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

Anonymous Authors[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 rather than to directly implement the original BIRD framework. Given a fixed panel and biologically adjusted ordering, 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. Conditional on full-cohort-selected panels, five-fold held-out evaluation showed that BIRD-20 had higher CNF-CCCE hit rates than the KEGG-26 pathway-union comparator and similar or modestly higher rates than the text-derived GPT-20 comparator, especially for two- and three-gene interventions. The resulting profiles should be interpreted as sample-level counterfactual hypotheses for downstream biological validation, not as experimentally verified causal mechanisms.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

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, copy number, and drug-structure 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 DR-Preter 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 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 dimensionality. Transcriptomic datasets contain tens of thousands of genes, making direct causal modeling unstable 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 should reduce the gene space while preserving mechanism diversity.

We address these challenges with a mechanism-guided framework inspired by Bayesian inference from abduction and deduction (BIRD). The original BIRD framework was introduced as a trustworthy Bayesian inference framework for large language models (Feng et al., 2025). However, our setting differs from the original setting in an important way: we do not have labeled gene-level or sample-level causal-driver annotations. Thus, we do not treat the original BIRD method as a supervised posterior learner. Instead, we use a BIRD-inspired conditional probability table (CPT) integration layer as a transparent evidence aggregation scheme for unlabeled pharmacogenomic data. Biological literature and pathway knowledge define Erlotinib mechanism families, whereas expression, mutation, and pathway features provide structured gene-level evidence. These evidence scores are integrated with a family-aware optimization step to produce BIRD-20, a compact BIRD-inspired Erlotinib panel designed to preserve mechanism coverage and reduce within-family redundancy.

On top of this panel, we construct a biologically adjusted directed acyclic graph (DAG), train a CNF using a biologically motivated ordering, and approximate CCCE-style scores for one-, two-, and three-gene intervention sets. Because the response-reducing direction is not known in advance, each candidate intervention is evaluated at both low- and high-expression anchors, and the direction that most reduces predicted Erlotinib logIC50 is selected. Our contributions are threefold. First, we adapt BIRD-inspired evidence integration to an unlabeled omics setting by combining expression, mutation, KEGG topology, and curated Erlotinib biology into a compact, mechanism-balanced gene panel. Second, we approximate CCCE-style individual attribution with CNFs, extending counterfactual causal attribution to continuous transcriptomic and drug-response variables and to higher-order gene sets. Third, we compare BIRD-20 with both a text-centric GPT-20 panel and a pathway-union KEGG-26 panel under the same downstream held-out sample-specific counterfactual evaluation, conditional on fixed panels. Overall, the framework combines text-derived biological structure with data-driven evidence integration and generative counterfactual hypothesis prioritization.

2. Method

2.1. Problem setup and data

Our goal is to prioritize genes or gene sets that plausibly contribute to an individual cell line's Erlotinib response under the assumed model and to report them as counterfactual hypotheses. Let $Y \in \mathbb{R}^n$ denote the continuous drug-response vector, defined as Erlotinib logIC50, and let $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ denote transcriptomic features for the same cell lines. The prepared analysis matrix contains transcripts per million (TPM) expression for $n = 571$ GDSC-matched cancer cell lines and $p = 19{,}193$ genes, matched to Erlotinib drug-response measurements from Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2013). Cell-line metadata were aligned through common identifiers, and gene annotation was standardized using human gene nomenclature (HGNC)-based mapping. Hotspot mutation evidence, KEGG pathway topology, and curated EGFR tyrosine kinase inhibitor (EGFR-TKI) biology were used only as auxiliary gene-level evidence for panel construction. We denote one-, two-, and three-gene intervention orders as CR1, CR2, and CR3, respectively. In the absence of experimentally validated sample-specific driver labels, we interpret the reported CR1-CR3 profiles as model-based counterfactual hypotheses conditioned on the selected panel, graph ordering, and CNF model.

Panel construction was performed once using the full matched cohort to define fixed BIRD-20, GPT-20, and KEGG-26 panels before downstream five-fold CNF training; BIRD-20 refers to the BIRD-inspired panel proposed in this study. Thus, the cross-validation results evaluate the downstream CNF-CCCE attribution procedure conditional on these fixed panels; they are not a nested or end-to-end held-out validation of the full gene-panel selection pipeline.

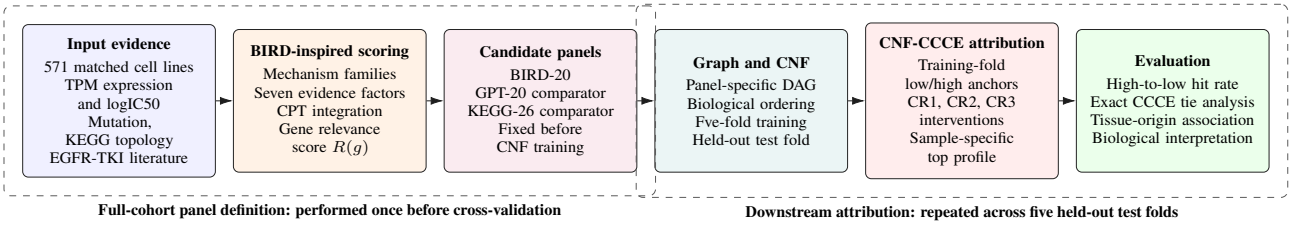

Figure 1. Overall analysis workflow. Full-cohort BIRD-inspired evidence integration defines the fixed BIRD-20 panel, together with GPT-20 and KEGG-26 comparators, after which each panel is evaluated by five-fold CNF-based counterfactual attribution on held-out high-IC50 samples.

2.2. Overview of the framework

The framework has two main stages. The first stage performs drug-related gene selection and DAG construction using knowledge-guided, BIRD-inspired evidence integration (Feng et al., 2025). The second stage performs individual-level CCCE approximation using CNF-based counterfactual generation (Javaloy et al., 2023; Zhao et al., 2023). The pipeline is designed for an unlabeled omics setting: mechanism families are specified from prior biological knowledge, evidence scores are computed from expression, mutation, and pathway structure, and a generative causal model is then used for continuous counterfactual analysis.

The workflow is summarized in Fig. 1. First, a BIRD-inspired evidence integration step combines curated Erlotinib biology with expression-response association, mutation support, and KEGG pathway topology. Second, a family-aware compression step constructs a compact 20-gene panel and a biologically adjusted DAG. Third, a CNF is trained for continuous counterfactual generation, and CCCE-style scores are approximated for one-, two-, and three-gene intervention sets.

2.3. Gene selection by BIRD-inspired Bayesian evidence integration

2.3.1. Mechanism families and input evidence

For Erlotinib, prior knowledge was organized into six mechanism families used as coverage constraints: direct EGFR signaling (f_1), bypass RTK activation (f_2), MAPK-mediated escape (f_3), PI3K/AKT/PTEN-related survival signaling (f_4), EMT/AXL-associated resistance (f_5), and apoptotic priming including the BIM axis (f_6), where BIM is encoded by *BCL2L11*. Although an EGFR-centered expression-program seed list was also curated, it substantially overlapped with the direct EGFR dependency family. We therefore used it only as auxiliary expression evidence and did not count it as an additional coverage family. These families were used both to define evidence factors and to constrain the final gene panel so that it covered multiple important mechanisms rather than overrepresenting a single pathway.

