DISCOVER, FALSIFY, REVISE: AUDITING INPUT-USE CLAIMS FROM SOURCE CODE TO PREDICTIVE CONTRIBUTION IN AGENT-DISCOVERED CELL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI virtual cells aim to predict cellular responses to specified interventions, yet held-out predictive performance alone does not establish use of the supplied perturbation information. This prediction–claim gap is especially consequential in agentic model discovery, where language-model agents generate and revise predictors using predominantly score-based feedback. We introduce CELLAUDIT, which audits registered input-use claims through three distinct questions: whether the input can enter the cited computation, whether fitted predictions depend on it, and whether that dependence improves prediction of observed responses. On a paired morphology–transcriptomics perturbation benchmark (BBBC047), an agent-selected predictor achieves a mean held-out Global Pearson correlation coefficient (PCC) of 0.3153 yet remains exactly invariant to compound replacement; a control-profile-only predictor reaches 0.3142. Source inspection identifies a compound-query pathway blocked by singleton key–value attention; the invariance persists after refitting with physically disjoint control wells for inputs and target references. In a stratified audit of 48 generated candidates across two linked tasks, 47 change predictions under compound replacement on both held-out folds, but only 20 show target-loss gains with intervals above zero on both folds. On BBBC047, falsification-guided revisions recover positive mean compound contributions while retaining predictive gains over the control-profile-only baseline. In matched sci-Plex searches, audit-enriched feedback shows higher mean held-out predictive performance and larger mean compound and dose contributions across five paired trajectories, although paired intervals span zero. Refitting fixed designs on an independently acquired cohort further shows predictive generalization need not imply generalization of input-use claims: dose contribution persists, whereas support for compound identity does not persist. CELLAUDIT therefore adds a falsification layer to agentic model discovery, moving from generate–score–revise toward discover–falsify–revise.

1 INTRODUCTION

AI virtual cells aim to predict cellular responses to chemical and genetic interventions and ultimately support intervention design (Bunne et al., 2024). Yet a model can predict well by exploiting cellular context, control profiles, or other systematic structure while making little or no use of the supplied perturbation information (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026). Predictive performance alone therefore does not establish perturbation use (D'Amour et al., 2022; Geirhos et al., 2020; Lapuschkin et al., 2019; DeGrave et al., 2021). We call a mismatch between a registered claim about input use and the model's implementation or fitted behavior the *prediction–claim gap*.

This gap poses an additional challenge for agentic model discovery (Huang et al., 2024; Li et al., 2024; Jiang et al., 2025; Lu et al., 2026). Systems such as CellScientist (Li et al., 2026), CellForge (Tang et al., 2025), and HarmonyCell (Huang et al., 2026) generate executable predictors, train them, observe validation feedback, and iteratively revise or select candidates. When feedback primarily rewards predictive performance, search can favor high-scoring predictors without testing whether they use the inputs specified by their designs. As generated models become more numerous and diverse, manual verification becomes increasingly difficult. Agentic discovery therefore needs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scalable tests of the claims attached to its selected models, not only mechanisms for proposing better predictors.

We introduce CELLAUDIT, a framework that links registered input-use claims to evidence from source code and fitted models. It asks three non-equivalent questions: can the input enter the cited computation, do fitted predictions depend on it, and does that dependence improve prediction of the observed response? These correspond to *source consumption*, *fitted dependence*, and *target-relevant predictive contribution*. A cited pathway can be mathematically inactive; an executable pathway may have little effect after fitting; and input dependence need not provide predictive benefit. To distinguish these cases, CELLAUDIT combines source inspection with matched input replacement (Fisher et al., 2019; Chamma et al., 2023) on frozen checkpoints, testing both whole-model dependence and target-relevant contribution. During development, the same diagnostics can guide revision; final claims are evaluated only after the selected source, checkpoint, and audit rules are frozen.

CELLAUDIT reveals these distinctions in agent-generated cellular-response predictors. On a paired morphology–transcriptomics perturbation task built from BBBC047 (Bray et al., 2017; Haghighi et al., 2022), an agent-selected model reaches a mean held-out Global PCC of 0.3153, compared with 0.3142 for a control-profile-only predictor, while remaining exactly invariant to compound replacement. Source inspection identifies a singleton key–value attention pathway whose query cannot transmit compound information. The model remains invariant after refitting with physically disjoint control wells for inputs and target references. A broader audit separates fitted dependence from evidence of predictive contribution. In a stratified sample of 48 generated candidates across the two linked tasks, 47 change predictions under compound replacement on both held-out folds, but only 20 have registered target-loss scaffold intervals that lie entirely above zero on both folds.

These diagnostics also distinguish what a revision needs to address. Making a perturbation pathway executable does not by itself ensure a target-relevant contribution. On BBBC047, falsification-guided revisions learn a residual over a control-profile-only predictor, recovering positive compound contributions together with predictive gains. In matched sci-Plex searches (Srivatsan et al., 2020), audit-enriched feedback produces higher mean held-out predictive performance and larger mean compound and dose contributions than score-only feedback. However, paired confidence intervals for these differences across five trajectories include zero, so the evidence for improved discovery remains suggestive rather than conclusive.

Finally, we distinguish *predictive generalization* from *claim generalization*. The latter asks whether a registered input-use claim remains supported when a fixed model design is refit on an independently acquired cohort. When LINCS-selected designs are fixed and refit on LKCP (Keenan et al., 2018; Subramanian et al., 2017; Weisbart et al., 2024), predictive gains and dose contribution persist, whereas compound-identity contribution changes substantially. A separate CRISPRa evaluation applies the same auditing principle to unseen genetic combinations (Norman et al., 2019). Together, these studies trace input-use claims from source code through fitted dependence to predictive contribution, use the resulting evidence to guide revision, and re-evaluate those claims on new data.

2 RELATED WORK

Perturbation prediction and evaluation. Cellular perturbation models use latent transfer, compositional representations, neural transport, genetic graphs, and foundation models to predict chemical or genetic responses (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Roohani et al., 2024; Cui et al., 2024). Benchmarks standardize evaluation across splits and baselines (Wu et al., 2025; Wenteler et al., 2025), while recent studies examine linear references, shared response variation, calibrated metrics, and distribution shifts (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026; Miller et al., 2025; Mao et al., 2026). These works ask how well perturbation responses can be predicted and evaluated; CELLAUDIT asks whether registered input-use claims are supported by a model's source code and fitted behavior.

Discovery, reliance, and falsification. Scientific agents generate and revise executable hypotheses, programs, and models (Boiko et al., 2023; Romera-Paredes et al., 2024; Jiang et al., 2025). Permutation reliance, conditional replacement, and behavioral testing probe fitted-model behavior under targeted input interventions (Fisher et al., 2019; Chamma et al., 2023; Ribeiro et al., 2020),

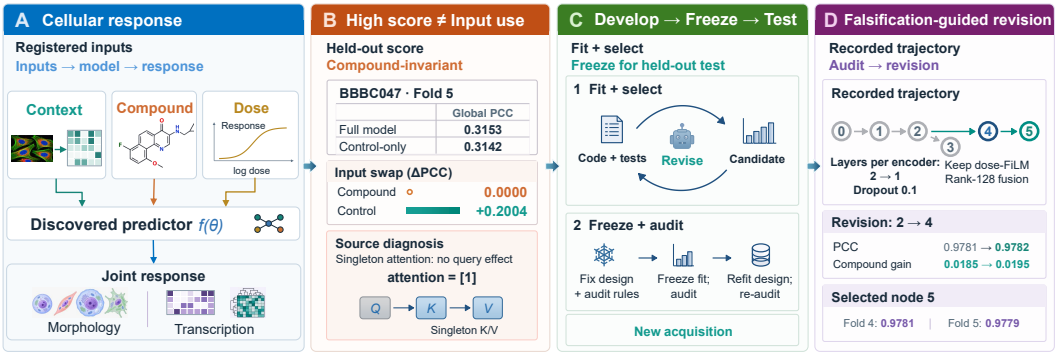


Figure 1: **Discover, falsify input-use claims, and revise.** (A) Registered context, compound, and dose inputs predict morphology and transcriptional responses. (B) CELLAUDIT identifies a high-scoring BBBC047 predictor with zero measured dependence on compound identity. (C) Development diagnostics guide revision; the selected source, checkpoint, and audit rules are frozen before held-out evaluation. (D) A recorded sci-Plex revision reduces encoder depth while retaining dose-conditioned FiLM and rank-128 fusion. Node 0 is the shared anchor, node 2 $\rightarrow$ 4 the highlighted revision, and node 5 the final selected endpoint for held-out evaluation. Appendix O gives the complete trajectories and source changes.

while ConceptSMILE studies explanation reliability and POPPER automates hypothesis falsification (Mollapour et al., 2026; Huang et al., 2025). Unlike reliance tests that begin from prediction behavior, CELLAUDIT starts with a registered input-use claim and its cited computation, then follows it through source consumption, fitted dependence, and target-relevant predictive contribution. This evidence can also guide model revision. Appendix K provides additional context.

3 PROBLEM SETUP: PREDICTIVE PERFORMANCE AND REGISTERED INPUT-USE CLAIMS

Each example contains biological context c, perturbation representation p, optional attributes a such as dose, and response $y = [y^{(1)}, \ldots, y^{(M)}]$. Source s and fitted parameters θ define $f_s(c, p, a; \theta)$; outputs may include imaging, molecular, or other readouts. Before search, the task registers input coordinates and admissible interventions. Each audit records whether an input-use claim comes from a candidate or the task contract and, after selection, fixes its cited computation and test mapping. A *prediction–claim gap* occurs when the cited source contradicts a registered claim or fitted predictions are invariant to input replacement. Predictive performance affects the importance of the case, not the definition; target-relevant contribution is separate.

Data processing, budgets, evaluation, and selection rules are fixed before search. Folds 1–2 fit candidates; Fold 3 provides feedback and selects checkpoints. Selected checkpoints are evaluated on Folds 4 and 5, except prediction-score discovery also selects a source across policies on Fold 4 and refits it for Fold 5. Norman instead selects sources before Fold 4 and refits the same sources on Fold 5 without reselection. Appendices B and N.1 specify study-specific refitting and reuse rules.

4 METHOD

4.1 AGENTIC MODEL SEARCH WITH FIXED DATA AND EVALUATION

The existing CELLSCIENTIST policy (Li et al., 2026) starts from a common predictor h_0. It receives the current model, development diagnostics, execution status, and compact history, and proposes a hypothesis and executable revision within the registered permissions. Shared preflight and training code provide up to three repairs per failed proposal; persistent failures consume a slot. Each executable candidate trains once, and a fixed Fold-3 rule selects the endpoint. Records preserve parent links, emitted hypotheses and declarations, source, diagnostics, repairs, checkpoint, and costs. The language model proposes revisions; input-use tests are deterministic.

4.2 CHECKING REGISTERED INPUT-USE CLAIMS IN SOURCE CODE

Given registered inputs and cited source locations, CELLAUDIT checks the computation attached to each input-use statement. Its abstract-syntax-tree checker returns *implemented*, *implementation-contradicted*, or *unresolved*. Coverage is the fraction of checked candidates whose cited locations are all resolved; unresolved code remains eligible for behavioral testing. With one key–value pair, attention has constant normalized weight and cannot transmit compound information through its query (Vaswani et al., 2017). This localizes a recognized implementation defect. Whole-model replacement then establishes whether fitted predictions depend on the input, including through other routes; target-loss contrasts test whether that dependence helps prediction.

4.3 TESTING FITTED DEPENDENCE AND PREDICTIVE CONTRIBUTION

With checkpoint and targets fixed, CELLAUDIT replaces one registered input x_q while retaining the others. Prediction distance measures target-blind dependence; zero identifies invariance on the tested replacements. To test predictive benefit, it computes the decrease in a higher-is-better score m and the increase in loss:

$$\Delta m_q = m(f(x), y) - m(f(x_{-q}, \tilde{x}_q), y), \quad \Delta\mathcal{L}_q = \mathcal{L}(f(x_{-q}, \tilde{x}_q), y) - \mathcal{L}(f(x), y). \tag{1}$$

Positive effects indicate benefit from the observed input relative to its registered replacements. Replacement distributions and their observed-support coverage are specified by task. We distinguish single-coordinate contrasts Δm_q from factorial effects E_q, which average over the other input's two states in the 2×2 compound–context audit. Loss effects $E_q^{\mathcal{L}}$ use the same averaging. The transfer analysis separates compound and dose allocations (Lundberg & Lee, 2017). Appendix L gives the full contrasts.

Effect size, uncertainty, and stability. Maps are repeated measurements within a checkpoint, averaged before across-model inference. Continuous contributions have paired training-seed or trajectory intervals; fixed-model source-group resampling tests variation across biological groups. Known-truth checks distinguish fixed-dataset and population uncertainty (Appendix F.1). BBBC, LINCS, and LKCP separately report whether the effect quantile interval lies *above*, *below*, or *across* a reference threshold. A positive mean contribution can occur in any category. Appendix B defines the threshold, quantiles, and original category labels; Norman uses a separate loss-based criterion (Appendix D).

4.4 USING FALSIFICATION EVIDENCE TO GUIDE THE NEXT SEARCH

The audit distinguishes two reasons to revise a model. A contradicted source path motivates an executable route; weak target contribution motivates learning an increment over context. Path-constrained discovery addresses the first by compiling structured design cards into models with an explicit compound route. BBBC falsification-guided discovery addresses the second through residual modeling and input-use feedback:

$$\hat{y} = g(c) + \alpha h(c, p, a), \tag{2}$$

where $g(c)$ is fitted on Folds 1–2 and fixed. The compiler enforces centered perturbation/dose branches, context gates, and bias-free readouts, giving zero residual at their joint reference input. Training and Fold-3 feedback use prediction error and replacement loss changes. BBBC047 selects by predictive score among qualifying candidates; BBBC036 prioritizes compound loss gain among predictively noninferior candidates. Each trajectory evaluates the anchor plus nine proposals; the harness selects scale α and checkpoint on Fold 3 and freezes that model for Folds 4 and 5. These formulations test the joint design changes; Appendix B specifies their rules.

Separating feedback from changes to the model space. The sci-Plex comparison instead shares a context-plus-dose anchor $g(c, a)$ and permissions for direct predictors or anchor residuals. Score feedback supplies metrics and learning curves; Audit adds grouped errors, source checks, and development-set compound/dose replacements. After each fit, CELLAUDIT computes these diagnostics and inserts them into the agent's next prompt, linking input-path findings and measured contributions to model revision. Loss, optimizer, training limits, and PCC-first selection stay fixed. Paired outcomes evaluate the complete audit-enriched feedback package; selected sources and checkpoints are frozen before both held-out evaluations.

5 EXPERIMENTAL DESIGN

5.1 DEVELOPMENT ON LINKED MULTIMODAL PERTURBATION TASKS

Development uses linked tasks from one paired cellular-profile release: BBBC036, the known-bioactive subset, and its parent BBBC047 (Haghighi et al., 2022; Bray et al., 2017; Subramanian et al., 2017). Each predicts joint morphology–transcription responses from a same-plate control profile, compound representation, and dose. Response-blind Bemis–Murcko scaffold folds (Bemis & Murcko, 1996) hold out compound families. Global PCC on concatenated responses selects models; MSE and modality-specific metrics diagnose gains. Appendix A gives provenance, transformations, fold inventories, and a complementary plate-family split.

