism-axis profiles

Because individual CR3 gene-triplet labels were sparse, each top-ranked hit profile was mapped to the combination of mechanism axes represented by its genes. We then tested the association between these axis-level labels and cell-line lineage among held-out CR3 hit samples. The analysis included lineage categories represented by at least 15 high-logIC50 cell lines. The mechanism-axis definitions and panel-level axis coverage are summarized in Appendix A.5.

Figure 6 shows the distribution of mechanism-axis profiles across the included lineages, and Table 4 reports normalized mutual information (NMI), permutation p values, and Benjamini–Hochberg (BH)-adjusted q values. BIRD-20

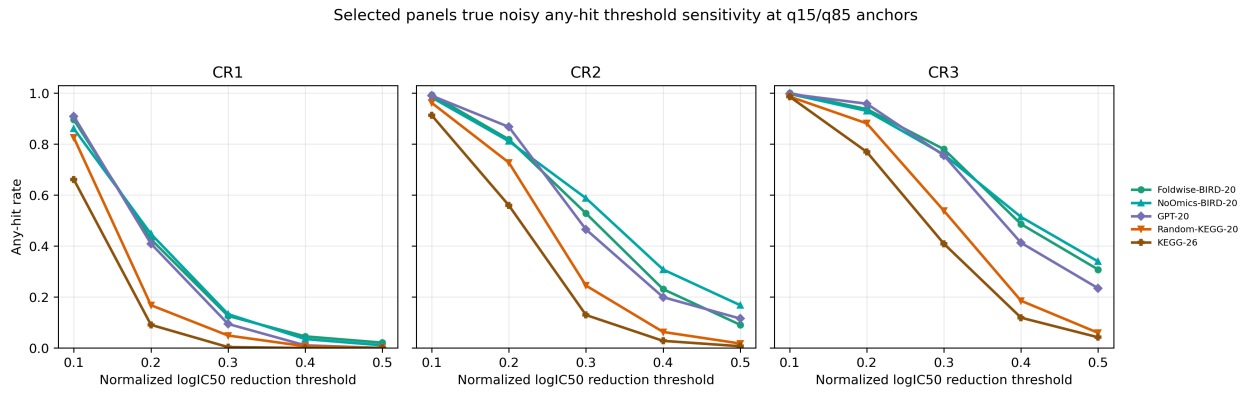

Figure 4. Sensitivity of noise-perturbed any-hit rates to normalized logIC50 reduction thresholds of 0.10, 0.20, 0.30, 0.40, and 0.50 using q15/q85 anchors.

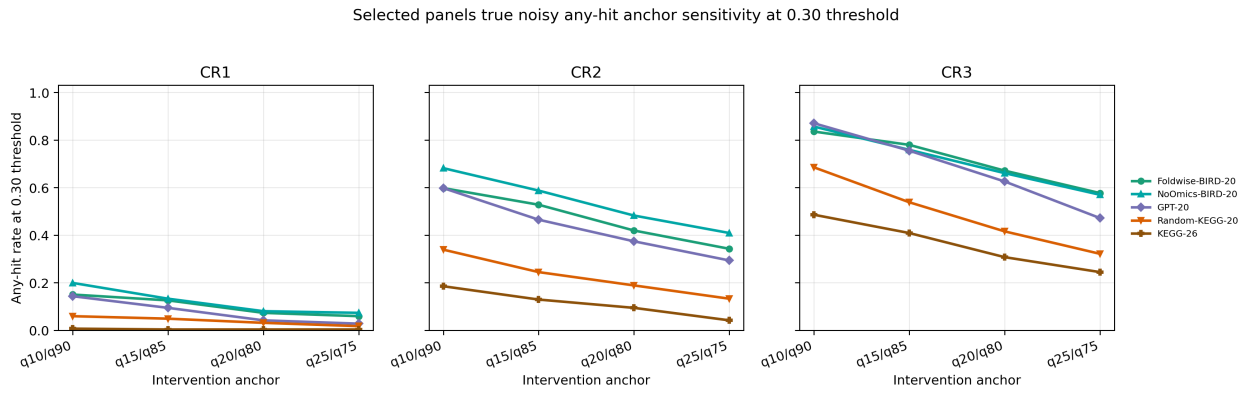

Figure 5. Sensitivity of noise-perturbed any-hit rates to q10/q90, q15/q85, q20/q80, and q25/q75 anchor definitions at normalized logIC50 reduction threshold 0.30.

and NoOmics-BIRD-20 showed significant lineage associations after multiple-testing correction (Table 4). GPT-20 had a similar NMI magnitude but did not reach significance, whereas the KEGG-derived controls showed lower NMI values and nonsignificant permutation tests. The NMI values indicate partial rather than deterministic lineage structure in the BIRD-inspired profiles, as also reflected in the profile distributions in Figure 6. A matched no-hit analysis is reported in Appendix A.7; none of the no-hit strata remained significant after BH correction.