For each gene $g \in \{1, \ldots, p\}$, evidence was summarized into seven fixed factors:

- Expression association (F_1): tissue-adjusted association between gene expression and Erlotinib logIC50.
- Mutation association (F_2): mutation-enrichment support where available; missing or sparse mutation evidence was treated conservatively.
- KEGG centrality (F_3): graph centrality within Erlotinib-relevant KEGG signaling context.
- EGFR proximity (F_4): graph proximity to the EGFR/ERBB core.
- Literature membership (F_5): membership in curated Erlotinib resistance or sensitivity mechanism families.
- Family activity support (F_6): support inherited from the activity of the gene's mechanism family.
- Multi-omics consistency (F_7): agreement between transcriptomic, mutation, pathway, and literature evidence.

2.3.2. Genome-wide BIRD-inspired CPT integration

Table 1 shows a representative evidence-factor matrix for a small set of genes. The table is intended to illustrate the structure of the inputs to the BIRD-inspired CPT integration step: each column corresponds to a gene, and each row corresponds to one normalized evidence factor. Missing or unavailable mutation evidence is shown as "–" and was treated conservatively during scoring.

For each evidence factor F_j, the continuous evidence score was discretized into one of three states: low, medium, or high. The relevance outcome O was represented by three classes: highly relevant, moderately relevant, and weakly relevant. The conditional probability table (CPT) encoded the probability of observing each factor state given a relevance outcome, $P(F_j = s \mid O = o)$. We converted it by

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-omics consistency	0.33	0.17	0.17	0.17	0.17	0.17

Bayes rule:

$$P(O = o \mid F_j = s) = \frac{P(F_j = s \mid O = o)P(O = o)}{\sum_{o'} P(F_j = s \mid O = o')P(O = o')}.$$

Because each gene has one observed state for each factor, $P(F_j = s \mid g)$ was deterministic: it is one for the gene's assigned state and zero otherwise. The gene-level posterior was then computed as a weighted mixture over factors:

$$P(O = o \mid g) = \sum_j \alpha_j \sum_s P(O = o \mid F_j = s)P(F_j = s \mid g),$$

where α_j is the normalized factor weight. The integrated relevance score of g was

$$R(g) = P(O = h \mid g) + 0.5P(O = m \mid g) + 0P(O = w \mid g),$$

where h, m, and w denote highly, moderately, and weakly relevant outcomes. This genome-wide pass produced a ranked list over 19,193 genes and a mechanism-filtered 25-gene shortlist. Because no ground-truth causal-driver labels were available, the resulting posterior should be interpreted as an evidence score rather than as a calibrated probability of true causal relevance.

2.3.3. Family-aware Panel20 compression

To compress the candidate list from 25 genes to the final 20-gene BIRD-inspired panel, we used a constrained heuristic objective rather than a formula derived from a previously published probabilistic model. The objective was designed to retain genes with high BIRD-inspired CPT relevance scores while controlling mechanism-family coverage and within-family redundancy.

For a candidate subset S of 20 genes, we used

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

where $R(g)$ is the BIRD-inspired CPT score, $Q(g)$ is the family representativeness score, and $n_f(S)$ is the number of selected genes from family f. The constraint was that at least one gene from each of the six mechanism families had to be retained. The weights were used only to break ties and control redundancy during the compression from 25 candidates to 20 genes. They were not interpreted as probabilistic parameters.

The removed genes were *SRC*, *FGFR1*, *FGFR2*, *KRAS*, and *IGF1R*. The final BIRD-20 panel was:

EGFR, AXL, ERBB3, ERBB2, MAPK3, PIK3CA, GAS6, AKT1, HGF, MET, ERBB4, DUSP4, DUSP6, PDGFRB, PDGFRA, FN1, BCL2L11, NRAS, BRAF, PTEN.

This panel retained six bypass RTK genes, five MAPK escape genes, four direct EGFR/ERBB genes, three PI3K/AKT/PTEN genes, one EMT/AXL marker, and one apoptotic priming gene.

2.4. DAG construction and comparator panels

The DAG was not learned from the data. Instead, we specified a fixed prior graph for each panel to provide a biologically constrained ordering for CNF training and counterfactual generation. For BIRD-20, the graph was constructed as a compressed Erlotinib signaling graph using three principles. First, direct KEGG or literature-supported ligand–receptor and pathway edges among selected genes were retained. Second, when selected genes represented endpoints of a known pathway but intermediate adaptor nodes were not included in the compact panel, we added compressed pathway edges to preserve the expected upstream-to-downstream signaling direction. Third, feedback edges from downstream effectors to upstream receptors were excluded to maintain acyclicity. Thus, the graph should be interpreted as a prior biological ordering used by the CNF model, rather than as a data-learned causal discovery result.

The resulting BIRD-20 graph contained 20 gene nodes and 27 directed internal edges, with no isolated genes. Comparator graphs for GPT-20 and KEGG-26 were constructed using the same pathway-prior principle and evaluated with the identical downstream CNF-CCCE procedure. The resulting BIRD-20 DAG contains 20 gene nodes and 27 directed internal edges, has no isolated genes, and is acyclic. Its topological order is

GAS6, HGF, NRAS, PTEN, AXL, MET, EGFR, ERBB2, ERBB4, PDGFRA, PDGFRB, BRAF, FN1, ERBB3, MAPK3, PIK3CA, BCL2L11, DUSP4, DUSP6, AKT1.

Two comparators were built for evaluation. GPT-20 is a compact text-centric literature panel selected without BIRD-inspired CPT scoring or family-aware optimization. KEGG-26 is a pathway-union panel derived from Erlotinib-relevant KEGG signaling context. Because BIRD-20 and GPT-20 contain 20 genes whereas KEGG-26 contains 26 genes,

KEGG-26 is interpreted as a pathway-union comparator rather than a size-matched control. All panels were evaluated with the same downstream CNF-CCCE procedure.

2.5. CNF for continuous generative causal modeling

Let $X = (X_1, \ldots, X_m)$ denote the selected genes and let Y denote Erlotinib logIC50. We trained a CNF on the $(m+1)$-node system (X, Y) to model the joint observational distribution while respecting the causal ordering or graph structure. The purpose of the model is not only predictive accuracy but also counterfactual tractability: once trained, the model should support interventions and sample-specific counterfactual generation under continuous-valued gene expression variables (Javaloy et al., 2023).

Although the biology-adjusted DAG provides a plausible prior structure among genes, the transcriptomic measurements and IC50 values were obtained from observational cell-line assays. Therefore, we do not assume that the directed gene–outcome relations in the graph are identifiable from the data alone. We initially evaluated a graph-constrained CNF setting, in which the gene–gene graph was preserved and all selected genes were allowed to point to the outcome node. However, in practice, this graph-based model did not learn reliable average treatment effect (ATE) gaps for downstream counterfactual analysis (Appendix C).

For this reason, we used the *ordering + no regularization* CNF variant as the main model. The ordering was obtained from the topological order of the biology-adjusted gene-level DAG, and the Erlotinib logIC50 node was placed last. This choice preserves a biologically motivated variable ordering while avoiding strong graph-edge regularization that may overconstrain the model. We trained the model using five-fold cross validation. For each fold, all node values were normalized using the mean and standard deviation computed from the training split, and the same transformation was applied to the corresponding held-out split.