Prediction-score discovery compares CellScientist, AIDE, CellForge, and HarmonyCell policies (Li et al., 2026; Jiang et al., 2025; Tang et al., 2025; Huang et al., 2026) under a matched initial model, executor, feedback, and ten-candidate budget, with ten trajectories per task and five paired refits per selected model. Both follow-up formulations retain the folds and budget and carry all ten selected designs through Folds 4 and 5; failures, costs, and repeated designs are retained (Appendix C).

Separating the control input from the target reference. For BBBC047, each plate's control wells are split into equal-size banks A/B. With the target reference fixed, shared and disjoint arms take the input profile from the same or the other bank, so their contrast isolates physical control-well overlap. Both orientations refit the score-selected source, a control-only source, and one fixed guided design under five seeds; all 60 checkpoints are fixed before evaluation on Folds 4 and 5 (Appendix A.1).

Matched feedback on sci-Plex. We predict 2,000-gene well-level responses in sci-Plex3 (Srivatsan et al., 2020), holding out compound–cell-line combinations, with whole treatment wells in one partition, while both marginals occur in training; gene selection and scaling use training wells only. Both conditions use the same DeepSeek serving model, five paired trajectory seeds, five new candidate slots per trajectory, and a shared anchor. All ten selected checkpoints are evaluated on two held-out combination partitions of this source. Compound replacements match dose and experimental context; dose replacements retain compound identity (Appendix N.1).

5.2 TESTING INPUT-USE CLAIMS ON NEW COHORTS AND PERTURBATION TYPES

The fixed falsification-guided protocol first runs on LINCS–Pilot1 (Keenan et al., 2018; Subramanian et al., 2017). Its ten selected designs and sources are then frozen and refit on independently acquired LKCP Batch 2 (Weisbart et al., 2024), with response- and fold-blind interface adaptation fixed before training. Compound permutations retain dose; dose permutations stay within compound (Appendix E). The Norman CRISPRa task predicts 5,045-gene responses for unseen double-perturbation combinations (Norman et al., 2019). Its registered input bundles multi-hot pair identity with the matching mean single-perturbation anchor. GEARS supplies a task-native reference (Roohani et al., 2024).

6 RESULTS

6.1 CELLAUDIT DETECTS COMPOUND INVARIANCE IN A HIGH-SCORING PREDICTOR

CELLAUDIT's automated audit identifies the BBBC047 source selected on Fold 4 as compound-invariant. Its source-selection score is 0.3035; mean Global PCC reaches 0.3153 on held-out replication Fold 5. Compound shuffling changes PCC by 0.0000 in every refit, whereas control replacement lowers PCC by 0.1932 on Fold 4 and 0.2004 on Fold 5 (Table 1). Control-only reaches 0.3142 on Fold 5; Full$-$control-only is $+0.0011$ (95% CI $[-0.0005, +0.0027]$). Prediction-space distance is exactly zero under compound replacement and large under control replacement (Appendix I).

The source checker identifies a concrete computation to revise: the cited compound-conditioned cross-attention has one key and one value, making its normalized weight constant. The behavioral test establishes complete-model invariance; the source result localizes an inactive cited pathway. A

compound-only positive control produces a shuffle effect of 0.0111 ± 0.0010, with all five refits above the reference threshold.

A source-stratified sample of 48 real candidates separates these outcomes further (Figure 2a,b). An implemented multi-tower source is exactly invariant; 47 candidates show fitted dependence on both folds, but only 20 have positive target-loss scaffold intervals on both. Source consumption and fitted dependence thus leave predictive contribution unresolved for many candidates (Appendix M.3). CELLSCIENTIST supplies a competitive discovery setting: mean Best@10 and frontier AUC lead both linked tasks, with intervals versus HarmonyCell crossing zero (Table 19).

Table 1: **Prediction, source implementation, and target-relevant compound use are distinct.** BBBC047: post-search Fold 4 and held-out replication Fold 5. Full/Control are mean Global PCC. E_{cmpd} and E_{ctrl} are factorial PCC effects averaged over the other input's two states; $E^{\mathcal{L}}_{\mathrm{cmpd}}$ is the corresponding loss effect. Reference counts locate effect quantile intervals relative to their thresholds. Rows summarize five prediction-score refits or ten endpoints per constrained search.

Search	Code check	Fold	Full	Control	E_{cmpd}	$E^{\mathcal{L}}_{\mathrm{cmpd}}$	E_{ctrl}	Reference Above/Below/Across
Prediction-score	Contradicted[a]	4	0.3035	0.3019	0.0000	0.00000	0.1932	0/5/0
		5	0.3153	0.3142	0.0000	0.00000	0.2004	0/5/0
Path-constrained	Implemented[b]	4	0.2822	–	0.0083	0.00024	0.1388	0/6/4
		5	0.2922	–	0.0074	0.00022	0.1477	0/7/3
Falsification-guided	Implemented[c]	4	0.3060	0.3023	0.0053	0.00018	0.1988	0/10/0
		5	0.3173	0.3145	0.0048	0.00017	0.2051	0/10/0

Control: control-only predictor. Reference: effect quantile interval above/below/across the reference threshold (Appendix B); distinct from a positive mean contribution. [a] One-key attention cannot transmit compound identity. [b] Compiler-enforced compound path. [c] Explicit compound-residual path.

6.2 THE MISMATCH PERSISTS WITH DISJOINT CONTROL REFERENCES

We test whether the mismatch persists when control inputs and CP target references use disjoint physical wells. Newly reconstructed shared and disjoint arms have identical targets within each bank orientation (Figure 2c,d). In the disjoint arm, the score-selected source reaches Global PCC 0.3085 and 0.3193, while control-only reaches 0.3082 and 0.3185 on Folds 4 and 5. The score-selected source remains exactly compound-invariant in every refit and evaluation. All six shared–disjoint PCC comparisons have five-seed intervals crossing zero. The detected mismatch therefore persists without physical control-well reuse at this reference-construction layer.

The fixed guided design retains both predictive and compound-use increments under the same separation (Table 6). Relative to disjoint control-only, its PCC gain is $+0.0028$ (95% paired-seed CI $[+0.0017, +0.0038]$) on Fold 4 and $+0.0025$ ($[+0.0012, +0.0037]$) on Fold 5. Compound replacement lowers PCC by 0.0043 and 0.0033 and increases target loss by 1.5106×10^{-4} and 1.1939×10^{-4}; each mean effect has a positive five-seed interval. Effects are slightly smaller than in the shared arm. Same-plate context and the original scaffold folds are retained (Appendix A.1).

6.3 AN EXPLICIT COMPOUND ROUTE AND ITS FITTED CONTRIBUTION

Path-constrained discovery addresses the source defect by compiling an explicit compound route: all 200 candidate evaluations pass source and interface checks without candidate-code repair. On BBBC047, mean factorial compound effects are 0.0083 and 0.0074, with positive mean loss effects on Folds 4 and 5. Relative to the reference threshold, Fold 4 has six below-threshold and four overlapping cases; Fold 5 has seven and three. Control-profile effects exceed the threshold throughout (Table 1). An executable route therefore admits compound dependence without ensuring a consistently above-reference contribution in each model.

6.4 GUIDED RESIDUAL DISCOVERY COMBINES PREDICTIVE GAINS WITH TARGET-RELEVANT COMPOUND CONTRIBUTION

Falsification-guided discovery next asks the perturbation residual to improve prediction beyond a fixed control-profile-only model. On BBBC047, the full predictor exceeds that baseline by $+0.0037$

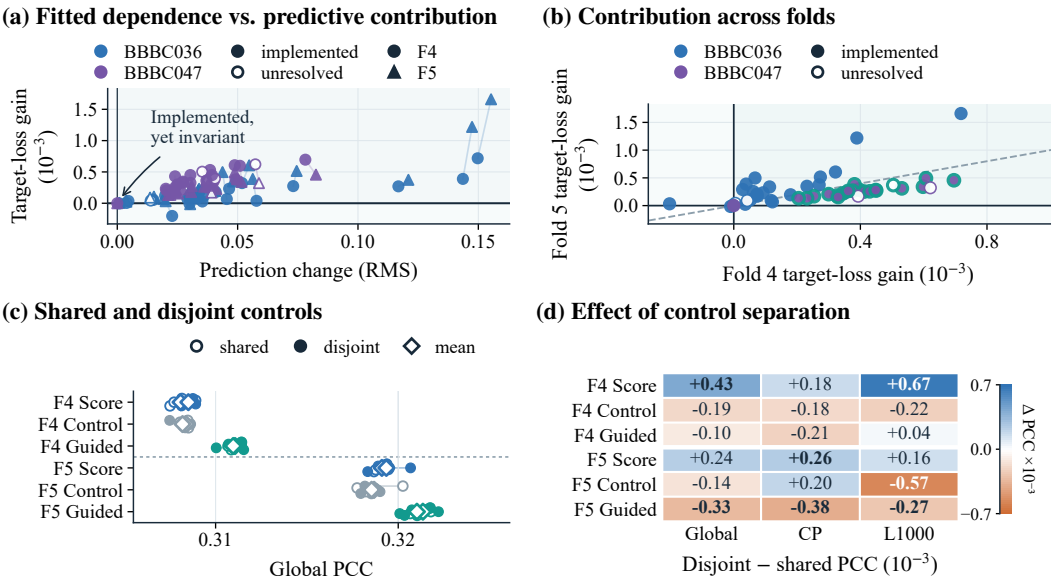


Figure 2: **Fitted dependence and predictive contribution are distinct.** (a,b) Forty-eight sources across the two linked tasks; filled/hollow markers denote implemented/unresolved paths, and green outlines mark the 20/48 sources with positive target-loss scaffold intervals on both folds. Lines in (a) pair folds. (c) Five paired seeds, averaging control-bank orientations within seed. (d) Mean disjoint-minus-shared PCC.

on Fold 4 and $+0.0028$ on Fold 5. Mean factorial compound effects are 0.0053 and 0.0048, with positive loss effects (Table 1); joint model–scaffold intervals for these increments are above zero. The explicit-route models have larger compound effects but lower predictive scores. Guided discovery combines higher scores with positive target-relevant effects, although none of its ten models exceeds the reference-relative criterion.

For the original guided checkpoints, observed-support replacements retain positive contributions on BBBC047 while holding context and dose fixed. They cover $3341/4380$ and $2927/3878$ rows on Folds 4 and 5: compound PCC drops are 0.00534 and 0.00404, and loss gains are 1.84×10^{-4} and 1.44×10^{-4}, all with positive source-scaffold intervals. The score-selected model remains exactly invariant on these same eligible rows (Table 57).

Fixed residual learners clarify the trade-off: Ridge has strong compound sensitivity but low joint-response scores, and on BBBC047 the residual MLP has larger compound effects but lower scores than the guided models, which lead both references in mean Global PCC at all four BBBC evaluations. On BBBC036 Fold 5, their gain over the context anchor is -0.0020, with model–scaffold intervals for predictive and compound increments crossing zero (Table 16).

Checking what these effects measure. Known-function experiments separate invariance, sensitivity without benefit, benefit, and harm across 6,400 independent simulated datasets. On the sensitive-without-benefit null, map-only intervals attain near-nominal coverage for the fixed-dataset conditional effect but severely undercover the population effect. Source-group intervals improve population coverage, with finite-group undercoverage remaining. Holding compound benefit fixed while changing reference-context variation can also change qualification. These checks support reporting continuous contribution, across-model uncertainty, and reference-relative status as distinct results (Appendix F.1).

6.5 Audit-enriched feedback yields higher mean held-out prediction

Do audit measurements help when returned to an agent? Figure 3a retains all five paired sci-Plex search frontiers. The context-plus-dose anchor already predicts much of the whole-expression profile, making incremental error reduction informative. Audit feedback yields higher mean PCC and

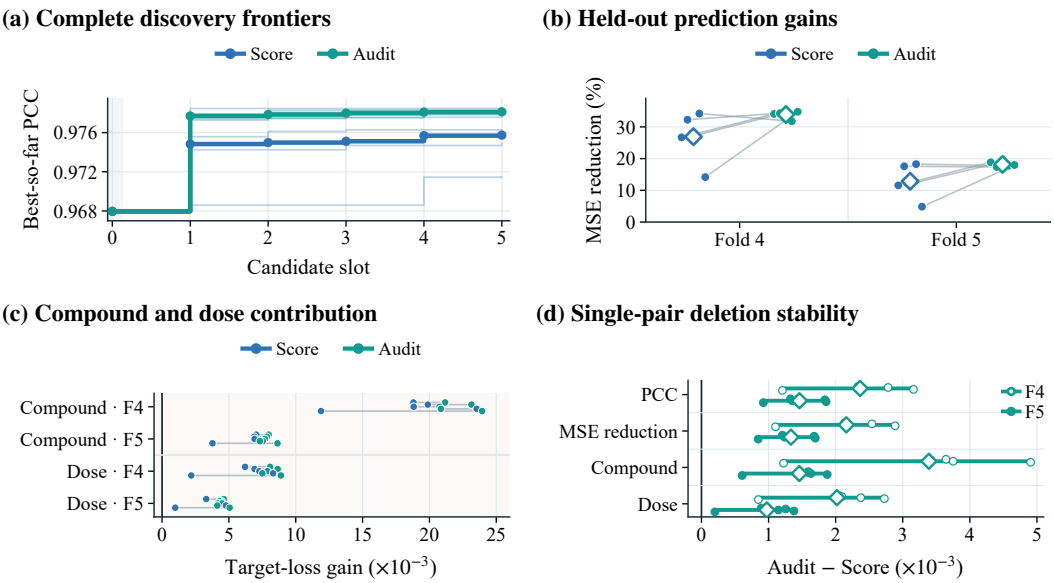


Figure 3: **Audit feedback connects search progress to held-out contributions.** (a) Five trajectories per condition and their mean frontiers; slot 0 is the anchor. (b) Relative MSE reduction against each paired anchor. (c) Compound/dose loss gains; lines pair trajectories. (d) Leave-one-pair-out means for the absolute Audit–Score differences; the MSE row is shown as a favorable reduction, so its scale is not the relative percentage scale in (b). Bars span the resulting means, not confidence intervals; diamonds summarize all five pairs.

lower mean MSE on both later partitions (Figure 3b): MSE is 9.58% lower than Score on Fold 4 and 6.10% lower on Fold 5. Both improve on the anchor.

The higher mean predictive performance of Audit is accompanied by higher mean compound and dose contributions (Figure 3c). Compound target-loss gain rises from 0.0186 to 0.0220 on Fold 4 and from 0.0064 to 0.0079 on Fold 5. Paired t intervals for these Audit–Score contribution contrasts include zero at both folds (Table 68). Correct inputs help prediction in all five endpoints of both conditions. Audit–Score PCC is $+0.0024$ (paired 95% CI $[-0.0013, +0.0060]$) on Fold 4 and $+0.0015$ ($[-0.0005, +0.0034]$) on Fold 5. Both means, and the favorable MSE and compound loss-gain directions, persist after removing any one trajectory pair (Figure 3d), whereas prediction-RMS differences can reverse after one omission. Percentile-bootstrap predictive intervals exclude zero even when resampling only trajectories, so uncertainty depends on interval construction at five pairs (Appendix N.2). Both conditions complete all 25 new candidates; Audit uses 38 versus 36 logical model calls and 592,347 versus at least 286,995 reported tokens (Table 65).