3.5. Counterfactual plausibility and sample-specificity diagnostics

We performed additional diagnostics to assess whether model-generated counterfactual profiles were plausible without treating them as experimentally observed perturbations. For CR3, all panels produced positive mean normalized logIC50 reductions from factual high-logIC50 samples, and the counterfactual expression states showed moderate correlation with observed low-logIC50 samples in the same held-out fold (Appendix A.8). Hit samples also had smaller absolute differences between counterfactual logIC50 and the nearest observed low-logIC50 neighbor than no-hit samples. We assessed cell-line specificity by reassigning each selected CR3 gene set to other held-out high-logIC50 samples within the same fold. The assigned gene set produced a larger predicted reduction in its original cell line than after reassignment for all panels, with the largest self-vs-other differences observed for NoOmics-BIRD-20 and BIRD-20.

4. Discussion

This study shows that compact, mechanism-guided panels provide more informative counterfactual intervention spaces than KEGG-derived controls within the CNF-CCCE framework. BIRD-20 and NoOmics-BIRD-20 achieved consistently higher hit rates than Random-KEGG-20 and KEGG-26 across CR1–CR3, while remaining competitive with GPT-20 for higher-order interventions (Figure 2 and Table 2). The broad comparison was maintained across alternative reduction thresholds and anchor definitions (Figures 4 and 5). The similar performance of BIRD-20 and NoOmics-BIRD-20 further indicates that curated mechanism structure

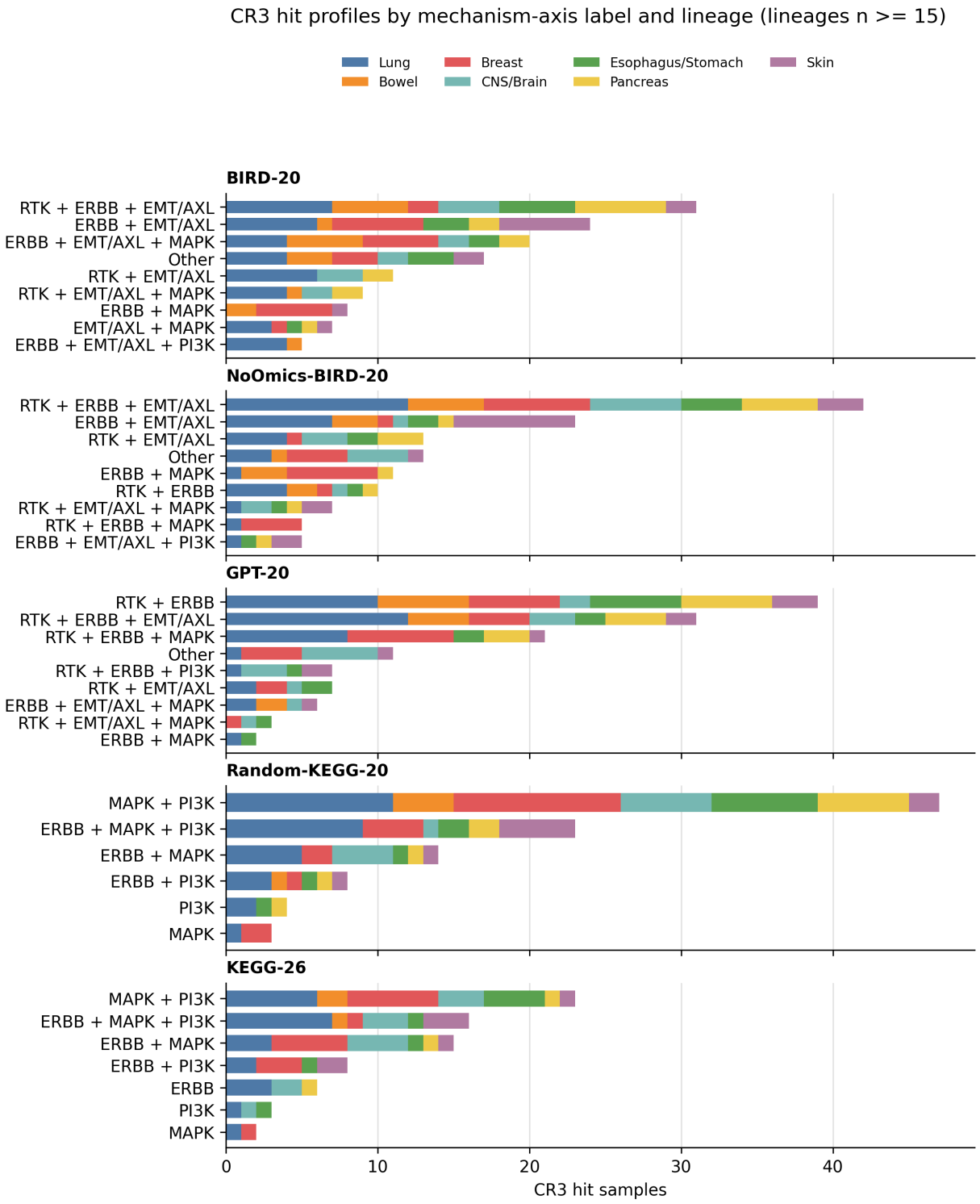

Figure 6. Distribution of CR3 mechanism-axis hit profiles across cell-line lineages. Only lineage categories represented by at least 15 high-logIC50 cell lines were included.

provides a strong foundation for defining the intervention space, even without cohort-derived omics evidence.