2.6. Individual-level counterfactual attribution with CCCE approximation

Our final goal is to assign a sample-specific counterfactual attribution score to each candidate gene or gene set for the observed Erlotinib response of an individual cell line. We build on CCCE (Zhao et al., 2023), but adapt it to continuous transcriptomic variables and continuous drug response. Let Y denote Erlotinib logIC50 and let X denote the vector of selected gene-expression variables. For notational consistency with the original CCCE formulation, we also define a binary high-IC50 event

$$Y^\star = \mathbb{I}(Y \geq \tau),$$

where τ is the training-fold mean Erlotinib logIC50 and $\mathbb{I}(\cdot)$ is the indicator function. Thus, $Y^\star = 1$ denotes that a cell line is in a high-IC50, relatively resistant state under the fold-specific training threshold.

For a candidate intervention set S, $X_S \subseteq X$ denotes the genes being intervened on. Because the response-reducing intervention direction is not known a priori, lowering expression is not assumed to be the only relevant perturbation. For each fold and each gene j, we therefore define two intervention anchors from the training split: a low-expression anchor x_j^{low} near the empirical $15\% \pm 5\%$ quantile and a high-expression anchor x_j^{high} near the empirical $85\% \pm 5\%$ quantile. For a gene set S, these anchors form the vectors x_S^{low} and x_S^{high}, and the candidate anchor set is

$$\mathcal{A}_S = \{x_S^{\text{low}}, x_S^{\text{high}}\}.$$

In the original binary CCCE setting, the contribution of a cause set can be written as the change in the probability of the observed positive outcome under a factual versus counterfactual cause state. For an observed high-IC50 sample with $X = x$ and $Y^\star = 1$, this motivates a removal-style quantity of the form

$$\begin{aligned}&\text{CCCE}(X_S \Rightarrow Y^\star \mid X = x, Y^\star = 1)\\ &\approx 1 - \frac{P(Y^\star_{do(X_S=a)} = 1 \mid X = x)}{P(Y^\star = 1 \mid X = x)},\end{aligned}$$

where $a \in \mathcal{A}_S$ denotes a low- or high-expression anchor vector for X_S. In discrete CCCE, a is usually the opposite or removed cause state. In our continuous setting, there is no unique "removal" value, so we evaluate both biologically interpretable low and high anchors in $\mathcal{A}_S$.

Directly estimating the observational denominator and the counterfactual numerator by enumerating configurations is computationally expensive and unstable here, because both X and Y are continuous and $X = x$ is a high-dimensional conditioning event. We therefore approximate both terms using the trained CNF. The denominator is approximated as

$$\begin{aligned}\widehat{p}_{\text{obs}}(x) &= \widehat{P}_{\text{CNF}}(Y^\star = 1 \mid X = x)\\ &= \int_\tau^\infty \widehat{p}_{\text{CNF}}(y \mid x)\, dy,\end{aligned}$$

where $\widehat{p}_{\text{CNF}}(y \mid x)$ is the conditional density of Erlotinib logIC50 estimated by the CNF.

For the counterfactual numerator, we evaluate both intervention anchors:

$$\widehat{p}_{\text{cf}}(S, x, a) = \widehat{P}_{\text{CNF}}\left(Y^\star_{do(X_S=a)} = 1 \mid X = x\right), \quad a \in \mathcal{A}_S.$$

The response-reducing anchor is selected using the continuous predicted response,

$$a_S^\star(x) = \arg\min_{a \in \mathcal{A}_S} \widehat{\mathbb{E}}_{\text{CNF}}\left[Y_{do(X_S=a)} \mid X = x\right],$$

and the binary high-IC50 numerator below is evaluated at this selected anchor. The resulting CNF-approximated CCCE score is defined as

$$\widehat{\mathrm{CCCE}}_{\mathrm{CNF}}(S;x) = 1 - \frac{\widehat{p}_{\mathrm{cf}}(S, x, a_S^\star(x))}{\max\{\widehat{p}_{\mathrm{obs}}(x), \epsilon\}},$$

where $\epsilon > 0$ is a small constant used to avoid numerical instability. A larger score indicates that the selected intervention anchor gives a lower counterfactual probability of remaining in the high-IC50 state, normalized by the sample's observational high-IC50 probability under the fitted CNF.

For each held-out high-IC50 cell line and each candidate intervention set S, we generated counterfactual outcomes under both $do(X_S = x_S^{\mathrm{low}})$ and $do(X_S = x_S^{\mathrm{high}})$ and retained the anchor that produced the larger reduction in predicted Erlotinib logIC50. Candidate gene sets were ranked by $\widehat{\mathrm{CCCE}}_{\mathrm{CNF}}(S;x)$, so the top-ranked set is the intervention set with the strongest model-predicted movement away from the resistant state for that sample. Separately, a hit was counted when the selected low- or high-anchor intervention produced at least a pre-specified 30% relative reduction in model-predicted Erlotinib logIC50 under the same fold-specific transformation used for evaluation. This 30% criterion is an operational model-comparison threshold, not a calibrated clinical response threshold. The procedure yields a sample-specific ranking of genes or gene sets whose counterfactual perturbation most strongly reduces the predicted Erlotinib resistance phenotype.

3. Results

3.1. Construction of KEGG-26 and GPT-20 comparator panels

To evaluate whether BIRD-20 provided advantages beyond known pathway membership or a compact text-derived panel, we compared it with KEGG-26 and GPT-20. Both comparator panels were converted into DAGs and evaluated with the same exact CCCE and CNF-approximated CCCE procedures as BIRD-20. For approximate CCCE, all panels used the ordering plus no-regularization CNF configuration, five-fold training, training-split intervention anchors, and held-out high-response test samples. Details of the comparator construction are provided in Appendix A and Appendix B. Because KEGG-26 has more genes than BIRD-20 and GPT-20, it should be read as a canonical pathway-union comparator rather than a size-matched baseline.

3.2. Exact CCCE hit rates

The exact CCCE of BIRD-20, GPT-20, and KEGG-26 was computed after full-data mean binarization for 285 Erlotinib-high samples. Table 2 reports exact hit rates together with maximum-score tie diagnostics.

Table 2. Diagnostic exact CCCE hit rates and maximum-score tie statistics after full-data mean binarization.

CR	Panel	Candi-dates	Hit rate	Top-tie	
				Fraction	Med. Count
CR1	BIRD-20	20	0.85	0.99	19
	GPT-20	20	0.60	1.00	19
	KEGG-26	26	0.61	1.00	25
CR2	BIRD-20	190	0.85	0.98	189
	GPT-20	190	0.80	0.95	189
	KEGG-26	325	0.77	0.99	324
CR3	BIRD-20	1140	0.94	0.91	1139
	GPT-20	1140	0.87	0.88	1139
	KEGG-26	2600	0.89	0.96	2599

Here, top-tie fraction is the fraction of samples for which more than one candidate gene set attains the maximum exact CCCE score, and median top-tie count is the median number of maximum-scoring candidates among those samples. Exact CCCE produced high hit rates but also very large maximum-score ties, often involving nearly all candidates in a CR order. Thus, exact CCCE was treated as a diagnostic rather than the primary ranking metric: when many candidates share the same maximum score, unique sample-specific top-profile selection becomes sensitive to candidate ordering or tie-breaking.