From automatic audit output to a recorded revision. In the first registered Audit trajectory, CELLAUDIT returns source checks, compound/dose replacement effects, grouped errors, and learning curves for the bilinear incumbent and a non-improving challenger. The agent returns to that incumbent and proposes simpler encoders with dropout, retaining dose-conditioned FiLM. Parameters decrease while selection PCC and compound loss gain increase (Figure 1D); search continues to a distinct final endpoint (Appendix O).

6.6 PREDICTIVE GAINS AND INPUT-USE SUPPORT GENERALIZE DIFFERENTLY UNDER INDEPENDENT ACQUISITION

To test claim generalization, the frozen LINCS-selected designs are refit on LKCP without source revision, model reselection, or audit-rule changes. They improve over the control-profile-only baseline by $+0.0104$ (95% CI $[+0.0103, +0.0106]$) on Fold 4 and $+0.0211$ (95% CI $[+0.0205, +0.0217]$) on Fold 5. Figure 4a,b separates compound identity from dose. On Folds 4 and 5, the dose-replacement Shapley allocation is $+0.0486$ and $+0.0467$ Global PCC on LINCS and $+0.0237$ and $+0.0341$ on LKCP, whereas the compound-identity allocation is $+0.0175$ and $+0.0126$ on LINCS but -0.000070 and $+0.0033$ on LKCP. All 50 trajectory-endpoint refits pass the dose-use crite-

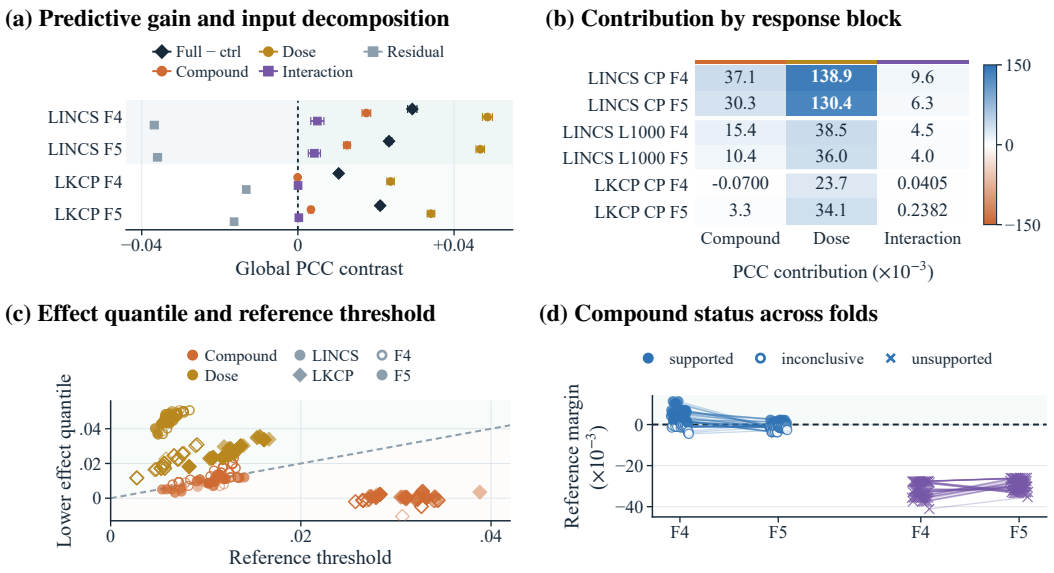


Figure 4: **Predictive and input-use evidence generalize differently across acquisitions.** (a) Means and 95% intervals over ten endpoint instances; interaction is separate from the additive allocations. (b) LINCS CP/L1000 and LKCP morphology. (c) Lower effect quantiles and thresholds for 400 input–checkpoint evaluations. (d) Paired compound reference margins; the left pair is LINCS and the right pair is LKCP, with lines joining the same endpoint–seed refit across folds. Marker states denote supported, inconclusive, and unsupported refits. Each cohort has ten endpoint instances with five refits each.

rion at every cohort–fold boundary, but none passes the compound-identity criterion at either LKCP boundary. LINCS shows why average effects and per-model decisions must remain separate: its compound-identity criterion is passed by 35/50 refits on Fold 4 but 10/50 on Fold 5 (mean lower effect quantile 0.0139 to 0.0090), although average compound contribution remains positive.

6.7 The audit detects dependence on the perturbation-pair input bundle

On Norman, all five CellScientist refits pass the registered pair-input-bundle criterion on both held-out folds; shuffling the multi-hot pair identity together with its matching single-perturbation anchor reduces PCC by 0.5578 and 0.6280. Mean PCC exceeds h_0 by $+0.0060$ and $+0.0045$ and the observed-single additive baseline by $+0.0209$ and $+0.0114$ (Table 26; Appendix D).

7 Discussion and conclusion

CellAudit shows that predictive success and claimed input use should be evaluated separately. Source consumption, fitted dependence, and target-relevant contribution can diverge: on BBBC047, a high-scoring source remains compound-invariant after disjoint control-reference refitting, while falsification-guided revisions recover positive compound contributions with predictive gains. The broader candidate audit shows this is not specific to singleton attention. In sci-Plex, audit-enriched feedback improves mean prediction and input contributions, but paired intervals over five trajectories span zero. Independent-acquisition refits further show that predictive generalization need not imply claim generalization.

These conclusions are conditional on the fitted model, evaluation setting, and registered replacement distribution. Positive contribution is distinct from passing the reference-relative criterion, and support for one fit should not be inherited by later refits. Observed-support replacements avoid unobserved combinations but do not establish biological exchangeability, mechanism, or causal effects. The sci-Plex study evaluates the complete audit-feedback package rather than individual diagnostics. CellAudit therefore adds a falsification layer to agentic model discovery, tracing input-use claims from cited computation to fitted behavior and predictive contribution.

AI USE STATEMENT

Generative AI was used as part of the experimental model-discovery workflows studied in this work. Language-model agents proposed model hypotheses, executable source-code revisions, and structured model-design revisions under the fixed experimental protocols described in Sections 4–5 and the appendices. The resulting candidates were trained and evaluated by the registered experimental harness, while held-out input-use audits and statistical analyses were performed by deterministic procedures. Generative AI was additionally used to assist with literature retrieval and review, and with language polishing and readability improvements. The authors take responsibility for the final methods, results, claims, and manuscript.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016. doi: 10.1038/nprot.2016.105. URL https://doi.org/10.1038/nprot.2016.105.

Mark-Anthony Bray, Sigrun M. Gustafsdottir, Mohammad H. Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L. Sokolnicki, Joshua A. Bittker, et al. A dataset of images and morphological profiles of 30,000 small-molecule treatments using the Cell Painting assay. *GigaScience*, 6 (12):giw014, 2017. doi: 10.1093/gigascience/giw014. URL https://doi.org/10.1093/gigascience/giw014.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x. URL https://doi.org/10.1038/s41592-023-01969-x.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25): 7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. URL https://doi.org/10.1016/j.cell.2024.11.015.

Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017. doi: 10.1038/nmeth.4397. URL https://doi.org/10.1038/nmeth.4397.

Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion. Statistically valid variable importance assessment through conditional permutations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67662–67685, 2023. doi: 10.52202/075280-2959. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d60e14c19cd6e0fc38556ad29ac8fbc9-Abstract-Conference.html.

Srinivas N. Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G. Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024. doi: 10.1038/s41592-024-02241-6. URL https://doi.org/10.1038/s41592-024-02241-6.

Elijah Cole, Geert-Jan Huizing, Sohan Addagudi, Nicholas Ho, Euxhen Hasanaj, Merel Kuijs, Toby Johnstone, Maria Carilli, Alec Davi, Caleb Ellington, Christoph Feinauer, Pan Li, Romain Menegaux, Shahin Mohammadi, Yanjun Shao, Josiah Zhang, Emma Lundberg, Le Song, Ziv Bar-Joseph, and Eric P. Xing. Foundation models improve perturbation response prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.18.706454. URL https://www.biorxiv.org/content/10.64898/2026.02.18.706454v1. Preprint.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7. URL https://doi.org/10.1038/s42256-021-00338-7.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038. URL https://doi.org/10.1016/j.cell.2016.11.038.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL https://www.jmlr.org/papers/v20/18-760.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. TabR: Tabular deep learning meets nearest neighbors. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rhgIgTSSxW.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Sd4wYYOhmY.

Marzieh Haghighi, Juan C. Caicedo, Beth A. Cimini, Anne E. Carpenter, and Shantanu Singh. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 19(12):1550–1557, 2022. doi: 10.1038/s41592-022-01667-0. URL https://doi.org/10.1038/s41592-022-01667-0.

Euxhen Hasanaj, Elijah Cole, Shahin Mohammadi, Sohan Addagudi, Xingyi Zhang, Le Song, and Eric P. Xing. Multimodal benchmarking of foundation model representations for cellular perturbation response prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661186. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661186v2. Preprint.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned mlps and boosted trees on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26577–26658, 2024. doi: 10.52202/079017-0837. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ee1c87245956e3eaa71aaba5f5753eb-Abstract.html.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24y.html.

Wei Huang and Hui Liu. Predicting single-cell cellular responses to perturbations using cycle consistency learning. *Bioinformatics*, 40(Suppl. 1):i462–i470, 2024. doi: 10.1093/bioinformatics/btae248. URL https://academic.oup.com/bioinformatics/article/40/Supplement_1/i462/7700878.

Wenxuan Huang, Mingyu Tsoi, Yanhao Huang, Xinjie Mao, Xue Xia, Hao Wu, Jiaqi Wei, Yuejin Yang, Lang Yu, Cheng Tan, et al. HarmonyCell: Automating single-cell perturbation modeling under semantic and distribution shifts. *arXiv preprint arXiv:2603.01396*, 2026. doi: 10.48550/arXiv.2603.01396. URL https://arxiv.org/abs/2603.01396.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025. doi: 10.48550/arXiv.2502.13138. URL https://arxiv.org/abs/2502.13138.

Alexandra B. Keenan, Sherry L. Jenkins, Kathleen M. Jagodnik, Simon Koplev, Edward He, Denis Torre, Zichen Wang, Anders B. Dohlman, Moshe C. Silverstein, Alexander Lachmann, et al. The library of integrated network-based cellular signatures NIH program: System-level cataloging of human cells response to perturbations. *Cell Systems*, 6(1):13–24, 2018. doi: 10.1016/j.cels.2017.11.001. URL https://doi.org/10.1016/j.cels.2017.11.001.

Benjamin Kuznets-Speck, Nitu Kumari, Irish Senthilkumar, Jaekwon Jung, Hector Manuel Lopez Rios, Leon Schwartz, Hanxiao Sun, Vidyesh Rao Anisetti, Changhu Wang, Qingyang Wang, Pornchanan Pholraksa, Emanuelle I. Grody, Madeline E. Melzer, Benjamin Haley, Ekta Prashnani, Jingyi Jessica Li, Akshit Goyal, Suriyanarayanan Vaikuntanathan, and Yogesh Goyal. Co-fluctuations of genes in single cells predict transcriptome-wide outcomes to perturbations. *bioRxiv*, 2025. doi: 10.1101/2025.06.27.661814. URL https://www.biorxiv.org/content/10.1101/2025.06.27.661814v2. Preprint, version 2.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4. URL https://doi.org/10.1038/s41467-019-08987-4.

Mengran Li, Bo Li, Jiaying Wang, Wenbin Xing, Yixuan Dong, Chengyang Zhang, Hongliang Zhang, Yuzhong Peng, Jinlin Wu, Bob Zhang, Bingo Wing-Kuen Ling, Fuji Yang, Zhen Lei, Jiebo Luo, and Zelin Zang. CellScientist: Dual-space hierarchical orchestration for closed-loop refinement of virtual cell models. *arXiv preprint arXiv:2605.07335*, 2026. doi: 10.48550/arXiv.2605.07335. URL https://arxiv.org/abs/2605.07335.

Michael Y. Li, Emily Fox, and Noah Goodman. Automated statistical model discovery with language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27791–27807. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24v.html.

Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637, 2012. doi: 10.1038/nmeth.2083. URL https://doi.org/10.1038/nmeth.2083.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517. URL https://doi.org/10.15252/msb.202211517.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Xinjie Mao, Songming Zhang, Qianhong Wen, Xiangyu Wen, Kedu Jin, Hao Wu, Shuizhou Chen, Yuqiang Li, Lei Bai, Qi Liu, Ning Ding, Siqi Sun, and Zhangyang Gao. Benchmarking virtual cell models for in-the-wild perturbation response. *arXiv preprint arXiv:2604.27646*, 2026. doi: 10.48550/arXiv.2604.27646. URL https://arxiv.org/abs/2604.27646.

Henry E. Miller, Gabriel M. Mejia, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Deep learning-based genetic perturbation models do outperform uninformative baselines on well-calibrated metrics. *bioRxiv*, 2025. doi: 10.1101/2025.10.20.683304. URL https://www.biorxiv.org/content/10.1101/2025.10.20.683304v1. Preprint.

Mohadeseh Mollapour, Koorosh Aslansefat, Zeinab Dehghani, Bhupesh Kumar Mishra, Tejal Shah, and Zhibao Mian. ConceptSMILE: Auditing the trustworthiness of concept-based explainable AI. *arXiv preprint arXiv:2607.09649*, 2026. doi: 10.48550/arXiv.2607.09649. URL https://arxiv.org/abs/2607.09649.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://doi.org/10.1126/science.aax4438. Data available under GEO accession GSE133344.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: Harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y. URL https://doi.org/10.1038/s41592-023-02144-y.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL https://doi.org/10.1038/s41586-023-06924-6.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://doi.org/10.1126/science.aax6234.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

Xiangru Tang, Zhuoyun Yu, Jiapeng Chen, Yan Cui, Daniel Shao, Weixu Wang, Fang Wu, Yuchen Zhuang, Wenqi Shi, Zhi Huang, et al. CellForge: Agentic design of virtual cell models. *arXiv preprint arXiv:2508.02276*, 2025. doi: 10.48550/arXiv.2508.02276. URL https://arxiv.org/abs/2508.02276.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8. URL https://doi.org/10.1038/s41587-025-02777-8.

Erin Weisbart, Ankur Kumar, John Arevalo, Anne E. Carpenter, Beth A. Cimini, and Shantanu Singh. Cell painting gallery: An open resource for image-based profiling. *Nature Methods*, 21(10):1775–1777, 2024. doi: 10.1038/s41592-024-02399-z. URL https://doi.org/10.1038/s41592-024-02399-z.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66633–66677. PMLR, 2025. URL https://proceedings.mlr.press/v267/wenteler25a.html.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 106937–106977, 2025. doi: 10.52202/085713-3225. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8aee537279a66ced96319dfca3c00002-Abstract-Datasets_and_Benchmarks_Track.html.

A Data provenance, transformations, and task statistics

BBBC036 is the known-bioactive subset of its parent BBBC047 screen; the two tasks are therefore linked cohorts from one paired cellular-profile release rather than independent acquisitions. Their source collections in `cpg0003-rosetta` are CDRPBIO-BBBC036-Bray and CDRP-BBBC047-Bray, and the abbreviated task names retain the formal BBBC accession mapping (Ljosa et al., 2012). Inputs come from released replicate-level Cell Painting and L1000 profile tables (Weisbart et al., 2024; Haghighi et al., 2022). The morphology screen originates from the CDRP/BBBC047 Cell Painting resource (Bray et al., 2017), and the transcriptional profiles use the L1000 platform (Subramanian et al., 2017). Both assays apply matched compound–dose conditions on parallel plates and are joined at the condition level.