The panel-construction strategies emphasized different resistance axes. The most frequent CR3 profile for BIRD-20 and NoOmics-BIRD-20 was bypass RTK + EGFR/ERBB +

EMT/AXL, occurring in 31 of 132 and 42 of 129 hit samples, respectively. Representative combinations included *AXL–EGFR–MET*, *AXL–ERBB2–MET*, and *AXL–EGFR–ERBB3*. In contrast, GPT-20 most frequently selected bypass RTK + EGFR/ERBB profiles, often involving *FGFR1*. These patterns are consistent with established EGFR-TKI resistance mechanisms. MET activation can maintain ERBB3-dependent PI3K signaling despite EGFR inhibition, while AXL activation and EMT-associated states can reduce erlotinib sensitivity (Engelman et al., 2007; Bean et al., 2007; Zhang et al., 2012). FGFR signaling provides a complementary bypass route and has motivated combined EGFR and FGFR inhibition in resistant models (Quintanal-Villalonga et al., 2019; Raoof et al., 2019). The BIRD-inspired and text-curated panels therefore captured related but distinct representations of receptor-level bypass signaling.

Several cell-line-level results were concordant with known RTK dependencies. The *AXL–MET–EGFR* profile was selected in EBC-1, a MET-amplified lung cancer model, and Hs 746T, a MET-driven gastric cancer model. Both cell lines have established sensitivity to MET pathway inhibition (Lutterbach et al., 2007; Smolen et al., 2006). These examples illustrate how the framework can recover known molecular states while adding sample-specific hypotheses involving additional resistance components. The recurrent high-*EGFR* direction in profiles such as low-*AXL*/low-*MET*/high-*EGFR* is best understood as an EGFR-dependent state with reduced bypass signaling. It does not represent a direct proposal to increase EGFR expression.

The selected CR2 and CR3 profiles also showed positive model-based non-additivity relative to additive single-gene benchmarks (Figure 3 and Table 3). The largest values were observed for NoOmics-BIRD-20, particularly among recurrent AXL–MET–EGFR/ERBB modules. This suggests that the fitted CNF represents joint intervention effects that are not captured by summing the strongest single-gene effects. The lineage analysis further showed that BIRD-20 and NoOmics-BIRD-20 profiles retained partial lineage structure after gene triplets were collapsed into mechanism-axis labels (Figure 6 and Table 4). At the same time, the dominant bypass RTK + EGFR/ERBB + EMT/AXL axis appeared across multiple lineages, indicating that related resistance states can recur across histologically distinct cancers. Nearest-low analyses further suggested that hit counterfactuals were closer in response value to observed low-logIC50 neighbors than no-hit counterfactuals, but same-lineage recovery was weak. Thus, we interpret cell-line specificity primarily through self-vs-other reassignment tests rather than through recovery of lineage-matched low-response counterparts.

Table 4. Association between cell-line lineage and CR3 mechanism-axis hit-profile labels. Permutation p values were adjusted using the Benjamini–Hochberg procedure across valid tests.

Panel	n	Lineages	Axis labels	NMI	p_{perm}	q_{BH}
BIRD-20	132	7	14	0.196	0.013	0.030
NoOmics-BIRD-20	129	7	14	0.211	0.001	0.006
GPT-20	127	7	14	0.190	0.090	0.134
Random-KEGG-20	99	7	13	0.092	0.703	0.715
KEGG-26	73	7	13	0.135	0.715	0.715

The framework remains dependent on the selected panel, mechanism-informed ordering, intervention anchors, and fitted CNF distribution. It is also based on observational cell-line data and deterministic latent-coordinate interventions. Foldwise panel reconstruction reduces outcome-informed selection bias, while random and pathway-derived controls help separate mechanism-guided selection from panel-size and pathway-membership effects. The resulting profiles provide a focused set of hypotheses for gene perturbation and drug-combination experiments, particularly in cell lines with known MET-, AXL-, ERBB-, or FGFR-associated signaling states.

5. Conclusion

We present a mechanism-guided framework for prioritizing sample-specific counterfactual gene-set hypotheses from continuous pharmacogenomic data. BIRD-inspired evidence integration defines compact, mechanism-diverse intervention panels, while CNF-based attribution evaluates one-, two-, and three-gene interventions in continuous gene-expression and erlotinib logIC50 space.

Under leakage-aware five-fold panel reconstruction, BIRD-20 and NoOmics-BIRD-20 showed consistently higher hit rates than the KEGG-derived controls and remained competitive with GPT-20, particularly for higher-order interventions. These comparisons were broadly preserved across alternative reduction thresholds and anchor definitions. The BIRD-inspired panels most frequently prioritized bypass RTK + EGFR/ERBB + EMT/AXL profiles, whereas GPT-20 more often emphasized bypass RTK + EGFR/ERBB combinations involving *FGFR1*. Selected multi-gene profiles also exhibited positive model-based non-additivity relative to additive single-gene benchmarks.