3.3. Approximate CNF CCCE hit rates

Because CNF-based approximate CCCE requires a trained CNF model, we evaluated the downstream attribution step using a five-fold cross-validation scheme conditional on the fixed full-cohort-selected panels. For each fold, the CNF was trained on the corresponding training split, and approximate CCCE was computed only for the held-out test samples. For each panel and CR order, the fold-level hit rate was computed as the number of hit held-out high-IC50 samples divided by the number of held-out high-IC50 samples in that fold. The fold-level results were then combined by reporting the mean and standard deviation across the five folds.

The number of held-out test samples in each fold, together with the number of Erlotinib-high samples used as denominators for CCCE evaluation, is reported in Table 3. Erlotinib-high samples were defined within each fold using the mean Erlotinib logIC50 value of the corresponding training split. The mean and standard deviation of the test-fold hit rates for CR1, CR2, and CR3 are reported in Table 4.

In contrast to exact CCCE, CNF-based approximate CCCE did not produce maximum-score ties. For all panels and causal orders, each evaluated sample had a single top-ranked

Table 3. Held-out test samples used for CNF-CCCE.

	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Test samples	115	114	114	114	114
High Erlotinib test samples	61	62	58	54	51

Table 4. Five-fold test-only CNF-CCCE hit rates conditional on fixed panels.

Panel	CR1	CR2	CR3
BIRD-20	0.10 ± 0.06	0.51 ± 0.23	0.76 ± 0.14
GPT-20	0.09 ± 0.05	0.47 ± 0.12	0.74 ± 0.05
KEGG-26	0.004 ± 0.008	0.12 ± 0.04	0.40 ± 0.07

candidate gene set, corresponding to a top-tie count of one. The separation between the top-ranked and second-ranked candidates, measured by the top1–top2 score margin, is summarized in Table 5. BIRD-20 had larger mean hit rates than KEGG-26 for CR2 and CR3 and only modestly higher rates than GPT-20; therefore, the BIRD-20 versus GPT-20 comparison should be interpreted as a small effect under this evaluation rather than a definitive separation.

3.4. Tissue-origin structure of hit and non-hit sample-specific counterfactual profiles

To evaluate whether sample-specific counterfactual profiles reflected tissue-specific structure, we stratified the held-out test samples according to the CCCE hit indicator. For each panel and causal order, we first pooled the five held-out test folds and then separated samples with hit from samples with no-hit. Within each stratum, the top-ranked CCCE gene set, including its intervention direction, was counted across cell lines and stacked by refined tissue-origin group. The origin mapping is provided in Table 8.

Among hit samples (Fig. 2), BIRD-20's most frequent CR1 profiles were AXL(0) and MET(0), whereas CR2 and CR3 were dominated by AXL/MET-centered combinations. GPT-20 showed a stronger FGFR1-centered pattern, especially in CR1 and CR2. KEGG-26 produced very few CR1 hits and therefore provided limited information at CR1, but CR2 and CR3 hits were mostly concentrated in EGFR/HRAS/PI3K-related sets.

To quantify the association between tissue origin and top counterfactual gene-set labels, we computed normalized mutual information (NMI) between the refined origin group and the top gene-set label. Statistical significance was assessed by 2,000 random permutations of the top gene-set labels within each panel and causal order, followed by Benjamini–Hochberg (BH) false-discovery-rate correction across valid hit-only tests. For consistency with the profile ranking, top gene-set labels outside the 12 most frequent labels were collapsed into an "Other top sets" category before the primary association test.

Table 5. Median top1–top2 CCCE score margin.

Panel	CR1	CR2	CR3
BIRD-20	0.033	0.016	0.014
GPT-20	0.038	0.020	0.013
KEGG-26	0.024	0.013	0.008

Table 6. Tissue-origin association among hit samples. BH q values were computed across valid hit-only tests.

Panel	CR	n	Origin groups	Top labels	NMI	p_{perm}	q_{BH}
BIRD-20	CR1	29	10	8	0.4714	0.2119	0.3390
BIRD-20	CR2	140	14	13	0.2723	0.0010	0.0080
BIRD-20	CR3	215	15	13	0.2012	0.0060	0.0240
GPT-20	CR1	26	11	5	0.3755	0.4248	0.4783
GPT-20	CR2	138	15	13	0.2679	0.0245	0.0653
GPT-20	CR3	212	15	13	0.1764	0.2839	0.3785
KEGG-26	CR1	1	1	1	–	–	–
KEGG-26	CR2	37	11	13	0.4845	0.4783	0.4783
KEGG-26	CR3	113	15	13	0.2662	0.1749	0.3390

The hit-only association results are summarized in Table 6. BIRD-20 showed tissue-origin association for CR2 and CR3 that remained significant after BH correction (CR2: NMI=0.2723, permutation $p = 0.0010$, $q = 0.0080$; CR3: NMI=0.2012, permutation $p = 0.0060$, $q = 0.0240$), indicating that the successful sample-specific profiles were not randomly distributed across tissue origins. GPT-20 showed a weaker nominal CR2 association (NMI=0.2679, permutation $p = 0.0245$, $q = 0.0653$), whereas GPT-20 CR1 and CR3 were not significant. KEGG-26 had only one CR1 hit, preventing a meaningful CR1 association test, and its CR2 and CR3 hit strata were not significant despite moderate raw NMI values. Because CR2 and CR3 generate many possible labels, the origin-by-profile contingency tables are sparse. Therefore, these permutation-based NMI results are interpreted as global evidence that tissue origin and top-profile labels are not independent, not as evidence for stable one-to-one mappings between individual origins and specific gene pairs or triplets. The corresponding non-hit origin-stratified profile counts and NMI results are provided in Appendix E. In the non-hit stratum, the NMI associations were not significant after BH correction, indicating that the tissue-origin structure was weaker outside the successful hit profiles.

3.5. Biological patterns in sample-specific profiles

Across held-out high-Erlotinib samples, BIRD-20 top profiles were dominated by biologically interpretable bypass-RTK and ERBB-family patterns. Frequent BIRD-20 CR1 candidates included *AXL*, *GAS6*, *ERBB3*, *MET*, and *EGFR*; frequent higher-order profiles extended these sig-