Table 2 reports both the released profile rows and the matched samples used for modeling. BBBC036 begins with 21,122 CP and 6,929 L1000 replicate profiles and retains 1,916 matched compound–dose conditions; BBBC047 begins with 153,386 CP and 68,120 L1000 profiles and retains 20,081 matched conditions. The matched subsets span 47 CP/22 L1000 treatment plates for BBBC036 and 273 CP/360 L1000 treatment plates for BBBC047. Before aggregation, 143 BBBC047 L1000 treatment profiles without a same-plate control are removed.

For each condition, c is the featurewise median of same-plate control-only CP profiles. Compound identity p is represented by a 2,048-bit radius-2 Morgan fingerprint (Rogers & Hahn, 2010), while the separate attribute a is the log-transformed observed dose. Before treatment filtering and condition aggregation, numeric profile columns are converted to float32 and any column containing a non-finite value in the released profile table is removed. CP treatment responses undergo the empirical same-plate-control mid-CDF transform, are centered by subtracting 0.5, and are aggregated by condition median; L1000 responses are centered by the same-plate control median and likewise aggregated by condition median. The resulting joint targets contain 591 CP and 977 L1000 outputs for BBBC036, and 701 CP and 977 L1000 outputs for BBBC047, yielding 1,568 and 1,678 dimensions. For training, each target coordinate is centered and scaled using its Folds 1–2 mean and population standard deviation; scales at or below 10^{-8} are replaced by one. Reported metrics invert this standardization, retaining the control-transformed response scale. Bemis–Murcko groups (Bemis & Murcko, 1996) are assigned without response values and remain disjoint across all five folds. Table 3 lists the complete fold inventory and experimental roles.

For the complementary plate-family/context transfer, connected components are formed from released CP treatment-plate identifiers, their same-plate control contexts, linked L1000 treatment plates, and matched-condition counts, without reading responses. The construction yields six components for BBBC036 and 64 for BBBC047, assigning every CP/context and linked L1000 plate to exactly one role. Fit/selection/audit/replication condition counts are 640/320/637/319 for BBBC036 and 8,264/3,868/3,826/4,123 for BBBC047. BBBC036 therefore provides a six-family transfer case, while BBBC047 supplies broader 64-family coverage.

Plate-family 2×2 audit. The plate-family audit crosses the observed or permuted compound representation with the observed or an alternative control profile from the same plate family. Each selected model is refit under five fixed seeds, and each refit averages 32 permutations constructed without response values. Compounds are permuted within the observed control-profile group whenever possible, covering every held-out BBBC036 row and 98.8% of BBBC047 rows; each remaining row receives a fixed donor with a different compound from the same held-out partition. Table 5 separates compound, control-profile, and interaction effects on Folds 4 and 5.

Table 2: **Paired morphology–transcriptomic task cohorts after consistent preprocessing.** Source counts are replicate-level profile rows from the released tables, not image or cell counts; parentheses give treatment/control rows. Matched conditions are compound–dose samples retained after condition-level matching across assays.

Cohort	CP source profiles	L1000 source profiles
A. Released source profiles		
BBBC036	21,122 (17,594/3,528)	6,929 (3,451/3,478)
BBBC047	153,386 (126,814/26,572)	68,120 (64,642/3,478)

Parentheses: treatment/control profile rows.

Cohort	Matched conditions	Unique SMILES	Murcko groups	CP/context dim.	L1000 dim.	Joint target dim.
B. Matched samples and model dimensions						
BBBC036	1,916	1,916	1,196	591	977	1,568
BBBC047	20,081	20,068	4,540	701	977	1,678

Table 3: **Fold-level inventory for the paired task cohorts.** Folds 1–2 fit models; Fold 3 supplies search feedback and within-policy selection. Fold 4 is held out from discovery and supplies the disclosed across-policy source selection in the prediction-score panel. That panel independently refits the selected source for Fold 5; constrained searches evaluate the same frozen checkpoint on both later folds (Appendix B). Scaffold groups and assay replicates are counted after filtering and condition matching.

Statistic	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
BBBC036					
Matched conditions	285	509	356	411	355
Murcko groups	206	244	258	265	223
CP replicate profiles	2,222	3,995	2,779	3,213	2,803
L1000 replicate profiles	519	919	626	739	648
BBBC047					
Matched conditions	3,788	3,921	4,114	4,380	3,878
Murcko groups	853	912	941	913	921
CP replicate profiles	15,891	17,060	17,263	18,390	16,429
L1000 replicate profiles	10,722	10,692	11,531	12,252	10,815

Joint residual references. Fixed residual baselines isolate a perturbation-specific function class while keeping the same task. They retain the selected control-profile predictor $g(c)$ and learn one joint residual $h(p, a)$ using either multi-output Ridge or a shallow multi-output MLP. Both baselines use Fold 3 for selection and are refit under the same five seeds. Their Fold-4 and Fold-5 predictive scores, compound effects, and target-loss gains are reported together in Table 48.

Two additional small-molecule cohorts test the final procedure beyond BBBC development. We first apply it to paired morphology–transcription LINCS–Pilot1, then transfer its ten trajectory-selected design instances to independently acquired Cell Painting LKCP Batch 2 under the same compound, dose, and control-profile inputs. Appendix E records data provenance, the order in which models and rules were fixed, the input permutations, and the compound–dose decomposition.

Table 4: Generalization of selected predictors to unseen plate families. The split is constructed without response values and models are selected before evaluation; each value is the mean [95% CI] over five paired refits. The same-plate control profile remains available as an input, while all Cell Painting treatment/control plates and linked L1000 plates are disjoint across split roles.

Frozen source	Audit Global PCC	Replication Global PCC	Audit CP PCC	Audit L1000 PCC	Parameters
BBBC036					
h_0	-0.0030 [-0.0400, 0.0339]	0.0484 [0.0223, 0.0746]	0.0255	-0.0130	1,144,864
CellScientist	**0.1921** [0.1903, 0.1939]	**0.1567** [0.1517, 0.1617]	0.2758	0.1568	3,999,776
AIDE	<u>0.1746</u> [0.1495, 0.1998]	<u>0.1497</u> [0.1193, 0.1800]	0.2520	0.1400	4,419,872
CellForge	-0.0091 [-0.0397, 0.0215]	0.0640 [0.0187, 0.1093]	0.0821	-0.0435	3,824,672
HarmonyCell	0.0657 [0.0378, 0.0936]	0.0102 [-0.0426, 0.0629]	0.0486	0.0719	2,061,343
BBBC047					
h_0	<u>0.1805</u> [0.1769, 0.1840]	<u>0.2014</u> [0.1941, 0.2087]	0.2688	0.0494	1,201,294
CellScientist	0.1781 [0.1716, 0.1846]	**0.2036** [0.1940, 0.2132]	0.2615	0.0469	1,777,486
AIDE	<u>0.1805</u> [0.1769, 0.1840]	<u>0.2014</u> [0.1941, 0.2087]	0.2688	0.0494	1,201,294
CellForge	0.1771 [0.1716, 0.1826]	0.2011 [0.1963, 0.2059]	0.2652	0.0444	2,531,982
HarmonyCell	**0.1808** [0.1699, 0.1916]	0.1963 [0.1921, 0.2006]	0.2689	0.0484	2,284,875

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 5: Compound/control audit on unseen plate families. Values are means [95% CI] across five paired training refits; each effect averages 32 fixed matched permutations within a refit. Positive compound and control effects denote a loss of Global PCC after the corresponding input intervention. Scientific notation preserves nonzero values below the four-decimal rounding threshold of 5×10^{-5}.

Model	Global PCC	Compound effect	Control effect	Interaction
BBBC036 / Audit				
h_0	-0.0030 [-0.0400, 0.0339]	0.0025 [0.0020, 0.0031]	0.0012 [-0.0126, 0.0150]	-0.0005 [-0.0008, -0.0002]
CellScientist	0.1921 [0.1903, 0.1939]	0.0000 [0.0000, 0.0000]	0.0146 [0.0114, 0.0178]	0.0000 [0.0000, 0.0000]
AIDE	0.1746 [0.1495, 0.1998]	0.0024 [0.0019, 0.0028]	0.0260 [0.0076, 0.0445]	-0.0008 [-0.0012, -0.0003]
CellForge	-0.0091 [-0.0397, 0.0215]	0.0018 [0.0011, 0.0025]	-0.0086 [-0.0169, -0.0004]	-0.0005 [-0.0010, 0.0001]
HarmonyCell	0.0657 [0.0378, 0.0936]	0.0019 [0.0002, 0.0037]	0.0768 [0.0632, 0.0905]	-0.0028 [-0.0044, -0.0013]
BBBC047 / Audit				
h_0	0.1805 [0.1769, 0.1840]	0.0046 [0.0039, 0.0053]	0.0249 [0.0228, 0.0271]	0.0008 [0.0005, 0.0012]
CellScientist	0.1781 [0.1716, 0.1846]	0.0000 [0.0000, 0.0000]	0.0295 [0.0243, 0.0346]	0.0000 [0.0000, 0.0000]
AIDE	0.1805 [0.1769, 0.1840]	0.0046 [0.0039, 0.0053]	0.0249 [0.0228, 0.0271]	0.0008 [0.0005, 0.0012]
CellForge	0.1771 [0.1716, 0.1826]	0.0050 [0.0039, 0.0061]	0.0208 [0.0163, 0.0252]	0.0014 [0.0004, 0.0023]
HarmonyCell	0.1808 [0.1699, 0.1916]	6.892×10^{-9} $[-1.390, 2.768] \times 10^{-8}$	0.0365 [0.0245, 0.0484]	1.211×10^{-8} $[-2.216, 4.637] \times 10^{-8}$
BBBC036 / Held-out replication				
h_0	0.0484 [0.0223, 0.0746]	0.0026 $[9.873 \times 10^{-6}, 0.0053]$	0.0043 [-0.0117, 0.0203]	0.0002 [-0.0006, 0.0009]
CellScientist	0.1567 [0.1517, 0.1617]	0.0000 [0.0000, 0.0000]	-0.0070 [-0.0088, -0.0053]	0.0000 [0.0000, 0.0000]
AIDE	0.1497 [0.1193, 0.1800]	0.0013 [-0.0004, 0.0031]	-0.0016 [-0.0168, 0.0136]	-0.0001 [-0.0004, 0.0002]
CellForge	0.0640 [0.0187, 0.1093]	0.0034 [0.0026, 0.0043]	0.0018 [-0.0087, 0.0123]	0.0001 [-0.0006, 0.0009]
HarmonyCell	0.0102 [-0.0426, 0.0629]	-0.0063 [-0.0100, -0.0025]	-0.0038 [-0.0350, 0.0274]	-0.0037 [-0.0080, 0.0005]
BBBC047 / Held-out replication				
h_0	0.2014 [0.1941, 0.2087]	0.0030 [0.0023, 0.0038]	0.0136 [0.0077, 0.0194]	0.0002 [-0.0010, 0.0013]
CellScientist	0.2036 [0.1940, 0.2132]	0.0000 [0.0000, 0.0000]	0.0231 [0.0168, 0.0293]	0.0000 [0.0000, 0.0000]
AIDE	0.2014 [0.1941, 0.2087]	0.0030 [0.0023, 0.0038]	0.0136 [0.0077, 0.0194]	0.0002 [-0.0010, 0.0013]
CellForge	0.2011 [0.1963, 0.2059]	0.0037 [0.0030, 0.0044]	0.0088 [0.0029, 0.0148]	-0.0004 [-0.0010, 0.0002]
HarmonyCell	0.1963 [0.1921, 0.2006]	3.073×10^{-9} $[-5.460, 11.61] \times 10^{-9}$	0.0205 [0.0142, 0.0268]	2.608×10^{-9} $[-4.632, 9.848] \times 10^{-9}$

A.1 PHYSICAL SEPARATION OF CONTROL INPUTS AND TARGET REFERENCES

Table 6: BBBC047 with shared control wells or disjoint control wells. Each arm uses newly constructed data and the same fixed model design.

Model	Fold	**Shared** PCC	**Disjoint** PCC	**Shared** PCC drop	**Disjoint** PCC drop	**Shared** Loss gain	**Disjoint** Loss gain
Score	F4	0.3081	0.3085	0	0	0	0
Control only	F4	0.3084	0.3082	0	0	0	0
Guided	F4	0.3111	0.3110	0.0047	0.0043	1.65×10^{-4}	1.51×10^{-4}
Score	F5	0.3191	0.3193	0	0	0	0
Control only	F5	0.3187	0.3185	0	0	0	0
Guided	F5	0.3214	0.3210	0.0036	0.0033	1.31×10^{-4}	1.19×10^{-4}

Means of five seeds, averaging directions A/B within seed; compound replacement uses 32 fixed maps. PCC drop and loss gain are correct-minus-replacement PCC and replacement-minus-correct MSE. Guided minus control-only disjoint-arm PCC: F4 0.0028 $[0.0017,\ 0.0038]$, F5 0.0025 $[0.0012,\ 0.0037]$ (paired-seed 95% t CIs).

Construction and scope. The shared and disjoint arms are newly constructed BBBC047 comparisons; the historical shared-control result is not their absolute baseline. Physical control wells are hash-sorted within each plate and well-row stratum and allocated to balanced, disjoint A/B banks (seed 2026091401). Within direction A, the target uses bank A as reference: shared context uses A, disjoint context uses B; direction B reverses these roles. CP targets are treated-well empirical mid-CDF values relative to the direction's reference bank, centered by 0.5 and aggregated by condition medians. The 701 CP features are fixed from the historical schema; no new feature selection is performed. Shared/disjoint arms have byte-identical targets and row identities within direction, and L1000 targets are unchanged. A/B are averaged within each seed, not counted as independent replicas. This intervention separates physical wells at the reconstructed CP reference layer; it does not establish that all upstream preprocessing or biological dependence has been removed. No historical qualification label is changed.

Table 7: Construction and inference units for the new control-well comparison.

Quantity	Value	Interpretation
Predictor refits / anchors	60 / 20	80 training stages; 120 fold evaluations
Seeds / directions	5 / 2	A/B are paired inside each seed
F4 retained support	3341 / 4380	76.28% of evaluation rows
F5 retained support	2927 / 3878	75.48% of evaluation rows
F4 / F5 source scaffolds	913 / 921	Bootstrap over source-scaffold labels
Maps / scaffold draws	32 / 2000	All refits and donor maps held fixed
Donor kernel total variation	0 (both folds)	Every eligible compound contributes one row
Guided coordinate signs	40 / 40 positive	RMS, loss gain and PCC drop, each record
Score/control coordinates	80 / 80 exact zero	All three compound coordinates
Physical controls, A / B	9144 / 9144	Unique plate/well identities; banks disjoint
CP plates / retained conditions	273 / 20081	0 conditions excluded

Only row-uniform maps were evaluated. An independent metadata calculation proves equality of row-uniform and compound-balanced donor probabilities on all 6268 retained source pools; pre-generated compound-balanced maps are not an additional evaluated sensitivity analysis. Kernel equality is not biological exchangeability.

Target transformation check. Separately from the direction-specific bank targets, a full-bank CP mid-CDF reconstruction was compared with historical normalized CP ranks. It recorded a maximum absolute discrepancy of 0.19140625 and mean absolute discrepancy 2.76×10^{-6}, with 99.9799% agreement at 10^{-6}. This discrepancy was retained, not repaired by choosing a favorable target transform.