Overall, the results indicate that mechanism-guided and text-curated panels provide complementary counterfactual search spaces, with greater utility than canonical pathway selection alone in the present CNF-based evaluation. The resulting profiles are intended as testable hypotheses for targeted gene perturbation and drug-combination studies rather than as direct evidence of causal mechanisms or treatment efficacy.

Acknowledgements

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No. RS-2025-25462804). The authors thank Hanbi Lee and Ahyoung Choi of the MOGAM Institute for Biomedical Research for their assistance with data collection and for providing biological expertise.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483 (7391):603–607, 2012. doi: 10.1038/nature11003. URL https://doi.org/10.1038/nature11003.

Bean, J., Brennan, C., Shih, J.-Y., Riely, G., Viale, A., Wang, L., Chitale, D., Motoi, N., Szoke, J., Broderick, S., Balak, M., Chang, W.-C., Yu, C.-J., Gazdar, A., Pass, H., Rusch, V., Gerald, W., Huang, S.-F., Yang, P.-C., Miller, V., Ladanyi, M., Yang, C.-H., and Pao, W. MET amplification occurs with or without T790M mutations in EGFR-mutant lung tumors with acquired resistance to gefitinib or erlotinib. *Proceedings of the National Academy of Sciences of the United States of America*, 104(52):20932–20937, 2007. doi: 10.1073/pnas.0710370104.

Deng, L., Cai, Y., Zhang, W., Yang, W., Gao, B., and Liu, H. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. *Journal of Chemical Information and Modeling*, 60(10):4497–4505, 2020. doi: 10.1021/acs.jcim.0c00331. URL https://doi.org/10.1021/acs.jcim.0c00331.

Engelman, J. A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J. O., Lindeman, N., Gale, C. M., Zhao, X., Christensen, J., Kosaka, T., Holmes, A. J., Rogers, A. M., Cappuzzo, F., Mok, T., Lee, C., Johnson, B. E., Cantley, L. C., and Jänne, P. A. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science*, 316(5827):1039–1043, 2007. doi: 10.1126/science.1141478.

Feng, Y., Zhou, B., Lin, W., and Roth, D. BIRD: A trustworthy Bayesian inference framework for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fAAaT826Vv.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., Mironenko, T., et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016. doi: 10.1016/j.cell.2016.06.017. URL https://doi.org/10.1016/j.cell.2016.06.017.

Javaloy, A., Sánchez-Martín, P., and Valera, I. Causal normalizing flows: From theory to practice. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=QIFoCI7ca1.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., and Ideker, T. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, 2020. doi: 10.1016/j.ccell.2020.09.014. URL https://doi.org/10.1016/j.ccell.2020.09.014.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8. URL https://doi.org/10.1038/s41592-019-0494-8.

Lutterbach, B., Zeng, Q., Davis, L. J., Hatch, H., Hang, G., Kohl, N. E., Gibbs, J. B., and Pan, B. S. Lung cancer cell lines harboring MET gene amplification are dependent on Met for growth and survival. *Cancer Research*, 67(5):2081–2088, 2007. doi: 10.1158/0008-5472.CAN-06-3495.

Quintanal-Villalonga, A., Molina-Pinelo, S., Cirauqui, C., Ojeda-Marquez, L., Marrugal, A., Suarez, R., Conde, E., Ponce-Aix, S., Enguita, A. B., Carnero, A., Ferrer, I., and Paz-Ares, L. FGFR1 cooperates with EGFR in lung cancer oncogenesis, and their combined inhibition shows improved efficacy. *Journal of Thoracic Oncology*, 14(4):641–655, 2019. doi: 10.1016/j.jtho.2018.12.021. URL https://doi.org/10.1016/j.jtho.2018.12.021.

Raoof, S., Mulford, I. J., Frisco-Cabanos, H., Nangia, V., Timonina, D., Labrot, E., Hafeez, N., Bilton, S. J., Drier, Y., Ji, F., et al. Targeting FGFR overcomes EMT-mediated resistance in EGFR mutant non-small cell lung cancer. *Oncogene*, 38(37):6399–6413, 2019. doi: 10.1038/s41388-019-0887-2. URL https://doi.org/10.1038/s41388-019-0887-2.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturba-

tions with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Schrod, S., Zacharias, H. U., Beißbarth, T., Hauschild, A.-C., and Altenbuchinger, M. CODEX: COunterfactual deep learning for the in silico EXploration of cancer cell line perturbations. *Bioinformatics*, 40(Supplement_1):i91–i99, 2024. doi: 10.1093/bioinformatics/btae261. URL https://doi.org/10.1093/bioinformatics/btae261.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 2019. doi: 10.1093/bioinformatics/btz318. URL https://doi.org/10.1093/bioinformatics/btz318.

Shin, J., Piao, Y., Bang, D., Kim, S., and Jo, K. DRPreter: Interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. *International Journal of Molecular Sciences*, 23 (22):13919, 2022. doi: 10.3390/ijms232213919. URL https://doi.org/10.3390/ijms232213919.