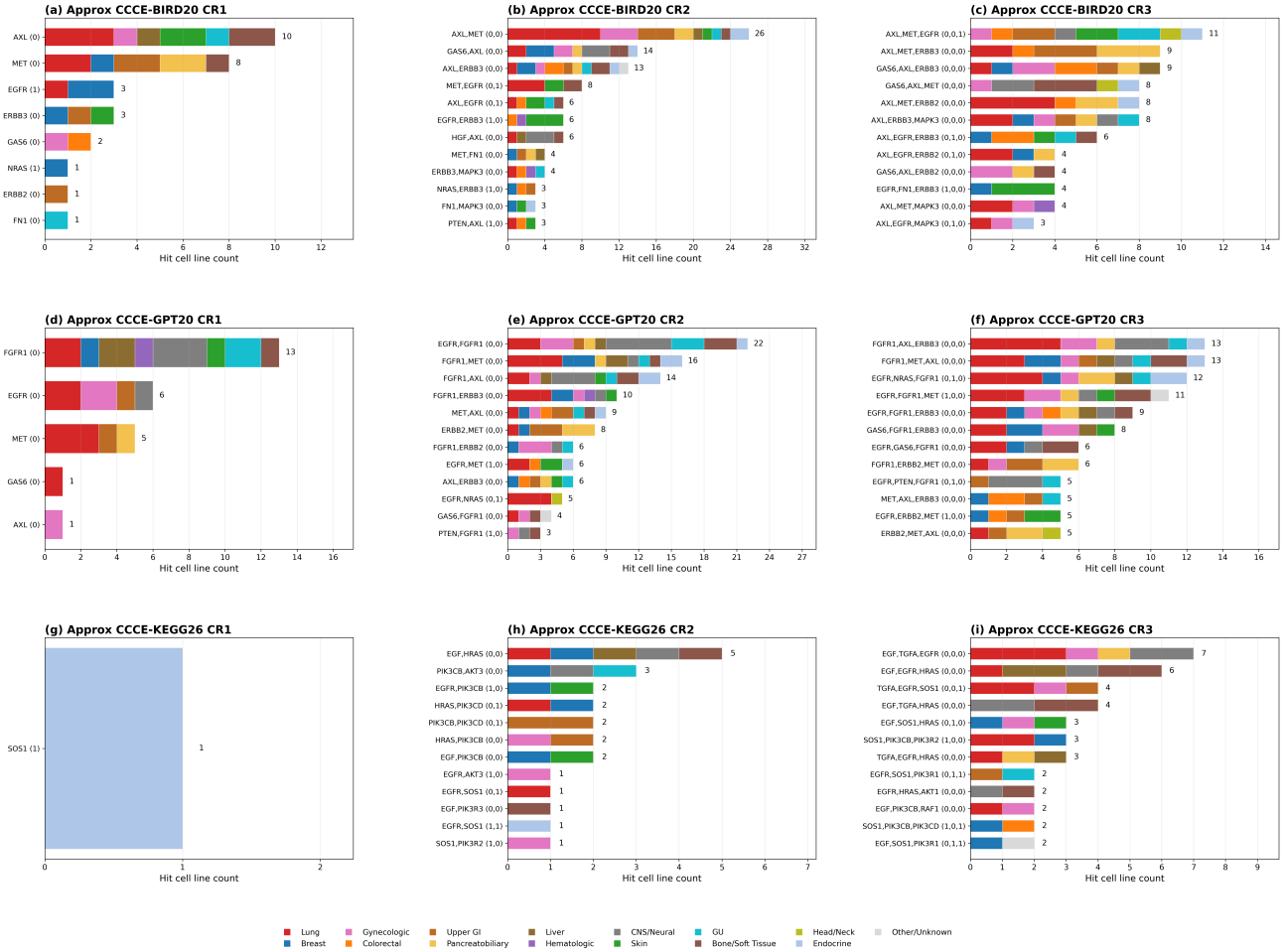

Figure 2. Tissue-origin composition of top-ranked approximate CCCE gene sets among hit held-out test samples. Samples were pooled across five test folds.

nals into *GAS6–AXL*, *AXL–ERBB3*, *AXL–MET*, and related triplets. When intervention direction was retained in the top-profile labels, *EGFR*(1), corresponding to the high-expression EGFR anchor, also appeared as a recurrent top-ranked BIRD-20 signal across CR1, CR2, and CR3. GPT-20 produced a more FGFR1- and SRC-adjacent pattern, whereas KEGG-26 was concentrated around canonical ligand/receptor and RAS/PI3K components such as *HRAS*, *EGFR*, *SOS1*, *PIK3CB*, *EGF*, and *TGFA*. These differences are consistent with the construction principles of the panels: BIRD-20 prioritizes mechanism-balanced Erlotinib biology, GPT-20 reflects a compact literature-style list, and KEGG-26 preserves dense canonical pathway membership.

The tissue-origin association analysis further showed that top counterfactual profiles were not determined solely by tissue origin. Instead, BIRD-20 hit profiles showed significant but moderate tissue-origin structure for CR2 and CR3, while KEGG-26 showed stronger tissue association mainly among non-hit profiles. This supports the interpretation that CNF-CCCE reflects both shared tissue-context effects and additional cell-line-specific heterogeneity in Erlotinib response.

4. Discussion

This study combines text-centric biological knowledge with data-driven counterfactual modeling for sample-specific Erlotinib response analysis. Because gene-level driver labels are unavailable, we used a BIRD-inspired evidence-integration step to define a compact mechanism space from literature, expression, mutation, and KEGG topology before applying CNF-CCCE attribution. This hybrid design aims to avoid two failure modes: text-only panels that are biologically plausible but insensitive to the cohort, and purely data-driven panels that may select redundant pathway hubs.

BIRD-20 should not be interpreted as the only biologically plausible panel. Several GPT-20 axes, including SRC and FGFR1, have independent Erlotinib relevance

through combination-treatment or bypass-resistance literature (Haura et al., 2010; Quintanal-Villalonga et al., 2019; Raoof et al., 2019). The distinction is that BIRD-inspired scoring and family-aware compression retained a more balanced set of Erlotinib mechanisms, including *HGF–MET*, *AXL–GAS6*, ERBB-family, MAPK, PI3K/AKT/PTEN, and PDGFR-related components, several of which overlap with clinically studied Erlotinib combination themes (Seto et al., 2014; Kawashima et al., 2022; Nakagawa et al., 2019; U.S. Food and Drug Administration, 2020; Sequist et al., 2011; Scagliotti et al., 2020; Spigel et al., 2011).

The recurrent high-*EGFR* BIRD-20 signal is biologically interpretable but should be read carefully. Among CR1 interventions, BIRD-20 was the only panel in which high-*EGFR* produced held-out high-to-low Erlotinib hits; in CR2 and CR3, *EGFR*-containing BIRD-20 hits also consistently used the high-*EGFR* anchor. This does not imply that experimentally increasing EGFR expression is therapeutic. Rather, under the learned CNF distribution, the high-*EGFR* anchor appears to mark an EGFR-dependency-like state associated with lower predicted Erlotinib logIC50, consistent with reports that EGFR pathway dependence and EGFR-related expression signatures can predict EGFR-TKI sensitivity (Balko et al., 2006; Cheng et al., 2020).

Finally, exact CCCE and CNF-approximated CCCE played different roles. Exact CCCE was useful as a diagnostic but produced extensive maximum-score ties, making top-candidate selection sensitive to ordering. CNF-approximated CCCE produced unique top-ranked profiles and nonzero top1–top2 margins, making the ranking more usable for sample-specific hypothesis generation. The analysis remains conditional on the selected panel, biological ordering, intervention anchors, and CNF configuration; because the data are observational, all reported profiles should be treated as model-based hypotheses requiring experimental validation.

5. Conclusion

We propose a mechanism-guided framework that converts high-dimensional Erlotinib transcriptomic data into compact, interpretable, sample-specific counterfactual profiles. BIRD-20 combines literature-defined mechanism families with expression, mutation, and KEGG-derived evidence through a BIRD-inspired scoring procedure, while CNF-approximated CCCE extends individual-level attribution to continuous drug-response data and multi-gene interventions. Conditional on fixed full-cohort-selected panels, held-out CNF-CCCE evaluation showed higher hit rates than the pathway-union KEGG-26 comparator and similar or modestly higher hit rates than the text-centric GPT-20 comparator, especially for CR2 and CR3. The results support the value of combining text-centric biological reasoning with data-driven evidence integration and generative counterfactual modeling for drug-response hypothesis generation, while leaving end-to-end panel validation and experimental causal testing for future work.