Reading the uncertainty. The guided model has positive compound RMS, target-loss gain and PCC drop in all 40 evaluations, whereas Score and control only are exactly invariant. The guided disjoint-arm PCC advantage over control only remains positive under both reported interval methods. However, all guided disjoint-minus-shared coordinate differences include zero under the five-seed interval while their fixed-refit source-scaffold intervals are negative; the two uncertainty scopes do not establish an unchanged effect. The F5 guided-minus-control MSE contrast likewise crosses zero under the source-scaffold interval despite a negative seed interval. A positive prediction RMS

Table 8: Complete predictive means for shared and disjoint control banks at both held-out folds (five seeds per row).

Model	Fold	Dir.	Arm	**Global** PCC	**MSE**	**CP** PCC	**CP** MSE	**L1000** PCC	**L1000** MSE
Fold 4									
Score	F4	A	Shared	0.3082	0.0460	0.3331	0.0532	0.2802	0.0408
Score	F4	A	Disjoint	0.3087	0.0460	0.3334	0.0532	0.2807	0.0408
Score	F4	B	Shared	0.3079	0.0460	0.3317	0.0532	0.2810	0.0408
Score	F4	B	Disjoint	0.3084	0.0460	0.3317	0.0532	0.2818	0.0408
Control only	F4	A	Shared	0.3093	0.0460	0.3327	0.0532	0.2828	0.0408
Control only	F4	A	Disjoint	0.3086	0.0460	0.3322	0.0533	0.2818	0.0408
Control only	F4	B	Shared	0.3074	0.0460	0.3303	0.0533	0.2813	0.0408
Control only	F4	B	Disjoint	0.3077	0.0460	0.3304	0.0533	0.2818	0.0408
Guided	F4	A	Shared	0.3122	0.0459	0.3384	0.0530	0.2823	0.0408
Guided	F4	A	Disjoint	0.3115	0.0459	0.3373	0.0531	0.2821	0.0408
Guided	F4	B	Shared	0.3100	0.0459	0.3350	0.0531	0.2814	0.0408
Guided	F4	B	Disjoint	0.3105	0.0459	0.3356	0.0531	0.2817	0.0408
Fold 5									
Score	F5	A	Shared	0.3219	0.0454	0.3492	0.0536	0.2895	0.0396
Score	F5	A	Disjoint	0.3222	0.0454	0.3500	0.0535	0.2889	0.0396
Score	F5	B	Shared	0.3163	0.0455	0.3389	0.0537	0.2894	0.0396
Score	F5	B	Disjoint	0.3165	0.0455	0.3386	0.0537	0.2903	0.0396
Control only	F5	A	Shared	0.3217	0.0454	0.3480	0.0536	0.2904	0.0396
Control only	F5	A	Disjoint	0.3218	0.0455	0.3493	0.0536	0.2889	0.0396
Control only	F5	B	Shared	0.3156	0.0455	0.3373	0.0538	0.2897	0.0396
Control only	F5	B	Disjoint	0.3153	0.0456	0.3364	0.0539	0.2901	0.0396
Guided	F5	A	Shared	0.3249	0.0454	0.3536	0.0534	0.2905	0.0396
Guided	F5	A	Disjoint	0.3236	0.0454	0.3525	0.0535	0.2892	0.0396
Guided	F5	B	Shared	0.3178	0.0455	0.3416	0.0537	0.2895	0.0396
Guided	F5	B	Disjoint	0.3184	0.0455	0.3420	0.0536	0.2903	0.0396

Table 9: Complete compound-coordinate means for shared and disjoint control banks at both held-out folds (five seeds per row).

Model	Fold	Dir.	Arm	**Compound** RMS	**Compound** loss gain	**Compound** PCC drop
Fold 4						
Score	F4	A	Shared	0	0	0
Score	F4	A	Disjoint	0	0	0
Score	F4	B	Shared	0	0	0
Score	F4	B	Disjoint	0	0	0
Control only	F4	A	Shared	0	0	0
Control only	F4	A	Disjoint	0	0	0
Control only	F4	B	Shared	0	0	0
Control only	F4	B	Disjoint	0	0	0
Guided	F4	A	Shared	0.0126	1.75×10^{-4}	0.0050
Guided	F4	A	Disjoint	0.0118	1.48×10^{-4}	0.0041
Guided	F4	B	Shared	0.0118	1.55×10^{-4}	0.0045
Guided	F4	B	Disjoint	0.0118	1.54×10^{-4}	0.0044
Fold 5						
Score	F5	A	Shared	0	0	0
Score	F5	A	Disjoint	0	0	0
Score	F5	B	Shared	0	0	0
Score	F5	B	Disjoint	0	0	0
Control only	F5	A	Shared	0	0	0
Control only	F5	A	Disjoint	0	0	0
Control only	F5	B	Shared	0	0	0
Control only	F5	B	Disjoint	0	0	0
Guided	F5	A	Shared	0.0125	1.43×10^{-4}	0.0039
Guided	F5	A	Disjoint	0.0116	1.17×10^{-4}	0.0032
Guided	F5	B	Shared	0.0117	1.19×10^{-4}	0.0033
Guided	F5	B	Disjoint	0.0117	1.21×10^{-4}	0.0034

Table 10: Predictive contrasts under control separation. Seed and source-scaffold uncertainty are reported separately; the two metric blocks retain their respective scale factors.

Contrast	Fold	Mean	Seed 95% CI	Scaffold 95% CI
Global PCC ($\times 10^3$)				
Score: D−S	F4	0.427	[−0.130, 0.984]	[−0.070, 0.945]
Score: D−S	F5	0.243	[−0.583, 1.069]	[−0.158, 0.638]
Control only: D−S	F4	−0.186	[−0.769, 0.398]	[−0.794, 0.338]
Control only: D−S	F5	−0.136	[−1.679, 1.407]	[−0.687, 0.393]
Guided: D−S	F4	−0.099	[−0.972, 0.773]	[−0.442, 0.225]
Guided: D−S	F5	−0.327	[−0.963, 0.309]	[−0.640, 0.004]
Score−control, S	F4	−0.295	[−1.223, 0.633]	[−1.724, 1.066]
Score−control, S	F5	0.428	[−0.662, 1.517]	[−0.938, 1.539]
Guided−control, S	F4	2.688	[2.059, 3.317]	[1.012, 4.541]
Guided−control, S	F5	2.674	[1.609, 3.739]	[0.272, 5.761]
Score−control, D	F4	0.318	[−0.514, 1.149]	[−0.853, 1.491]
Score−control, D	F5	0.807	[−0.037, 1.650]	[−0.554, 1.931]
Guided−control, D	F4	2.775	[1.731, 3.818]	[1.264, 4.468]
Guided−control, D	F5	2.483	[1.227, 3.739]	[0.187, 5.548]
MSE ($\times 10^5$)				
Score: D−S	F4	−1.079	[−2.884, 0.726]	[−2.847, 0.536]
Score: D−S	F5	−0.564	[−3.258, 2.131]	[−1.846, 0.775]
Control only: D−S	F4	1.923	[−1.011, 4.858]	[0.232, 3.870]
Control only: D−S	F5	1.490	[−4.983, 7.964]	[−0.383, 3.538]
Guided: D−S	F4	1.047	[−2.859, 4.954]	[−0.192, 2.309]
Guided: D−S	F5	1.780	[−0.662, 4.222]	[0.555, 2.980]
Score−control, S	F4	−0.598	[−3.662, 2.465]	[−4.858, 4.209]
Score−control, S	F5	−2.943	[−7.456, 1.569]	[−6.772, 1.784]
Guided−control, S	F4	−7.236	[−11.010, −3.463]	[−13.811, −1.332]
Guided−control, S	F5	−7.329	[−14.377, −0.282]	[−18.153, 0.990]
Score−control, D	F4	−3.601	[−6.091, −1.110]	[−7.476, 0.520]
Score−control, D	F5	−4.998	[−8.488, −1.507]	[−8.982, −0.105]
Guided−control, D	F4	−8.112	[−12.543, −3.681]	[−14.199, −2.832]
Guided−control, D	F5	−7.040	[−12.443, −1.636]	[−17.585, 0.849]

D−S: disjoint minus shared wells. Within-arm contrasts subtract control only. Seed CIs use five A/B-averaged paired values (t_4). Scaffold CIs share each draw across all arms/models/directions and fix all five refits, 32 maps and the donor pool; 2000 draws, seed $2026091433 + \text{fold}$. PCC is recomputed after pooling sufficient statistics, never averaged across scaffold PCCs. These are source-scaffold, not compound-cluster, intervals; scaffold labels do not ensure fully independent biological groups.

indicates input use, not target benefit. RMS is the mean over maps of the square root of mean squared prediction distance, not the square root after averaging maps.

Table 11: Compound sensitivity and target-loss contrasts under control separation. Each metric block retains its own scale; seed and source-scaffold intervals quantify distinct uncertainty.

Contrast	Fold	Mean	Seed 95% CI	Scaffold 95% CI
Compound RMS ($\times 10^3$)				
Score: D−S	F4	0	[0, 0]	[0, 0]
Score: D−S	F5	0	[0, 0]	[0, 0]
Control only: D−S	F4	0	[0, 0]	[0, 0]
Control only: D−S	F5	0	[0, 0]	[0, 0]
Guided: D−S	F4	−0.414	[−1.567, 0.738]	[−0.495, −0.333]
Guided: D−S	F5	−0.406	[−1.563, 0.750]	[−0.490, −0.327]
Score−control, S	F4	0	[0, 0]	[0, 0]
Score−control, S	F5	0	[0, 0]	[0, 0]
Guided−control, S	F4	12.220	[11.209, 13.231]	[11.249, 13.186]
Guided−control, S	F5	12.085	[11.011, 13.158]	[11.456, 12.679]
Score−control, D	F4	0	[0, 0]	[0, 0]
Score−control, D	F5	0	[0, 0]	[0, 0]
Guided−control, D	F4	11.805	[9.826, 13.784]	[10.884, 12.725]
Guided−control, D	F5	11.678	[9.740, 13.616]	[11.106, 12.221]
Compound loss gain ($\times 10^5$)				
Score: D−S	F4	0	[0, 0]	[-2.36×10^{-12}, 2.64×10^{-12}]
Score: D−S	F5	0	[0, 0]	[-2.71×10^{-12}, 1.87×10^{-12}]
Control only: D−S	F4	0	[0, 0]	[-2.91×10^{-12}, 1.53×10^{-12}]
Control only: D−S	F5	0	[0, 0]	[-1.46×10^{-12}, 3.05×10^{-12}]
Guided: D−S	F4	−1.403	[−4.179, 1.374]	[−2.134, −0.809]
Guided: D−S	F5	−1.163	[−3.432, 1.107]	[−1.938, −0.376]
Score−control, S	F4	0	[0, 0]	[-2.43×10^{-12}, 2.50×10^{-12}]
Score−control, S	F5	0	[0, 0]	[-1.39×10^{-12}, 3.26×10^{-12}]
Guided−control, S	F4	16.509	[13.377, 19.640]	[10.136, 24.392]
Guided−control, S	F5	13.102	[10.729, 15.475]	[5.985, 20.741]
Score−control, D	F4	0	[0, 0]	[-1.32×10^{-12}, 3.12×10^{-12}]
Score−control, D	F5	0	[0, 0]	[-2.57×10^{-12}, 2.01×10^{-12}]
Guided−control, D	F4	15.106	[9.906, 20.305]	[9.267, 22.336]
Guided−control, D	F5	11.939	[7.423, 16.455]	[5.390, 18.963]

D−S: disjoint minus shared wells. Within-arm contrasts subtract control only. Seed CIs use five A/B-averaged paired values (t_4). Scaffold CIs share each draw across all arms/models/directions and fix all five refits, 32 maps and the donor pool; 2000 draws, seed $2026091433 + \text{fold}$. PCC is recomputed after pooling sufficient statistics, never averaged across scaffold PCCs. These are source-scaffold, not compound-cluster, intervals; scaffold labels do not ensure fully independent biological groups.

Table 12: Compound PCC drop contrasts ($\times 10^3$): seed and source-scaffold uncertainty are distinct.

Contrast	Fold	Mean	Seed 95% CI	Scaffold 95% CI
Score: D−S	F4	0	[0, 0]	[-1.50×10^{-13}, 1.83×10^{-13}]
Score: D−S	F5	0	[0, 0]	[-1.89×10^{-13}, 1.78×10^{-13}]
Control only: D−S	F4	0	[0, 0]	[-1.61×10^{-13}, 1.83×10^{-13}]
Control only: D−S	F5	0	[0, 0]	[-2.33×10^{-13}, 1.44×10^{-13}]
Guided: D−S	F4	−0.424	[−1.191, 0.342]	[−0.627, −0.248]
Guided: D−S	F5	−0.337	[−0.944, 0.270]	[−0.559, −0.115]
Score−control, S	F4	0	[0, 0]	[-2.05×10^{-13}, 1.11×10^{-13}]
Score−control, S	F5	0	[0, 0]	[-2.28×10^{-13}, 1.33×10^{-13}]
Guided−control, S	F4	4.708	[3.832, 5.583]	[2.949, 6.840]
Guided−control, S	F5	3.625	[2.955, 4.296]	[1.650, 5.741]
Score−control, D	F4	0	[0, 0]	[-2.22×10^{-13}, 1.33×10^{-13}]
Score−control, D	F5	0	[0, 0]	[-1.89×10^{-13}, 1.67×10^{-13}]
Guided−control, D	F4	4.283	[2.823, 5.744]	[2.673, 6.238]
Guided−control, D	F5	3.288	[2.042, 4.535]	[1.471, 5.215]

D−S: disjoint minus shared wells. Within-arm contrasts subtract control only. Seed CIs use five A/B-averaged paired values (t_4). Scaffold CIs share each draw across all arms/models/directions and fix all five refits, 32 maps and the donor pool; 2000 draws, seed $2026091433 + \text{fold}$. PCC is recomputed after pooling sufficient statistics, never averaged across scaffold PCCs. These are source-scaffold, not compound-cluster, intervals; scaffold labels do not ensure fully independent biological groups.

B EXACT PROTOCOL, DECISION RULES, AND STATISTICAL UNITS

The fixed executor owns joint inputs (c, p, a), the response target, data transformations, folds, evaluator, checkpoint rule, and training configuration; candidate code cannot alter data loading, targets, folds, or metrics. Discovery accesses Folds 1–3 only. Global PCC is one Pearson correlation after flattening all post-search rows and all dimensions of the inverse-standardized joint output. CP and L1000 PCC use the same operation within each response block. MSE is averaged over every element of the inverse-standardized joint output. At search step t, the policy maps the current model, Fold-3 diagnostics, and compact search history to a proposed revision,

$$s_{t+1} = \mathcal{D}(s_t, r_t, \mathcal{H}_t), \tag{3}$$

where edits are limited to architecture, fusion, response readout, loss, optimizer, and scheduler. The experiment log stores the proposal's parent, hypothesis, source or structured design, declared input use, execution and repair results, checkpoint, metrics, provider usage, and compute. Within each ten-slot prediction-score trajectory, the incumbent changes only under a strict improvement in Fold-3 Global PCC. After all ten trajectories finish, the first, within-policy selection chooses one source per search policy from that policy's complete executable candidate pool by maximum Fold-3 Global PCC, with lower Fold-3 MSE, fewer parameters, and stable candidate identity as tie-breakers. Neither Fold 4 nor Fold 5 enters this within-policy selection. The four policy-selected agent sources and three distinct diagnostic sources are then refit under five paired seeds and evaluated on Fold 4; the h_0 source is also displayed under its simple-concatenation role. A second, prespecified across-policy score rule chooses the agent source with the highest mean Fold-4 PCC for transfer to Fold 5. Within source selection, Fold 4 enters only this disclosed across-policy choice, while Fold 5 never enters. The path-constrained and falsification-guided studies instead carry all ten Fold-3 winners through Folds 4 and 5 without Fold-4 reselection. Before Fold 4, each study freezes the candidate set and registers the source-check rules, input-intervention construction, map seeds, threshold-construction rule, status rule, and aggregation. The realized model-specific thresholds are computed from the 32 reference scores on the fold being audited. For the prediction-score panel, Fold 5 independently refits the frozen source on Folds 1–2, selects its checkpoint on Fold 3 under the same paired-seed schedule, and evaluates Fold 5 once. Thus Fold 5 does not reuse Fold-4 weights and never changes the selected source. LINCS and LKCP subsequently test the final procedure outside BBBC development.