Smolen, G. A., Sordella, R., Muir, B., Mohapatra, G., Barmettler, A., Archibald, H., Kim, W. J., Okimoto, R. A., Bell, D. W., Sgroi, D. C., Christensen, J. G., Settleman, J., and Haber, D. A. Amplification of MET may identify a subset of cancers with extreme sensitivity to the selective tyrosine kinase inhibitor PHA-665752. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2316–2321, 2006. doi: 10.1073/pnas.0508776103.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., and Garnett, M. J. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1): D955–D961, 2013. doi: 10.1093/nar/gks1111. URL https://doi.org/10.1093/nar/gks1111.

Zhang, Z., Lee, J. C., Lin, L., Olivas, V., Au, V., LaFramboise, T., Abdel-Rahman, M., Wang, X., Levine, A. D., Rho, J. K., Choi, Y. J., Choi, C. M., Kim, S. W., Jang, S. J., Park, Y. S., Kim, W. S., Lee, D. H., Lee, J. S., Miller, V. A., Arcila, M., Ladanyi, M., Moonsamy, P., Sawyers, C., Boggon, T. J., Ma, P. C., Costa, C., Tenen, D. G., Halmos, B., and Bivona, T. G. Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genetics*, 44(8):852–860, 2012. doi: 10.1038/ng.2330.

Zhao, R., Zhang, L., Zhu, S., Lu, Z., Dong, Z., Zhang, C., Xu, J., Geng, Z., and He, Y. Conditional counterfactual causal effect for individual attribution. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2519–2528, 2023.

A. Additional implementation details

A.1. Panel construction and graph artifacts

Table 5 summarizes the five panels used in the main evaluation. BIRD-20 was reconstructed within each training fold, whereas NoOmics-BIRD-20, GPT-20, Random-KEGG-20, and KEGG-26 followed their predefined panel-construction protocols. For all panels, the graph was used to provide a variable ordering for CNF training. No edge-wise graph regularization was used.

Table 5. Panel and graph artifacts used in the five-fold CNF-CCCE evaluation.

Panel	Size	Fold-specific	Construction and graph artifact
BIRD-20	20	Yes	BIRD-inspired evidence scores and 25-to-20 compression were recomputed using the training split of each fold. The fold-specific graph matrices are stored as fold-level BIRD-20 CNF graph CSV artifacts.
NoOmics-BIRD-20	20	No	Cohort-derived expression, mutation, family-activity, and multi-source consistency factors were removed from CPT integration; pathway/topology and curated mechanism priors were retained. The same 20-gene panel was evaluated across folds.
GPT-20	20	No	Text-curated EGFR-TKI response and resistance panel evaluated with the same CNF-CCCE procedure and q15/q85 intervention anchors.
Random-KEGG-20	20	No	One fixed size-matched panel sampled without replacement from the KEGG-26 candidate space using selection seed 2026.
KEGG-26	26	No	Canonical pathway-union panel derived from erlotinib-relevant KEGG signaling context. It is a pathway-union comparator rather than a size-matched control.

Table 6. Gene composition of comparator panels used in the five-fold CNF-CCCE evaluation.

Panel	Genes
NoOmics-BIRD-20	*HGF, NRAS, KRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, IGF1R, PDGFRB, SRC, BRAF, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP6, AKT1*
GPT-20	*EGFR, ERBB2, ERBB3, MET, HGF, AXL, GAS6, KRAS, NRAS, BRAF, MAP2K1, MAPK1, MAPK3, PIK3CA, AKT1, PTEN, BCL2L11, IGF1R, FGFR1, SRC*
Random-KEGG-20	*EGF, TGFA, EGFR, GRB2, SOS1, SOS2, HRAS, KRAS, PIK3R3, PIK3CB, PIK3CD, PIK3R1, PIK3R2, ARAF, RAF1, AKT1, AKT2, AKT3, MAP2K2, BAD*
KEGG-26	*EGF, TGFA, EGFR, GRB2, SOS1, SOS2, HRAS, KRAS, PIK3R3, PIK3CB, PIK3CD, PIK3R1, PIK3R2, ARAF, RAF1, AKT1, AKT2, AKT3, MAP2K2, BAD, NRAS, PIK3CA, BRAF, MAP2K1, MAPK1, MAPK3*

A.2. BIRD-inspired CPT and panel-compression summary

The BIRD-inspired panel was implemented as a deterministic scoring and compression pipeline rather than as a learned Bayesian posterior model. For stage-1 scoring, normalized evidence factors were discretized into low, medium, and high states and combined with prespecified weakly informative CPTs over high, moderate, and weak relevance classes. The factor-specific relevance distributions were linearly pooled with fixed weights, yielding the score $R(g)$ used. These CPT values and weights were fixed before downstream CNF-CCCE evaluation and were not tuned on held-out hit rates.

The main evidence factors were expression–response association, mutation–response association, KEGG centrality, EGFR/ERBB proximity, curated mechanism membership, family-level activity support, and multi-source consistency. The last term was treated as a corroboration score rather than an independent data source because it is derived from the other evidence channels. NoOmics-BIRD-20 removed the cohort-derived expression, mutation, family-activity, and multi-source consistency factors, leaving topology and curated mechanism evidence. For stage-2 scoring, the curated 25-gene shortlist was reranked using the same CPT-style integration idea with panel-level topology, proximity, family support, and representativeness features.