References

Balko, J. M., Potti, A., Saunders, C., Stromberg, A., Haura, E. B., and Black, E. P. Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics*, 7:289, 2006. doi: 10.1186/1471-2164-7-289. URL https://doi.org/10.1186/1471-2164-7-289.

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.

Cheng, C., Zhao, Y., Schaafsma, E., Weng, Y.-L., and Amos, C. I. An EGFR signature predicts cell line and patient sensitivity to multiple tyrosine kinase inhibitors. *International Journal of Cancer*, 147(9):2621–2633, 2020. doi: 10.1002/ijc.33053. URL https://doi.org/10.1002/ijc.33053.

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.

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.

Haura, E. B., Tanvetyanon, T., Chiappori, A., Williams, C., Simon, G., Antonia, S., Gray, J., Litschauer, S., Tetteh, L., Neuger, A., et al. Phase I/II study of the Src inhibitor dasatinib in combination with erlotinib in advanced non-small-cell lung cancer. *Journal of Clinical Oncology*, 28(8):1387–1394, 2010. doi: 10.1200/JCO.2009.25.4029. URL https://doi.org/10.1200/JCO.2009.25.4029.

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 36*, pp. 1–32, 2023. URL https://openreview.net/forum?id=QIFoCI7ca1.

Kawashima, Y., Fukuhara, T., Saito, H., Furuya, N., Watanabe, K., Sugawara, S., Iwasawa, S., Tsunezuka, Y., Yamaguchi, O., Okada, M., et al. Bevacizumab plus erlotinib versus erlotinib alone in japanese patients with advanced, metastatic, EGFR-mutant non-small-cell lung cancer (NEJ026): overall survival analysis of an open-label, randomised, multicentre, phase 3 trial. *The Lancet Respiratory Medicine*, 10(1):72–82, 2022. doi: 10.1016/S2213-2600(21)00166-1. URL https://doi.org/10.1016/S2213-2600(21)00166-1.

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.

Nakagawa, K., Garon, E. B., Seto, T., Nishio, M., Ponce Aix, S., Paz-Ares, L., Chiu, C.-H., Park, K., Novello, S., Nadal, E., et al. Ramucirumab plus erlotinib in patients with untreated, EGFR-mutated, advanced non-small-cell lung cancer (RELAY): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet Oncology*, 20(12):1655–1669, 2019. doi: 10.1016/S1470-2045(19)30634-5. URL https://doi.org/10.1016/S1470-2045(19)30634-5.

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 perturbations 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.

Scagliotti, G., Moro-Sibilot, D., Kollmeier, J., Favaretto, A., Cho, E. K., Grosch, H., Kimmich, M., Girard, N., Tsai, C.-M., Hsia, T.-C., et al. A randomized-controlled phase 2 study of the MET antibody emibetuzumab in combination with erlotinib as first-line treatment for EGFR mutation-positive NSCLC patients. *Journal of Thoracic Oncology*, 15(1):80–90, 2020. doi: 10.1016/j.jtho.2019.10.003. URL https://doi.org/10.1016/j.jtho.2019.10.003.

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.

Sequist, L. V., von Pawel, J., Garmey, E. G., Akerley, W. L., Brugger, W., Ferrari, D., Chen, Y., Costa, D. B., Gerber, D. E., Orlov, S., et al. Randomized phase II study of erlotinib plus tivantinib versus erlotinib plus placebo in previously treated non-small-cell lung cancer. *Journal of Clinical Oncology*, 29(24):3307–3315, 2011. doi: 10.1200/JCO.2010.34.0570. URL https://doi.org/10.1200/JCO.2010.34.0570.

Seto, T., Kato, T., Nishio, M., Goto, K., Atagi, S., Hosomi, Y., Yamamoto, N., Hida, T., Maemondo, M., Nakagawa, K., et al. Erlotinib alone or with bevacizumab as first-line therapy in patients with advanced non-squamous non-small-cell lung cancer harbouring EGFR mutations (JO25567): an open-label, randomised, multicentre, phase 2 study. *The Lancet Oncology*, 15(11):1236–1244, 2014. doi: 10.1016/S1470-2045(14)70381-X. URL https://doi.org/10.1016/S1470-2045(14)70381-X.

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.

Spigel, D. R., Burris, H. A., Greco, F. A., Shipley, D. L., Friedman, E. K., Waterhouse, D. M., Whorf, R. C., Mitchell, R. B., Daniel, D. B., Zangmeister, J., et al. Randomized, double-blind, placebo-controlled, phase II trial of sorafenib and erlotinib or erlotinib alone in previously treated advanced non-small-cell lung cancer. *Journal of Clinical Oncology*, 29(18):2582–2589, 2011. doi: 10.1200/JCO.2010.30.7678. URL https://doi.org/10.1200/JCO.2010.30.7678.

U.S. Food and Drug Administration. FDA approves ramucirumab plus erlotinib for first-line metastatic NSCLC. FDA approval notice, 2020. Accessed 2026-05-08.

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.

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 Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 2519–2528, 2023.

A. Construction of the KEGG-derived 26-gene comparator

To evaluate whether the proposed BIRD-inspired 20-gene panel provides a more mechanism-balanced representation of Erlotinib response than a direct pathway-union baseline, we constructed a comparator panel and its pathway-union DAG. They were constructed by merging three KEGG pathways: EGFR tyrosine kinase inhibitor resistance (`hsa01521`), non-small cell lung cancer (`hsa05223`), and colorectal cancer (`hsa05210`). We refer to this comparator as KEGG-26. This panel was designed to represent a broader pathway-derived baseline, rather than a family-optimized compact panel.

KEGG-derived gene set. The final KEGG-26 panel contained the following 26 genes:

EGF, *TGFA*, *EGFR*, *GRB2*, *SOS1*, *SOS2*, *HRAS*, *KRAS*, *NRAS*, *PIK3CA*, *PIK3CB*, *PIK3CD*, *PIK3R1*, *PIK3R2*, *PIK3R3*, *ARAF*, *BRAF*, *RAF1*, *AKT1*, *AKT2*, *AKT3*, *MAP2K1*, *MAP2K2*, *BAD*, *MAPK1*, *MAPK3*.

Here, *PIK3CA* denotes phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha.

Excluded features. Two genes or features were removed from the initial pathway-derived candidate set before constructing KEGG-26. First, *BAX* was excluded because it became isolated after restricting the pathway union to the common retained genes. In the original KEGG pathway edges, *BAX* was connected to genes such as *BCL2L11*, *BCL2*, *BCL2L1*, and *TP53*. However, these neighboring genes were not retained in the common pathway-union gene set. As a result, *BAX* had no remaining edges in the reduced union graph and did not contribute to the comparator DAG.

Second, the feature *P3R3URF–PIK3R3* was excluded because it was highly sparse in the intermediate expression-feature table in which this readthrough feature was available: 343 of 363 evaluable records were zero-valued and only 20 had nonzero values. This intermediate sparsity check is distinct from the final 571-cell-line matched Erlotinib analysis cohort. In addition, *PIK3R3* itself was already included as a standard gene feature. We therefore treated *P3R3URF–PIK3R3* as a sparse readthrough or merged feature and excluded it from the KEGG-26 panel.