The compound intervention first seeks a different-fingerprint derangement within rows sharing the same observed control-profile vector. If that class derangement is impossible, a deterministic partition-wide donor with a different fingerprint is used and recorded. Only the compound representation changes: the source control profile and dose remain fixed. The resulting effect is measured under this control-stratified, dose-preserving replacement distribution. The control intervention maps each row to a different observed control-vector class; donor reuse is allowed to ensure that every control vector changes. Dose maps replace only the scalar dose, with observed-dose support checked against the source compound's released BRD identity and available cell/time annotations. None of these maps uses responses, model predictions, or evaluation scores to choose donors.

For the primary scaffold-fold compound maps, within-control-group derangements cover $411/411$ and $355/355$ rows on BBBC036, $4330/4380$ and $3831/3878$ on BBBC047, $847/847$ and $868/868$ on LINCS, and $378/405$ and $378/432$ on LKCP, respectively on Folds 4 and 5. The remaining rows use partition-wide donors. These counts specify control matching; the response contrasts retain the source dose, including when the donor compound was not observed at that dose. The plate-family analysis uses its separately specified maps in Appendix A.

For the 2×2 audit, the model-specific reference values are derived from the intervened predictions rather than from responses used to choose a permutation. For compound use they are $b^{(\mathrm{chem})}_{j,r} = \frac{1}{2}(m_{\tilde{p}c,r} + m_{\tilde{p}\tilde{c},r})$; for control-profile use, $b^{(\mathrm{context})}_{j,r} = \frac{1}{2}(m_{p\tilde{c},r} + m_{\tilde{p}\tilde{c},r})$; and for their interaction, $b^{(\mathrm{int})}_{j,r} = m_{\tilde{p}c,r} - m_{\tilde{p}\tilde{c},r}$. Dose uses the prediction score under a within-compound dose permutation. In every case, $\tau_{j,q}$ is the 95th percentile of absolute pairwise differences among the 32 fixed reference values. The lower and upper bounds used for the status are the empirical 2.5th and 97.5th percentiles of the 32 effect values; the interaction uses absolute effects. A model is qualified when the lower bound exceeds $\tau_{j,q}$, unsupported when the upper bound does not exceed $\tau_{j,q}$, and inconclusive otherwise. Inputs without enough legal interventions are reported as not identifiable. For the five-refit prediction-score panel, each refit receives its own status and an aggregate status

requires agreement in at least four of five refits; if no category reaches four of five, the aggregate is inconclusive. Path-constrained and falsification-guided results report status counts across ten selected models without a vote. The 50-refit coordinate analysis is descriptive at the refit level and uses the ten trajectory-selected endpoint instances as the policy-distribution units. These instances contain six unique configurations; repeated configurations are retained because they are independent trajectory selections. For the BBBC, LINCS, and LKCP panels, the 32 permutations are repeated measurements within a trained model, not independent sample units. The sci-Plex feedback study uses 32 development maps and 128 final held-out maps, which are likewise repeated measurements; Norman uses a separate set of 32 complete derangements (Appendix D).

The falsification-guided trainer evaluates $\alpha \in \{0.05, 0.10, 0.20, 0.35, 0.50, 0.75, 1.00\}$ on Fold 3. Predictive noninferiority allows at most a 0.003 decrease from the control-only predictor in Global PCC, a 0.005 decrease in either response block, and a 0.0005 increase in MSE. Search qualification uses eight fixed Fold-3 replacement maps. The empirical 2.5th percentile of their compound-related $\Delta\mathcal{L}$ values (increases in inverse-standardized joint-output MSE) must exceed both the 95th percentile of absolute pairwise differences among the replaced-input losses and 0.001 times the reference loss. Here reference loss is the current candidate's Fold-3 MSE with correct inputs. Dose uses the same rule when eligible, with reference loss restricted to eligible rows. These eight development maps are separate from the 32 held-out audit maps. The discovery code groups equal Morgan fingerprints for its replacement maps. Its dose gate requires at least 32 eligible rows and eight fingerprint groups in both the fit partition and Fold 3, with at least 0.1 log-dose contrast. For each candidate, epoch/scale selection maximizes Fold-3 Global PCC among qualifying states, breaking ties by lower MSE; if none qualifies, the best predictive state is retained with its qualification status recorded.

Trajectory selection is distinct from this within-candidate checkpoint rule. BBBC047 selects the highest-PCC qualifying candidate, with lower MSE as tie-breaker, or reports no qualified endpoint. BBBC036 prefers predictively noninferior candidates and ranks them by compound-related loss increase, then Global PCC, lower MSE, and stable candidate identity. If none is noninferior, that ordering retains a predictive-fallback endpoint from the executable candidates. Strict qualification is reported separately for BBBC036. These cohort-specific rules select the ten sources before either post-search boundary; held-out results never reselect them.

Held-out dose testing requires at least 64 eligible rows and coverage of at least 10% of the partition. The prediction-score and path-constrained tests accept any observed within-compound dose change; the falsification-guided test additionally requires eight compounds and an absolute log-dose difference of 0.1. Compound identity here is the BRD compound identifier, with sample/batch suffixes removed, rather than a fingerprint-equivalence class. On BBBC047, the resulting coverage is $0/4380$ rows on Fold 4 and $14/3878$ rows from seven compounds on Fold 5 (0.3610%); neither boundary meets the minimum, and neither contains a legal 0.1-log-dose contrast. Dose is therefore not identifiable in the BBBC047 panels (Table 50). Equal-fingerprint grouping alone would merge distinct released identities. On LINCS and LKCP, all replacement dose values in the 32 frozen maps are observed for the source BRD compound and its available cell/time annotations; the original input tensors and effect estimates are retained. The held-out PCC threshold is computed from fold-specific reference-score variation and is separate from the Fold-3 loss gate.

The prediction-score discovery study contains 80 ten-slot trajectories: two tasks, four search methods, and ten independent trajectory seeds. Each trajectory's budget includes the common initial predictor, and discovery intervals treat paired trajectory identifiers as the statistical unit because all policies share the same candidate-training seed schedule within an identifier. Selected-model comparisons use five paired fitting seeds on the relevant held-out fold. Path-constrained and falsification-guided studies retain all ten trajectory winners and report their distributions. All summaries are computed from fixed experiment logs that retain failed slots, repairs, and provider attempts.

For chemical-generalization contrasts, complete held-out Murcko-scaffold clusters are jointly resampled across all five paired refits, and Global PCC is recomputed from cluster sufficient statistics. This preserves within-scaffold response dependence and quantifies uncertainty over held-out chemical families. The Holm step-down adjustment applies to the three registered discovery-policy contrasts reported in Table 18; the broader frozen-predictor comparison is descriptive and reports its unadjusted paired intervals explicitly (Holm, 1979).

Table 13: Pre-specified paired comparisons of agentic model discovery. Each comparison pairs methods by the shared trajectory seed ($n = 10$ matched blocks). Differences are CELLSCIENTIST minus the listed method; CIs and two-sided p-values use paired t tests without multiplicity adjustment.

Comparator	Best@10 Δ [95% CI]; p	Frontier AUC Δ [95% CI]; p
BBBC036		
AIDE	+0.0084 [+0.0044, +0.0124] $p = 9.638 \times 10^{-4}$	+0.0054 [-0.0003, +0.0111] $p = 6.157 \times 10^{-2}$
CELLFORGE	+0.0151 [+0.0107, +0.0196] $p = 3.24 \times 10^{-5}$	+0.0106 [+0.0056, +0.0155] $p = 9.601 \times 10^{-4}$
HARMONYCELL	+0.0046 [-0.0007, +0.0100] $p = 8.064 \times 10^{-2}$	+0.0025 [-0.0016, +0.0066] $p = 1.962 \times 10^{-1}$
BBBC047		
AIDE	+0.0214 [+0.0105, +0.0324] $p = 1.636 \times 10^{-3}$	+0.0180 [+0.0078, +0.0282] $p = 3.18 \times 10^{-3}$
CELLFORGE	+0.0181 [+0.0057, +0.0304] $p = 9.051 \times 10^{-3}$	+0.0159 [+0.0046, +0.0271] $p = 1.086 \times 10^{-2}$
HARMONYCELL	+0.0028 [-0.0168, +0.0223] $p = 7.56 \times 10^{-1}$	+0.0094 [-0.0033, +0.0220] $p = 1.293 \times 10^{-1}$

Table 14: Paired Fold-4 Global-PCC contrasts for frozen sources/designs. Differences are CELLSCIENTIST minus the listed model; each uses five shared fitting seeds evaluated on Fold 4, a two-sided paired t interval, and an unadjusted paired t test.

Comparator	Δ Global PCC	95% CI	p	n
BBBC036				
HARMONYCELL	-0.0015	[-0.0037, +0.0007]	1.228×10^{-1}	5
AIDE	+0.0117	[+0.0030, +0.0205]	2.027×10^{-2}	5
CELLFORGE	+0.0158	[+0.0095, +0.0222]	2.307×10^{-3}	5
TabM	+0.0048	[+0.0016, +0.0081]	1.445×10^{-2}	5
RealMLP	+0.0118	[-0.0103, +0.0338]	2.13×10^{-1}	5
Standard MLP	+0.0193	[+0.0072, +0.0313]	1.128×10^{-2}	5
h_0	+0.0207	[+0.0130, +0.0283]	1.733×10^{-3}	5
Joint TabR	+0.0369	[+0.0343, +0.0395]	2.576×10^{-6}	5
Ridge	+0.2324	[+0.2301, +0.2347]	1.033×10^{-9}	5
BBBC047				
HARMONYCELL	+0.0012	[-0.0083, +0.0108]	7.394×10^{-1}	5
AIDE	+0.0337	[+0.0283, +0.0390]	6.271×10^{-5}	5
CELLFORGE	+0.0293	[+0.0256, +0.0330]	2.57×10^{-5}	5
TabM	+0.0181	[+0.0150, +0.0211]	7.867×10^{-5}	5
RealMLP	+0.0202	[+0.0154, +0.0250]	3.062×10^{-4}	5
Standard MLP	+0.0337	[+0.0275, +0.0398]	1.103×10^{-4}	5
h_0	+0.0337	[+0.0283, +0.0390]	6.271×10^{-5}	5
Joint TabR	+0.0500	[+0.0451, +0.0549]	9.339×10^{-6}	5
Ridge	+0.2541	[+0.2530, +0.2552]	3.077×10^{-11}	5

For the central falsification-guided study, effects are first computed from all 32 fixed within-model input permutations. The conditional interval resamples held-out Murcko scaffolds while holding the ten selected models fixed; the hierarchical sensitivity interval jointly resamples selected models and scaffolds. Models and scaffolds are the resampling units, whereas the 32 permutations remain fixed repeated measurements within a model.

The larger BBBC047 cohort retains positive hierarchical intervals for all three quantities on both held-out folds: Full−control-only PCC, compound-shuffle PCC drop, and compound target-loss gain. On BBBC036, Fold-4 compound and target-loss intervals remain positive, whereas Fold-5 intervals include zero. Scaffold resampling therefore shows which compound-family effects persist without selecting models or permutations again.

Statistical reporting and multiplicity. Trajectory-level Best@10 Global PCC is the primary discovery outcome, with frontier AUC as the co-primary search-efficiency summary. Each contrast pairs the shared trajectory seed and reports its effect estimate, two-sided paired interval, and complete ten-trajectory distribution. Holm adjustment is applied to the three pre-specified policy contrasts within each task and outcome (Table 18). Response blocks, response-sensitive metrics, cost

Table 15: Murcko-scaffold cluster bootstrap for selected-model contrasts. Each interval resamples held-out scaffold groups jointly across five paired refits.

Contrast	Evaluation	Δ Global PCC [95% CI]
BBBC036		
CellScientist − control-only	Audit	-0.0002 [-0.0020, +0.0015]
CellScientist − HarmonyCell	Audit	-0.0014 [-0.0032, +0.0003]
score-selected − h_0	Replication	+0.0344 [+0.0195, +0.0514]
BBBC047		
CellScientist − control-only	Audit	+0.0016 [+0.0001, +0.0032]
CellScientist − HarmonyCell	Audit	+0.0052 [+0.0034, +0.0069]
score-selected − h_0	Replication	+0.0366 [+0.0308, +0.0424]

Table 16: Scaffold-aware uncertainty for falsification-guided discovery. The conditional interval resamples held-out Murcko scaffolds jointly across all ten models; the hierarchical interval resamples both models and scaffolds. Compound PCC and loss effects are the context-averaged factorial contrasts E_{cmpd} and $E^{\mathcal{L}}_{\mathrm{cmpd}}$. All 32 permutations remain fixed.

Quantity	Estimate	Scaffold 95% CI	Model×scaffold 95% CI
BBBC036 / Fold 4			
Full − anchor PCC	+0.0046	[-0.0042, +0.0165]	[-0.0043, +0.0164]
Compound PCC drop	+0.0091	[+0.0001, +0.0211]	[$+1.78 \times 10^{-5}$, +0.0213]
Compound target-loss gain	+0.00039	[+0.00002, +0.00089]	[+0.00002, +0.00090]
BBBC036 / Fold 5			
Full − anchor PCC	-0.0020	[-0.0104, +0.0052]	[-0.0102, +0.0053]
Compound PCC drop	+0.0022	[-0.0050, +0.0090]	[-0.0051, +0.0092]
Compound target-loss gain	+0.00009	[-0.00022, +0.00038]	[-0.00022, +0.00039]
BBBC047 / Fold 4			
Full − anchor PCC	+0.0037	[+0.0018, +0.0060]	[+0.0018, +0.0060]
Compound PCC drop	+0.0053	[+0.0033, +0.0078]	[+0.0033, +0.0078]
Compound target-loss gain	+0.00018	[+0.00011, +0.00026]	[+0.00011, +0.00026]
BBBC047 / Fold 5			
Full − anchor PCC	+0.0028	[+0.0003, +0.0061]	[+0.0002, +0.0061]
Compound PCC drop	+0.0048	[+0.0024, +0.0079]	[+0.0023, +0.0079]
Compound target-loss gain	+0.00017	[+0.00009, +0.00026]	[+0.00009, +0.00026]

Table 17: Search-method comparison under a common executable model interface. Every method starts from the same h_0, receives the same Fold-3 feedback, evaluates ten candidates including h_0, and uses the same data, folds, evaluator, training budget, seed rule, repair limit, and model-selection rule.