The 25-gene mechanism shortlist used before the final compression step is shown in Table 7. This shortlist was predefined from the 69 unique curated seed genes to cover the main erlotinib response and resistance axes. It was then reranked within each training fold using the stage-2 CPT scores.

The final 25-to-20 compression was performed by exhaustive enumeration. The implemented objective retained high

stage-2 scores, required at least one gene from each coverage family, rewarded broader family coverage and within-family representativeness, and penalized overrepresentation of any single family. Because the shortlist contained 25 genes, the implementation enumerated all $\binom{25}{5} = 53{,}130$ possible five-gene removal sets, discarded subsets that failed the family-coverage constraint, and selected the feasible subset with the largest objective.

A.3. Relation to the original BIRD framework

The original BIRD framework was proposed for trustworthy Bayesian inference with large language models (Feng et al., 2025). In that setting, an LLM is used to generate factor descriptions and factor values for a scenario, sample complete factor-value combinations, and provide verbal probability judgments such as "unlikely," "likely," or "very likely." Those verbal probabilities are mapped to numerical supervision values and used to learn CPT parameters for a Bayesian inference model.

Our use of BIRD is more limited. We do not use LLM-generated factor combinations, LLM verbal-probability supervision, or learned CPT parameters. Instead, we adapt the factor-based evidence-integration idea to pharmacogenomic panel construction: biological knowledge defines the evidence factors and mechanism families, while prespecified CPT-style mappings convert heterogeneous gene-level evidence into a relevance score. Thus, BIRD-20 should be interpreted as a BIRD-inspired, mechanism-guided panel rather than as a direct implementation of the original BIRD learning algorithm.

A.4. Foldwise BIRD-20 panel variability

The foldwise BIRD-20 panels shared a stable core and varied in a small number of positions across folds. The five fold-specific panels were:

- Fold 0: *GAS6, HGF, KRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, FGFR2, PDGFRB, BRAF, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1*.
- Fold 1: *GAS6, HGF, NRAS, KRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, IGF1R, SRC, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1*.
- Fold 2: *GAS6, HGF, NRAS, KRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, PDGFRB, SRC, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1*.
- Fold 3: *GAS6, HGF, NRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, FGFR2, PDGFRB, BRAF, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1*.
- Fold 4: *GAS6, HGF, NRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, FGFR2, PDGFRB, BRAF, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1*.

Thus, the foldwise construction retained the major erlotinib mechanism axes while allowing a small number of training-split-dependent substitutions.

A.5. Comparator panel gene composition and axis coverage

Table 6 lists the non-foldwise comparator panels used in the main evaluation. For BIRD-20, the fold-specific gene lists are given in Appendix A.4.

Table 8 summarizes how many genes from each panel map to the six erlotinib mechanism axes used in the lineage analysis. The same gene-to-axis mapping was used for Figure 6. Because each BIRD-20 fold had the same axis counts, BIRD-20 is shown as a single row for the axis-coverage summary.

Here, BIM denotes apoptotic priming/BIM-axis genes, RTK denotes bypass receptor tyrosine kinase signaling, EGFR denotes direct EGFR/ERBB signaling, EMT denotes EMT/AXL-associated state, MAPK denotes MAPK escape, and PI3K denotes PI3K/AKT/PTEN survival signaling. This summary highlights that the BIRD-inspired and GPT-20 panels distribute genes across all six axes, whereas the KEGG-derived controls are concentrated in canonical EGFR, MAPK, PI3K, and BIM-axis components and contain no bypass-RTK or EMT/AXL-state genes under this mapping.

A.6. CNF training and implementation details

All panels used the same ordering-based CNF configuration. The flow used neural spline flow transforms with four layers, hidden dimensions $[32, 32, 32]$, exponential linear unit (ELU) activations, a standard normal base distribution, and no adjacency or edge-wise regularization. Models were trained with Adam (learning rate 10^{-3}, weight decay 10^{-5}, batch size 256) for 400 epochs with random seed 42 and no early stopping; checkpoints were saved every epoch. For downstream counterfactual scoring, the checkpoint with the lowest evaluation negative log-likelihood was used when available, otherwise the final checkpoint was used. All variables were standardized using training-fold statistics before CNF training and inverse-transformed to the original logIC50 scale before computing normalized reductions. The numerical constant in the reduction denominator was $\epsilon = 10^{-8}$.

A.7. No-hit counterpart of the CR3 axis-lineage analysis

We repeated the CR3 axis-lineage association analysis after restricting to held-out high-logIC50 samples whose selected CR3 profile did not satisfy the primary any-hit criterion. The same lineage filter, gene-to-axis mapping, NMI statistic, 2,000 label permutations, and BH correction family were used as in the hit-profile analysis. Figure 7 shows the no-hit axis-lineage distributions, and Table 9 reports the corresponding NMI permutation tests.

Unlike the hit-profile analysis, none of the no-hit strata in Table 9 showed a significant lineage association after BH correction. These results support interpreting the main lineage signal as a property of successful model-generated counterfactual profiles rather than as a global property of all selected CR3 labels.