KEGG-26 DAG. The resulting KEGG-26 graph contained 26 gene nodes and one Erlotinib response node. The gene–gene backbone contained 63 pathway-union edges derived from the integrated KEGG pathways. As in the relaxed outcome setting used for the BIRD-inspired graph, all 26

Table 7. Mechanism-family counts for BIRD-20 and KEGG-26.

Panel	BIM	RTK	EGFR	EMT/ AXL	MAPK	PI3K/AKT/ PTEN
BIRD-20	1	6	4	1	5	3
KEGG-26	1	0	3	0	13	9

genes were connected to the Erlotinib outcome node. This produced 26 relaxed outcome edges. Thus, the KEGG-26 comparator graph contained 27 total nodes, 63 pathway-union edges, and 26 outcome-parent edges.

This comparator was not optimized for mechanism-family balance or redundancy reduction. Instead, it was intended to represent a direct pathway-union baseline that preserves a broader set of canonical signaling components from KEGG.

B. Comparison of BIRD-20 and KEGG-26 mechanism coverage

We compared BIRD-20, the proposed BIRD-inspired panel, with the KEGG-derived 26-gene comparator to clarify how the two candidate spaces differ biologically. This comparison is not meant to show that KEGG-derived genes are irrelevant. Rather, it shows that direct pathway-union construction and family-aware panel construction emphasize different parts of Erlotinib biology. In the BIRD-inspired evidence table, only seven KEGG-26 genes were present in the mechanism-enriched Stage-2 shortlist: *EGFR*, *KRAS*, *NRAS*, *PIK3CA*, *BRAF*, *AKT1*, and *MAPK3*. The remaining 19 KEGG-26 genes were absent from that shortlist:

EGF, *TGFA*, *GRB2*, *SOS1*, *SOS2*, *HRAS*, *PIK3R3*, *PIK3CB*, *PIK3CD*, *PIK3R1*, *PIK3R2*, *ARAF*, *RAF1*, *AKT2*, *AKT3*, *MAP2K1*, *MAP2K2*, *BAD*, *MAPK1*.

The family-count comparison shows that KEGG-26 was strongly concentrated in mitogen-activated protein kinase (MAPK) and phosphoinositide 3-kinase/protein kinase B/phosphatase and tensin homolog (PI3K/AKT/PTEN) signaling components. In contrast, BIRD-20 covered all six predefined Erlotinib-related mechanism families, including bypass receptor tyrosine kinase activation and epithelial–mesenchymal transition/AXL-associated resistance, which were not represented in KEGG-26 under the family mapping used in this study.

This comparison highlights the difference between the two panel-construction strategies. KEGG-26 preserves a broader canonical pathway-union structure, but it is concentrated in a few dense signaling families and does not explicitly optimize mechanism diversity. BIRD-20, by contrast, was selected using BIRD-inspired evidence integration to retain mechanism coverage while limiting within-family redundancy. Therefore, KEGG-26 serves as a useful pathway-derived baseline, whereas BIRD-20 better reflects the objective of constructing a compact and mechanism-balanced panel for causal normalizing flow modeling.

Table 8. Mapping from detailed cell-line origin labels to broad origin groups used in the tissue-origin association analysis.

Broad origin group	Detailed origin
Lung	Lung
Breast	Breast
Gynecologic	Ovary, Uterus, Cervix
Colorectal	Colorectal
Upper GI	Gastric, Esophagus
Pancreatobiliary	Pancreas, Bile Duct
Liver	Liver
Head/Neck	Upper Aerodigestive
Hematologic	Blood, Lymphocyte, Plasma Cell
CNS/Neural	Central Nervous System, Peripheral Nervous System
Skin	Skin
GU	Kidney, Urinary Tract, Prostate
Bone/Soft Tissue	Bone, Soft Tissue
Endocrine	Thyroid
Other/Unknown	Embryo, Unknown

C. CNF training

We adopt a neural spline flow (NSF)-based architecture. The model uses hidden widths of $[32, 32, 32]$ and 4 flow layers. We consider four training variants that differ by whether ordering constraints or graph constraints are used, and whether structural regularization is applied: `ordering + no regularization`, `ordering + regularization`, `graph + no regularization`, and `graph + regularization`. We use 5-fold training, 400 epochs per fold, checkpointing at each epoch, and no early stopping.

D. Origin group

E. Non-hit tissue-origin structure

We repeated the tissue-origin association analysis after restricting samples to `any_hit`=0. This analysis summarizes residual structure among top-ranked counterfactual profiles that did not satisfy the Erlotinib hit criterion.

Table 9. Permutation-based association between refined tissue origin and top CCCE gene-set labels among non-hit held-out test samples. The same NMI permutation procedure as in Table 6 was applied after restricting samples to `any_hit`=0; BH q values were computed across the nine non-hit tests.

Panel	CR	n	Origin groups	Top labels	NMI	p_{perm}	q_{BH}
BIRD-20	CR1	257	15	13	0.1899	0.0005	0.0023
BIRD-20	CR2	146	14	13	0.2747	0.0010	0.0023
BIRD-20	CR3	71	14	13	0.3623	0.2764	0.2764
GPT-20	CR1	260	15	13	0.1945	0.0005	0.0023
GPT-20	CR2	148	14	13	0.2452	0.0455	0.0585
GPT-20	CR3	74	13	13	0.3515	0.1124	0.1265
KEGG-26	CR1	285	15	13	0.1735	0.0010	0.0023
KEGG-26	CR2	249	15	13	0.1793	0.0060	0.0108
KEGG-26	CR3	173	14	13	0.2133	0.0360	0.0540