Method	Information used	Revision rule	Main difference
CellScientist	Incumbent; compact four-record modeling history; diagnostic agenda.	Feedback-conditioned, source-local revision of architecture, fusion, readout, objective, optimizer, or scheduler.	Diagnostic history; incumbent-conditioned revision.
AIDE	Incumbent; compact four-record score history.	Greedy source-local code improvement under the shared model interface.	Generic improvement; no CellScientist diagnostic agenda.
CellForge	Incumbent; shared Fold-3 feedback.	One-call synthesis of modeling and critic perspectives, followed by one constrained implementation.	Committee framing; no extra proposals or shared CellScientist candidates.
HarmonyCell	Independent two-level UCB state; shared Fold-3 feedback.	Hierarchical action selection, followed by constrained leaf implementation.	Tree-directed sequencing; no CellScientist history or candidates.

records, code-check outcomes, and input-shuffle effects retain their own estimands and statistical units rather than being pooled into one omnibus claim. Input permutations are repeated transformations within a trained refit and are never treated as independent biological replicates.

Table 18: Multiplicity reporting for the pre-specified agentic-discovery comparisons. Holm adjustment is applied separately within each task and outcome across the three method contrasts. Values derive from paired t tests over ten shared trajectory seeds.

Comparator	Best@10 p_{Holm}	Frontier AUC p_{Holm}
BBBC036		
AIDE	1.9×10^{-3}	1.231×10^{-1}
CellForge	1×10^{-4}	2.9×10^{-3}
HarmonyCell	8.06×10^{-2}	1.962×10^{-1}
BBBC047		
AIDE	4.9×10^{-3}	9.5×10^{-3}
CellForge	1.81×10^{-2}	2.17×10^{-2}
HarmonyCell	7.56×10^{-1}	1.293×10^{-1}

C SEARCH TRAJECTORIES, RELIABILITY, AND COST

Table 19 summarizes the competitive discovery setting, and Tables 20 and 21 give the complete trajectory and selected-predictor comparisons. Fixed-model references include the pre-tuned RealMLP (Holzmüller et al., 2024), TabM's parameter-efficient ensemble (Gorishniy et al., 2025), and a joint-output adaptation of TabR (Gorishniy et al., 2024). Table 23 reports valid candidates, failures, repairs, success by budget, model-provider calls, tokens, GPU training time, and trajectory wall time; failed slots and provider attempts remain charged. The RBDE score provides a joint efficiency summary alongside its separate quality and cost columns.

Table 19: **Competitive discovery under a matched ten-slot budget.** Ten-trajectory means ± 95% t-interval half-widths. Best@k: best Global PCC by slot k; AUC: mean best-so-far PCC; Valid: executable-slot fraction. Per-search calls and tokens include all slots. Complete predictor and cost results: Appendix C.

Search	Best@3 ↑	Best@5 ↑	Best@10 ↑	AUC ↑	Valid ↑	Calls	Tokens (K)
BBBC036							
CELLSCIENTIST	**0.2833** ± 0.0075	**0.2856** ± 0.0052	**0.2924** ± 0.0037	**0.2845** ± 0.0046	0.990 ± 0.023	10.50 ± 1.59	51.7 ± 6.4
AIDE	0.2765 ± 0.0036	0.2804 ± 0.0033	0.2840 ± 0.0030	0.2791 ± 0.0023	1.000 ± 0.000	11.10 ± 0.98	41.8 ± 3.8
CELLFORGE	0.2708 ± 0.0016	0.2757 ± 0.0025	0.2772 ± 0.0027	0.2739 ± 0.0015	0.910 ± 0.092	16.30 ± 2.39	44.2 ± 7.9
HARMONYCELL	0.2803 ± 0.0037	0.2845 ± 0.0045	0.2877 ± 0.0038	0.2820 ± 0.0036	0.950 ± 0.051	11.80 ± 1.46	38.8 ± 4.7
BBBC047							
CELLSCIENTIST	**0.3052** ± 0.0127	**0.3084** ± 0.0127	**0.3097** ± 0.0119	**0.3058** ± 0.0111	0.940 ± 0.077	12.50 ± 3.20	59.2 ± 10.4
AIDE	0.2875 ± 0.0015	0.2882 ± 0.0014	0.2883 ± 0.0012	0.2878 ± 0.0013	0.990 ± 0.023	12.30 ± 1.17	45.2 ± 4.4
CELLFORGE	0.2891 ± 0.0015	0.2901 ± 0.0010	0.2917 ± 0.0011	0.2899 ± 0.0008	0.940 ± 0.090	12.90 ± 2.75	36.1 ± 7.4
HARMONYCELL	0.2918 ± 0.0043	0.2965 ± 0.0035	0.3070 ± 0.0083	0.2964 ± 0.0034	0.950 ± 0.051	13.20 ± 1.99	41.6 ± 5.8

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 20: **CellScientist supplies a competitive discovery frontier.** Each cell summarizes ten independent trajectories with ten evaluated-candidate slots including h_0; values are mean $\pm$ 95% t-interval half-width. AIDE, CellForge, and HarmonyCell denote the common-executor policy instantiations defined in Section 5.

Method	Best@3 PCC ↑	Best@5 ↑	Best@10 ↑	Best MSE ↓	CP PCC ↑	L1000 PCC ↑	Frontier AUC ↑	Lift AUC ↑
BBBC036								
CELLSCIENTIST	**0.2833** ± 0.0075	**0.2856** ± 0.0052	**0.2924** ± 0.0037	**0.0607** ± 0.0001	**0.2824** ± 0.0015	**0.2965** ± 0.0052	**0.2845** ± 0.0046	**0.0172** ± 0.0046
AIDE	0.2765 ± 0.0036	0.2804 ± 0.0033	0.2840 ± 0.0030	0.0612 ± 0.0001	0.2762 ± 0.0028	0.2875 ± 0.0034	0.2791 ± 0.0023	0.0118 ± 0.0028
CELLFORGE	0.2708 ± 0.0016	0.2757 ± 0.0025	0.2772 ± 0.0027	0.0617 ± 0.0003	0.2687 ± 0.0038	0.2810 ± 0.0032	0.2739 ± 0.0015	0.0066 ± 0.0020
HARMONYCELL	0.2803 ± 0.0037	0.2845 ± 0.0045	0.2877 ± 0.0038	0.0609 ± 0.0002	0.2821 ± 0.0034	0.2905 ± 0.0045	0.2820 ± 0.0036	0.0147 ± 0.0038
BBBC047								
CELLSCIENTIST	**0.3052** ± 0.0127	**0.3084** ± 0.0127	**0.3097** ± 0.0119	**0.0451** ± 0.0004	0.3351 ± 0.0048	**0.2806** ± 0.0212	**0.3058** ± 0.0111	**0.0201** ± 0.0108
AIDE	0.2875 ± 0.0015	0.2882 ± 0.0014	0.2883 ± 0.0012	0.0458 ± 0.0000	0.3269 ± 0.0028	0.2433 ± 0.0025	0.2878 ± 0.0013	0.0021 ± 0.0014
CELLFORGE	0.2891 ± 0.0015	0.2901 ± 0.0010	0.2917 ± 0.0011	0.0457 ± 0.0001	0.3315 ± 0.0040	0.2449 ± 0.0049	0.2899 ± 0.0008	0.0042 ± 0.0016
HARMONYCELL	0.2918 ± 0.0043	0.2965 ± 0.0035	0.3070 ± 0.0083	0.0452 ± 0.0003	**0.3373** ± 0.0041	0.2724 ± 0.0151	0.2964 ± 0.0034	0.0107 ± 0.0039

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 21: **Selected sources remain competitive after five-seed refitting.** All models use one joint multi-output predictor and the same input, target, partition, preprocessing, and evaluator contract; Fold-3-selected sources are evaluated on Fold 4. Values are mean ± 95% t-interval half-width.

Model	Global PCC ↑	MSE ↓	CP PCC ↑	L1000 PCC ↑	Parameters	Train (s)	Selected epoch
BBBC036							
CELLSCIENTIST	<u>0.3101</u> ± 0.0023	**0.0586** ± 0.0001	0.3107 ± 0.0025	<u>0.3098</u> ± 0.0032	4.00M	7.2 ± 1.2	21.8 ± 17.0
AIDE	0.2984 ± 0.0091	0.0593 ± 0.0006	0.3085 ± 0.0043	0.2950 ± 0.0124	4.42M	6.3 ± 1.2	4.8 ± 1.0
CELLFORGE	0.2942 ± 0.0043	0.0593 ± 0.0002	0.3052 ± 0.0048	0.2902 ± 0.0056	3.82M	6.2 ± 1.3	2.0 ± 0.0
HARMONYCELL	**0.3116** ± 0.0013	**0.0586** ± 0.0001	0.3117 ± 0.0024	**0.3115** ± 0.0020	2.06M	6.7 ± 1.4	17.2 ± 12.5
h_0	0.2894 ± 0.0081	0.0597 ± 0.0003	0.2984 ± 0.0048	0.2860 ± 0.0098	1.14M	5.9 ± 1.2	3.2 ± 0.6
Standard MLP	0.2908 ± 0.0114	0.0594 ± 0.0004	0.3058 ± 0.0089	0.2845 ± 0.0125	2.42M	5.8 ± 1.2	1.8 ± 0.6
RealMLP	0.2983 ± 0.0202	0.0592 ± 0.0007	<u>0.3170</u> ± 0.0129	0.2902 ± 0.0260	3.47M	5.9 ± 0.7	–
TabM	0.3053 ± 0.0039	<u>0.0589</u> ± 0.0001	**0.3175** ± 0.0063	0.3004 ± 0.0044	2.37M	11.4 ± 1.7	2.8 ± 1.0
Joint TabR	0.2732 ± 0.0033	0.0604 ± 0.0001	0.2835 ± 0.0036	0.2687 ± 0.0044	2.17M	2.6 ± 0.3	7.6 ± 1.1
Ridge	0.0777 ± 0.0000	0.2048 ± 0.0000	0.0812 ± 0.0000	0.0763 ± 0.0000	4.14M	1.5 ± 0.4	–
BBBC047							
CELLSCIENTIST	**0.3033** ± 0.0011	**0.0457** ± 0.0000	0.3254 ± 0.0010	**0.2787** ± 0.0015	1.78M	24.3 ± 3.7	36.6 ± 10.5
AIDE	0.2696 ± 0.0047	0.0468 ± 0.0002	0.3167 ± 0.0062	0.2125 ± 0.0101	1.20M	9.5 ± 0.7	2.0 ± 0.9
CELLFORGE	0.2740 ± 0.0035	0.0467 ± 0.0001	0.3178 ± 0.0061	0.2216 ± 0.0060	2.53M	9.6 ± 0.8	1.8 ± 1.0
HARMONYCELL	<u>0.3021</u> ± 0.0098	<u>0.0458</u> ± 0.0003	<u>0.3265</u> ± 0.0016	<u>0.2743</u> ± 0.0196	2.28M	26.0 ± 9.2	41.8 ± 27.1
h_0	0.2696 ± 0.0047	0.0468 ± 0.0002	0.3167 ± 0.0062	0.2125 ± 0.0101	1.20M	9.3 ± 0.9	2.0 ± 0.9
Standard MLP	0.2696 ± 0.0059	0.0468 ± 0.0002	0.3167 ± 0.0062	0.2127 ± 0.0067	2.53M	9.6 ± 0.9	1.2 ± 0.6
RealMLP	0.2831 ± 0.0042	0.0463 ± 0.0001	0.3259 ± 0.0038	0.2326 ± 0.0049	3.62M	13.7 ± 0.8	–
TabM	0.2852 ± 0.0033	0.0463 ± 0.0001	**0.3282** ± 0.0023	0.2350 ± 0.0105	2.51M	16.7 ± 2.1	2.6 ± 0.7
Joint TabR	0.2533 ± 0.0049	0.0475 ± 0.0002	0.2951 ± 0.0050	0.2027 ± 0.0066	2.25M	10.8 ± 1.1	4.6 ± 0.7
Ridge	0.0492 ± 0.0000	0.3880 ± 0.0000	0.0547 ± 0.0000	0.0438 ± 0.0000	4.62M	1.3 ± 0.2	–

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 22: Model-discovery configuration. All agentic discovery calls covered by this configuration use the recorded DeepSeek V4 Flash model identifier through an OpenAI-compatible chat-completions interface. No fallback is enabled. Each request explicitly sets the model, messages, temperature, and maximum output length; other sampling fields use provider defaults. For reproducibility, prompt-construction code, search history, request/response SHA-256 hashes, returned code or design, response model, and token usage are logged for every call.

Formulation	Generated artifact	Max output	Slots
Prediction-score policies	Executable source	9000	10
Path-constrained	Structured design card	2400	10
Falsification-guided	Semantic design card	1800	10

Shared settings: model `deepseek-v4-flash`; temperature 0.2; reasoning effort not set; fallback disabled. Max output is the token limit per request.

Table 23: Complete agentic-discovery execution and cost receipts. Values are mean ± 95% t-interval half-width across ten trajectories. Success@k counts trajectories (out of ten) whose best Fold3 Global PCC by candidate slot k strictly exceeds that of the initial h_0; the candidate-slot budget includes h_0. RBDE is the registered valid-candidate rate times the non-negative 95% lower confidence bound of Frontier-Lift AUC per ten minutes of total trajectory wall time. Tokens are thousands and exclude only trajectories without a complete provider-usage receipt; all other fields have $n = 10$.

A. Execution quality and budget efficiency

Method	RBDE ↑	Valid rate ↑	Failed slots ↓	Repairs ↓	Success@3/5/10 ↑
BBBC036					
CELLSCIENTIST	0.0104	0.990 ± 0.023	0.10 ± 0.23	1.50 ± 1.59	8/10/10
AIDE	0.0096	1.000 ± 0.000	0.00 ± 0.00	2.10 ± 0.98	9/10/10
CELLFORGE	0.0044	0.910 ± 0.092	0.90 ± 0.92	7.30 ± 2.39	8/10/10
HARMONYCELL	0.0102	0.950 ± 0.051	0.50 ± 0.51	2.80 ± 1.46	9/10/10
BBBC047					
CELLSCIENTIST	0.0064	0.940 ± 0.077	0.60 ± 0.77	3.50 ± 3.20	7/8/9
AIDE	0.0008	0.990 ± 0.023	0.10 ± 0.23	3.30 ± 1.17	5/8/9
CELLFORGE	0.0027	0.940 ± 0.090	0.60 ± 0.90	3.90 ± 2.75	8/9/9
HARMONYCELL	0.0057	0.950 ± 0.051	0.50 ± 0.51	4.20 ± 1.99	8/10/10

B. Language-model and compute costs

Method	Calls	Tokens (K)	GPU training (s)	Wall time (s)
BBBC036				
CELLSCIENTIST	10.50 ± 1.59	51.7 ± 6.4	19.8 ± 2.2	716.3 ± 101.3
AIDE	11.10 ± 0.98	41.8 ± 3.8	16.6 ± 2.5	559.0 ± 212.1
CELLFORGE	16.30 ± 2.39	44.2 ± 7.9	15.7 ± 1.4	578.0 ± 86.4
HARMONYCELL	11.80 ± 1.46	38.8 ± 4.7	16.8 ± 1.1	610.8 ± 158.2
BBBC047				
CELLSCIENTIST	12.50 ± 3.20	59.2 ± 10.4	111.7 ± 21.3	821.9 ± 116.0
AIDE	12.30 ± 1.17	45.2 ± 4.4	60.3 ± 3.9	533.0 ± 152.3
CELLFORGE	12.90 ± 2.75	36.1 ± 7.4	56.7 ± 5.8	533.0 ± 83.6
HARMONYCELL	13.20 ± 1.99	41.6 ± 5.8	72.9 ± 11.0	680.5 ± 228.0

Table 24: Diagnostic response-block metrics for the frozen joint predictors on Fold 4. Global PCC is the registered joint selection endpoint; CP and L1000 metrics localize response behavior by readout block. Top-K response dimensions are registered from Fold-1–2 response magnitude. Values are mean ± 95% t-interval half-width over five paired seeds.