A.8. Counterfactual plausibility and sample-specificity diagnostics

We compared CR3 counterfactual states with observed low-logIC50 samples as a plausibility diagnostic rather than as direct causal validation. For each held-out high-logIC50 sample, the nearest observed low-logIC50 sample was identified in the panel-gene expression space using Pearson correlation. Table 10 summarizes the main diagnostics. Hit counterfactuals were closer in logIC50 value to their nearest low-logIC50 neighbors than no-hit counterfactuals, while same-lineage recovery was weak across panels.

Figure 8 visualizes the self-vs-other reassignment diagnostic. For the reassignment test, each selected CR3 gene set was assigned to other high-logIC50 samples within the same fold, while the receiving sample's stored best low/high direction for that gene set was used. This is a conservative gene-set specificity diagnostic because the reassigned sample was allowed to re-optimize intervention direction.

Table 7. Curated 25-gene erlotinib/EGFR-TKI mechanism shortlist used before 25-to-20 compression.

Rank	Gene	Mechanism family	Short reason
1	*EGFR*	Direct EGFR dependency	Direct target
2	*ERBB2*	Direct EGFR dependency	ERBB family coreceptor
3	*ERBB3*	Direct EGFR dependency	ERBB family coreceptor
4	*ERBB4*	Direct EGFR dependency	ERBB family coreceptor
5	*AXL*	Bypass RTK activation	Bypass RTK / EMT
6	*GAS6*	Bypass RTK activation	AXL ligand
7	*MET*	Bypass RTK activation	Bypass RTK
8	*HGF*	Bypass RTK activation	MET ligand
9	*FGFR1*	Bypass RTK activation	Bypass RTK
10	*FGFR2*	Bypass RTK activation	Bypass RTK
11	*PDGFRA*	Bypass RTK activation	Bypass RTK
12	*PDGFRB*	Bypass RTK activation	Bypass RTK
13	*IGF1R*	Bypass RTK activation	Bypass RTK
14	*SRC*	Bypass RTK activation	Signaling bypass hub
15	*KRAS*	MAPK escape	MAPK escape
16	*NRAS*	MAPK escape	MAPK escape
17	*BRAF*	MAPK escape	MAPK escape
18	*MAPK3*	MAPK escape	MAPK signaling core
19	*DUSP4*	MAPK escape	MAPK feedback regulator
20	*DUSP6*	MAPK escape	MAPK feedback regulator
21	*PIK3CA*	PI3K/AKT/PTEN axis	PI3K pathway
22	*AKT1*	PI3K/AKT/PTEN axis	AKT pathway
23	*PTEN*	PI3K/AKT/PTEN axis	PI3K brake / resistance context
24	*BCL2L11*	Apoptotic priming/BIM axis	Apoptotic priming
25	*FN1*	EMT/AXL state	EMT / mesenchymal state

Table 8. Mechanism-axis coverage of evaluated panels. Counts denote the number of panel genes assigned to each axis.

Panel	BIM	RTK	EGFR	EMT	MAPK	PI3K
BIRD-20	1	4	4	3	5	3
NoOmics-BIRD-20	1	5	4	2	5	3
GPT-20	1	5	3	2	6	3
Random-KEGG-20	1	0	3	0	8	8
KEGG-26	1	0	3	0	13	9

Table 9. Association between cell-line lineage and CR3 mechanism-axis no-hit-profile labels. Permutation p values were adjusted using the Benjamini–Hochberg procedure across valid no-hit tests.

Panel	n	Lineages	Axis labels	NMI	p_{perm}	q_{BH}
BIRD-20	35	6	13	0.428	0.214	0.612
NoOmics-BIRD-20	38	5	15	0.367	0.412	0.612
GPT-20	40	7	16	0.411	0.612	0.612
Random-KEGG-20	68	7	7	0.141	0.565	0.612
KEGG-26	94	7	7	0.134	0.253	0.612