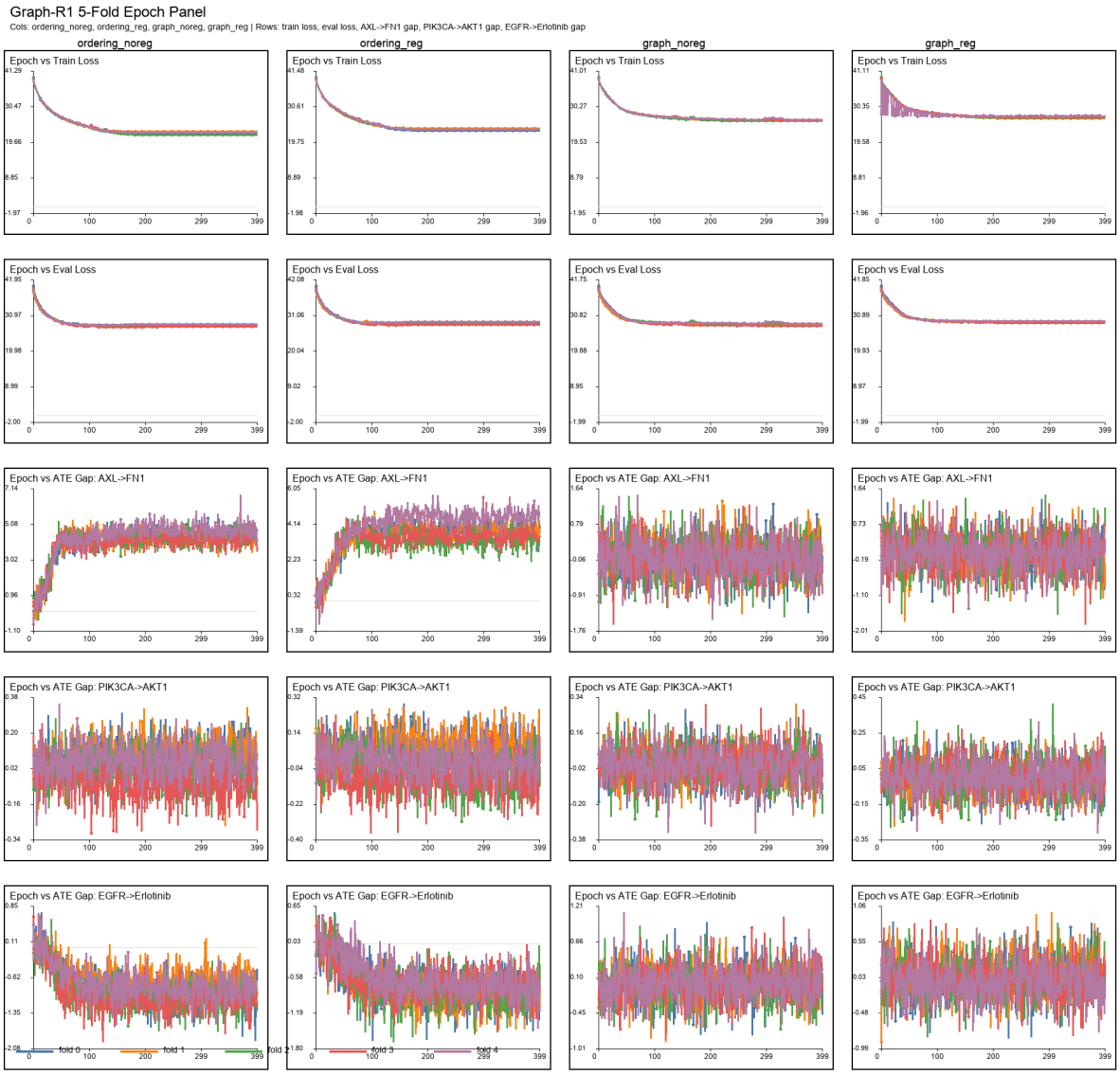

Figure 3. Five-fold epoch-level training dynamics and causal-direction diagnostics for the BIRD-20 CNF model. Epoch-wise curves are shown across five cross-validation folds and four model variants: ordering without regularization, ordering with regularization, graph without regularization, and graph with regularization. Rows summarize training loss, validation/evaluation loss, and ATE-gap diagnostics for representative directed relationships in the Erlotinib DAG: AXL $\rightarrow$ FN1, PIK3CA $\rightarrow$ AKT1, and EGFR $\rightarrow$ Erlotinib logIC50. The ATE gap is defined as the forward intervention effect minus the reverse intervention effect, so larger positive values indicate stronger agreement with the specified causal direction. Together, these panels assess both predictive training behavior and whether the learned flow preserves biologically expected directional effects across epochs and folds.

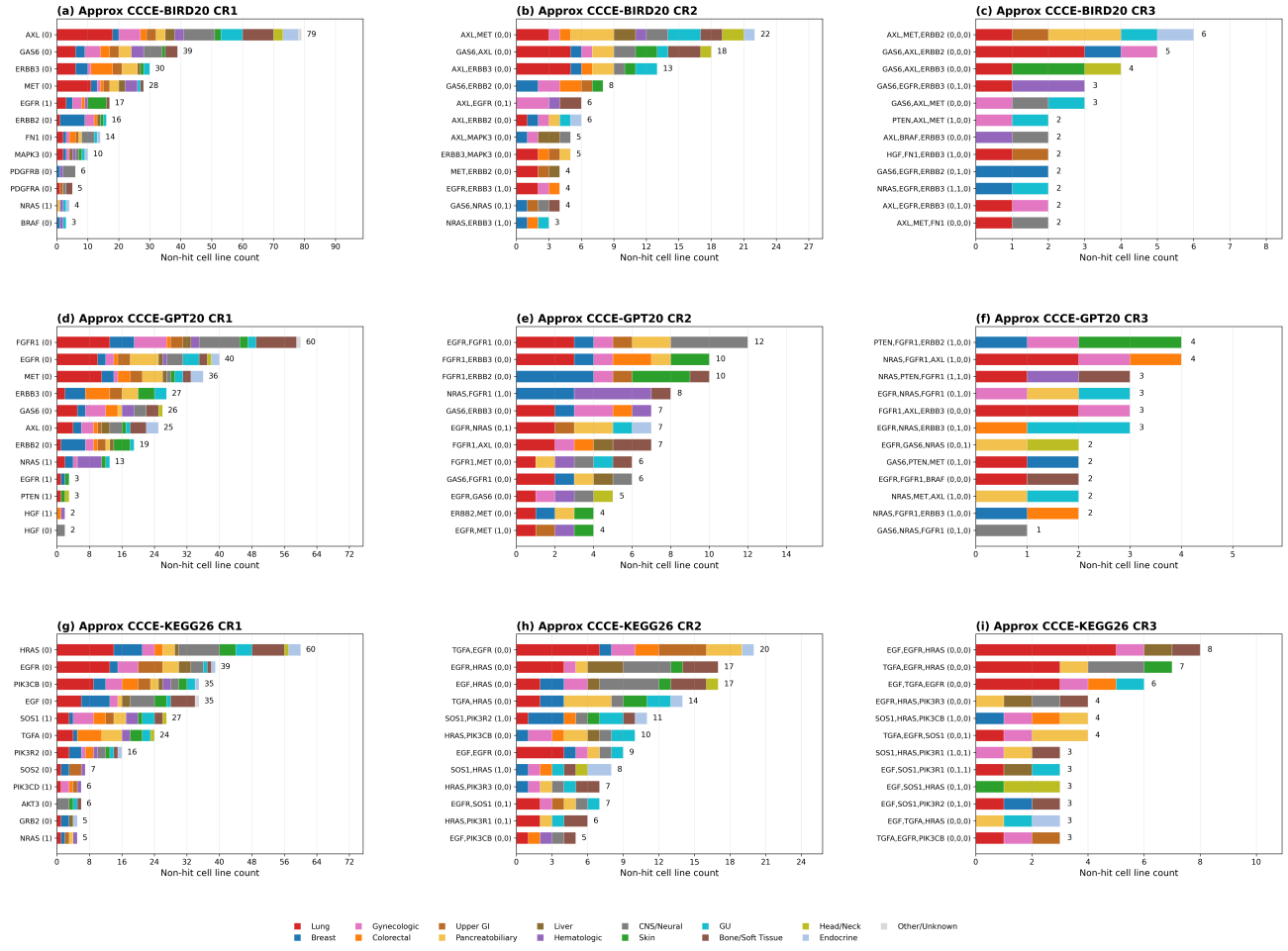

Figure 4. Tissue-origin composition of top-ranked approximate CCCE gene sets among non-hit held-out test samples. Samples were pooled across the five held-out test folds and restricted to `any_hit=0`. Bars show the number of non-hit cell lines assigned to each top counterfactual gene set, stacked by refined tissue-origin group.