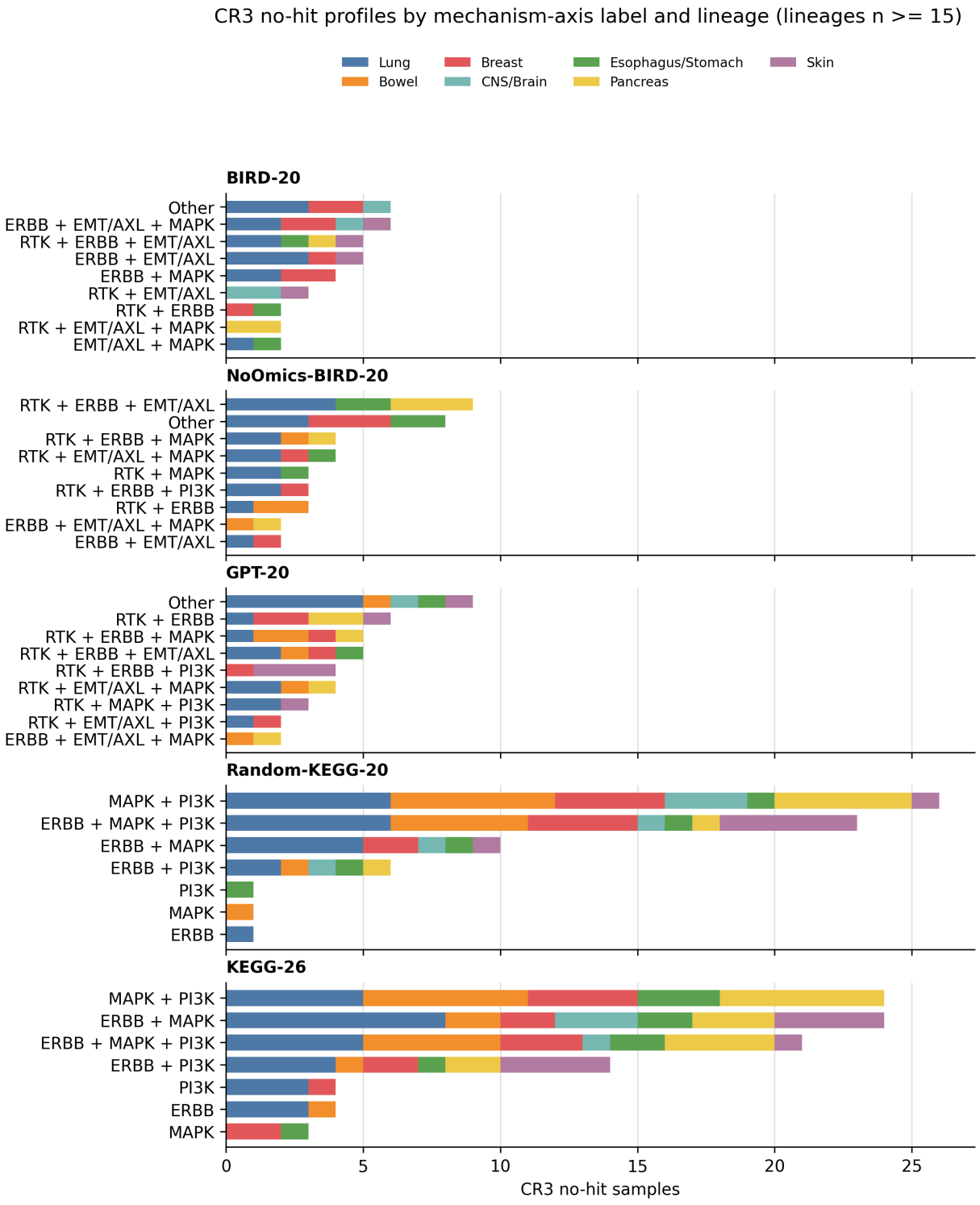

Figure 7. Distribution of CR3 mechanism-axis no-hit profiles across cell-line lineages. Only lineage categories represented by at least 15 high-logIC50 cell lines were included.

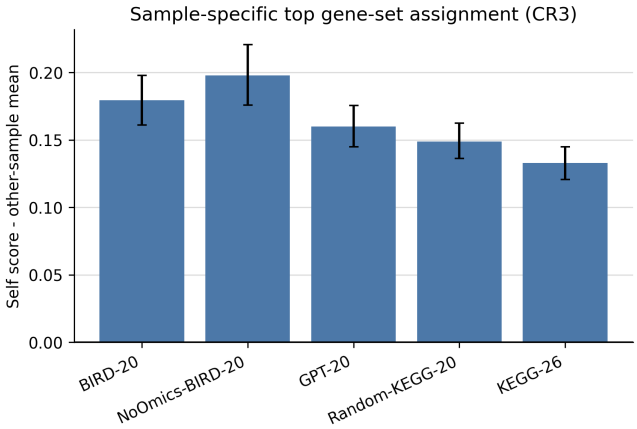

Figure 8. CR3 self-vs-other reassignment diagnostic. Bars show the mean difference between the selected gene set's normalized reduction in its assigned held-out high-logIC50 cell line and the mean reduction after reassignment to other held-out high-logIC50 samples in the same fold. Error bars denote 95% bootstrap confidence intervals.

Table 10. CR3 counterfactual plausibility and sample-specificity diagnostics. Reduction denotes mean normalized logIC50 reduction. Corr. denotes mean Pearson correlation to the nearest observed low-logIC50 sample. Error diff. is hit minus no-hit absolute difference between counterfactual logIC50 and nearest low-factual logIC50; negative values indicate smaller error among hit samples. Same lineage is the fraction of nearest low-logIC50 matches with the same OncoTree lineage. Self–other is the mean reduction advantage of each selected CR3 gene set in its assigned cell line versus reassignment to other held-out high-logIC50 samples.

Panel	Reduction	Corr.	Error diff.	Same lineage	Self–other
BIRD-20	0.378	0.593	-0.646	0.112	0.179
NoOmics-BIRD-20	0.408	0.591	-0.708	0.129	0.198
GPT-20	0.373	0.589	-0.562	0.098	0.160
Random-KEGG-20	0.303	0.577	-0.684	0.136	0.149
KEGG-26	0.262	0.534	-0.664	0.105	0.133