	CP response block				L1000 response block			
Model	**PCC @20** ↑	**PCC @50** ↑	**RMSE @20** ↓	**RMSE @50** ↓	**PCC @20** ↑	**PCC @50** ↑	**RMSE @20** ↓	**RMSE @50** ↓
BBBC036								
CELLSCIENTIST	0.4620 ± 0.0035	0.4454 ± 0.0033	0.2411 ± 0.0005	0.2385 ± 0.0004	0.2828 ± 0.0051	0.3167 ± 0.0043	0.5135 ± 0.0008	0.4497 ± 0.0007
AIDE	0.4636 ± 0.0024	0.4445 ± 0.0025	0.2408 ± 0.0006	0.2386 ± 0.0004	0.2611 ± 0.0200	0.3005 ± 0.0160	0.5190 ± 0.0054	0.4540 ± 0.0048
CELLFORGE	0.4607 ± 0.0050	0.4430 ± 0.0034	0.2412 ± 0.0006	0.2388 ± 0.0005	0.2598 ± 0.0081	0.2966 ± 0.0099	0.5176 ± 0.0011	0.4531 ± 0.0013
HARMONYCELL	0.4633 ± 0.0023	0.4464 ± 0.0024	0.2409 ± 0.0005	0.2383 ± 0.0004	0.2823 ± 0.0040	0.3170 ± 0.0031	0.5137 ± 0.0006	0.4497 ± 0.0004
h_0	0.4526 ± 0.0034	0.4357 ± 0.0032	0.2426 ± 0.0005	0.2399 ± 0.0005	0.2555 ± 0.0079	0.2983 ± 0.0034	0.5210 ± 0.0056	0.4545 ± 0.0031
Standard MLP	0.4608 ± 0.0077	0.4404 ± 0.0070	0.2412 ± 0.0011	0.2391 ± 0.0009	0.2473 ± 0.0048	0.2923 ± 0.0111	0.5197 ± 0.0006	0.4536 ± 0.0017
RealMLP	0.4668 ± 0.0081	0.4497 ± 0.0080	0.2402 ± 0.0013	0.2378 ± 0.0011	0.2476 ± 0.0332	0.2945 ± 0.0320	0.5200 ± 0.0053	0.4540 ± 0.0047
TabM	0.4640 ± 0.0023	0.4477 ± 0.0021	0.2408 ± 0.0005	0.2382 ± 0.0003	0.2603 ± 0.0079	0.3066 ± 0.0033	0.5180 ± 0.0031	0.4518 ± 0.0016
Joint TabR	0.4433 ± 0.0041	0.4235 ± 0.0045	0.2437 ± 0.0005	0.2414 ± 0.0006	0.2389 ± 0.0124	0.2812 ± 0.0120	0.5231 ± 0.0024	0.4567 ± 0.0020
Ridge	0.1172 ± 0.0000	0.1153 ± 0.0000	0.4678 ± 0.0000	0.4603 ± 0.0000	0.0807 ± 0.0000	0.0903 ± 0.0000	0.9852 ± 0.0000	0.8401 ± 0.0000
BBBC047								
CELLSCIENTIST	0.4003 ± 0.0028	0.4061 ± 0.0018	0.2689 ± 0.0002	0.2547 ± 0.0002	0.3262 ± 0.0025	0.3154 ± 0.0019	0.3703 ± 0.0004	0.3250 ± 0.0003
AIDE	0.4127 ± 0.0124	0.4078 ± 0.0083	0.2676 ± 0.0022	0.2549 ± 0.0015	0.2577 ± 0.0090	0.2486 ± 0.0060	0.3789 ± 0.0009	0.3323 ± 0.0006
CELLFORGE	0.4117 ± 0.0095	0.4069 ± 0.0032	0.2678 ± 0.0010	0.2550 ± 0.0007	0.2684 ± 0.0051	0.2583 ± 0.0041	0.3776 ± 0.0009	0.3312 ± 0.0007
HARMONYCELL	0.4029 ± 0.0049	0.4079 ± 0.0026	0.2686 ± 0.0009	0.2545 ± 0.0004	0.3234 ± 0.0156	0.3130 ± 0.0148	0.3705 ± 0.0021	0.3252 ± 0.0016
h_0	0.4127 ± 0.0124	0.4078 ± 0.0083	0.2676 ± 0.0022	0.2549 ± 0.0015	0.2577 ± 0.0090	0.2486 ± 0.0060	0.3789 ± 0.0009	0.3323 ± 0.0006
Standard MLP	0.4125 ± 0.0157	0.4083 ± 0.0089	0.2678 ± 0.0023	0.2548 ± 0.0013	0.2626 ± 0.0135	0.2531 ± 0.0123	0.3782 ± 0.0018	0.3316 ± 0.0015
RealMLP	0.4142 ± 0.0072	0.4137 ± 0.0043	0.2667 ± 0.0010	0.2536 ± 0.0005	0.2883 ± 0.0034	0.2798 ± 0.0028	0.3750 ± 0.0004	0.3287 ± 0.0003
TabM	0.4152 ± 0.0107	0.4142 ± 0.0050	0.2666 ± 0.0015	0.2536 ± 0.0007	0.2784 ± 0.0051	0.2685 ± 0.0067	0.3763 ± 0.0010	0.3300 ± 0.0008
Joint TabR	0.3751 ± 0.0078	0.3810 ± 0.0061	0.2757 ± 0.0007	0.2599 ± 0.0006	0.2482 ± 0.0101	0.2409 ± 0.0084	0.3806 ± 0.0012	0.3336 ± 0.0009
Ridge	0.0833 ± 0.0000	0.0808 ± 0.0000	0.7616 ± 0.0000	0.7251 ± 0.0000	0.0583 ± 0.0000	0.0583 ± 0.0000	1.1990 ± 0.0000	1.0275 ± 0.0000

Table 25: **Deterministic compilation eliminates candidate-code execution failures in the evaluated path-constrained runs.** PCC and resource values are trajectory-level means ± 95% interval half-width. Valid covers all candidate slots, repairs are candidate-code repair attempts per trajectory, and Linked is the number of candidates connected to the declared input in code. Provider retries remain included in calls and tokens.

Task	**Score** PCC	**Path** PCC	Δ	**Tokens** (K)	Valid	Repairs	Linked
BBBC036	0.2924 ± 0.0037	0.2925 ± 0.0019	+0.0001	21.9 ± 1.4	100.0%	0.00 ± 0.00	100/100
BBBC047	0.3097 ± 0.0119	0.3008 ± 0.0012	-0.0090	21.7 ± 1.2	100.0%	0.00 ± 0.00	100/100

D INPUT-USE CLAIMS IN COMBINATORIAL GENETIC-PERTURBATION PREDICTION

Table 26: **Registered pair-input dependence on unseen genetic combinations.** Five refits per predictor; all pass the input-use test. Anchored predictors replace the pair-input bundle; GEARS uses its native pair-identity replacement. Drops are reported without cross-interface ranking. Complete comparisons: Appendix D.

Boundary	Predictor	Global PCC ↑	Input ΔPCC ↑	Support
Fold 4 audit	GEARS	0.6819 [0.6714, 0.6924]	0.3029	5/5
	Initial model (h_0)	0.8960 [0.8909, 0.9011]	0.5425	5/5
	CELLSCIENTIST	0.9020 [0.8988, 0.9051]	0.5578	5/5
Fold 5 replication	GEARS	0.6676 [0.6568, 0.6784]	0.3599	5/5
	Initial model (h_0)	0.9051 [0.9023, 0.9079]	0.6193	5/5
	CELLSCIENTIST	0.9096 [0.9051, 0.9142]	0.6280	5/5

The transfer task is reconstructed from the count layer of the Norman K562 CRISPRa Perturb-seq experiment (GEO GSE133344) (Norman et al., 2019). Its source matrix contains 91,205 cells and 5,045 genes, including 7,353 control cells. Per-cell counts are library-size normalized to 10,000 and log-transformed, after which canonical condition means are centered by the control mean. The 284 source labels resolve to 237 canonical conditions: 105 observed singles, 131 doubles, and one control. The common discovery interface receives a pair-input bundle containing a 105-dimensional multi-hot perturbation identity and its matching mean single-perturbation anchor, and returns the full 5,045-gene response without external pathway features or cell-type covariates. GEARS retains its task-native graph inputs as described below.

The five-fold assignment is constructed without expression responses and balances pair count, source-cell count, and perturbation-gene incidence. All singles and controls remain fit-only with the Fold-1/2 doubles; Fold 3 supplies discovery feedback, and Folds 4 and 5 are successive held-out evaluation folds. Table 27 lists every held-out double-perturbation block. The comparison predicts unseen double combinations whose constituent single-gene responses are available during fitting.

Each held-out boundary contains 26 unseen double-perturbation conditions. Before Fold 4 is opened, we construct 32 distinct complete derangements of the 26 row positions; each is a one-to-one permutation with no fixed point and is reused positionally on Fold 5. For each source condition, the registered intervention replaces both coordinates of the pair-input bundle (its 105-dimensional pair-identity vector and matching mean single-perturbation anchor) with the donor pair's values while retaining the source response target. Replacing both coordinates prevents a residual predictor from retaining pair information through the anchor; consequently, the test measures dependence on the complete bundle rather than attributing the effect to the multi-hot coordinate alone. The maps use neither responses nor model outputs and do not require donor pairs to be gene-disjoint from source pairs.

For refit j and map r, let L_j be joint-output MSE under the correct pair-input bundle and $L_{j,r}^{\mathrm{shuf}}$ the MSE after the bundle shuffle. We define $g_{j,r} = L_{j,r}^{\mathrm{shuf}} - L_j$ and $\tilde{g}_{j,r} = g_{j,r} / \max(L_j, 10^{-12})$. A refit is qualified under the pair-input-use criterion when $32^{-1}\sum_r \tilde{g}_{j,r} > 0.001$ and $Q_{0.05}(\{g_{j,r}\}_{r=1}^{32}) > 0$. It is unsupported when the first quantity lies within $[-0.001, 0.001]$ and inconclusive otherwise. The Global-PCC drop is reported as a continuous effect but does not enter this categorical rule. The 32 maps are repeated interventions within a refit; the five refits are the statistical units for model-performance intervals and support counts.

GEARS is evaluated through the shared task boundary while retaining its original multigene-perturbation formulation (Roohani et al., 2024): GO and coexpression graphs, native GNN/decoder, optimization objective, and validation-selected checkpoint. It shares count normalization, response target, fold roles, and held-out metrics with the other predictors. Table 28 records the architecture, graph scope, optimizer, and checkpoint rule used in the comparison.

Across ten trajectories, all 77 nonduplicate CellScientist candidates pass source, shape, and training checks; the 23 duplicate design cards still consume their candidate slots, and no Fold-4 or Fold-5 response enters proposal or selection. The trajectories account for 154 provider calls, 403,762 tokens, 17.8 H100 fit-minutes, and 50.5 minutes of end-to-end time. Held-out input-use testing is deterministic and uses zero model-provider calls.

Table 27: **Norman combinatorial-perturbation fold inventory.** Source cells are counted before condition-level aggregation. All 105 observed single perturbations and the control condition remain fit-only; the table reports the held-out double-perturbation partition.

Fold	Role	**Double** perturbations	**Source** cells	**Component** genes
1	Fit	26	7,301	40
2	Fit	26	7,849	41
3	Discovery/selection	27	6,817	42
4	Independent audit	26	6,541	40
5	Final replication	26	6,937	38

Table 28: **GEARS alignment for the Norman transfer.** The original GEARS architecture and graph construction are retained, while normalization, response target, data-split roles, and held-out evaluator match the common 5,045-gene task.

Protocol item	Setting
Input and output	105 observed perturbation genes; one joint 5,045-gene response
Partition roles	Folds 1–2 fit; Fold 3 selects checkpoints; Folds 4/5 are audit/replication
Development graph	Fit and Fold-3 conditions only; 102/105 genes GO-annotated; three isolated genes receive self-loops; 2,043 GO edges
Architecture	Hidden size 64; one GO GNN layer; one gene GNN layer; decoder hidden size 16; 20 GO and 20 coexpression neighbors; coexpression threshold 0.4
Optimization	20 epochs; batch 128; Adam, learning rate 10^{-3}, weight decay 5×10^{-4}; StepLR$(1, 0.5)$; five seeds (41–45)
Checkpoint rule	Minimum Fold-3 DEG-20 validation MSE; frozen before either held-out boundary is opened
Input-use test	32 complete pair-identity derangements fixed without response values; pass requires mean relative MSE increase > 0.001 and positive 5th-percentile raw MSE increase

Table 29: **Search statistics for the Norman transfer.** Results use ten independent trajectories; the candidate budget includes h_0.

A. Search performance

B	Best@3	Best@5	Best@10	Frontier AUC	Δh_0 [95% CI]
10	0.9192	0.9201	0.9210	0.9194	+0.0062 [+0.0036, +0.0087]

B. Completion and resources

Positive	Executed	Calls	Tokens
10/10	77/100	154	403,762

The matched transfer analysis gives CellScientist, AIDE, CellForge, and HarmonyCell the same initial model, design language, Fold-3 feedback, physical budget, and model-selection rule; each policy generates its own candidates. CellScientist has the highest mean Best@10 and frontier AUC. Its paired Best@10 effect is $+0.0029$ versus AIDE (95% CI $[-0.0003, +0.0061]$), $+0.0058$ versus CellForge ($[+0.0037, +0.0080]$), and $+0.0006$ versus HarmonyCell ($[-0.0005, +0.0016]$). After selection, all sources are refit under five paired seeds and audited without provider calls. Global-PCC ordering varies across folds, while all four selected models pass the pair-input-bundle test in 5/5 refits; CellScientist leads condition-, DEG-20-, and DEG-50-PCC at Fold 5.

Table 30: **Complete held-out Norman response metrics.** Neural entries summarize five refits; deterministic references are evaluated once. All predictors return the same 5,045-gene joint response. Input-use support is the number of refits passing the predictor's registered perturbation-pair input test; deterministic references report their single result. Predictors with an observed-single anchor are tested by replacing the complete pair-input bundle, while GEARS retains its task-native pair-identity intervention.

Predictor	Global PCC ↑	MSE ↓	Condition PCC ↑	Rank corr. ↑	DEG-20 PCC ↑	DEG-50 PCC ↑	Full-state PCC ↑	Input-use support
Fold 4 audit								
Matching mean	0.8810	0.0059	0.8966	0.4902	0.9702	0.9634	0.9915	supported
Additive singles	0.8810	0.0046	0.8966	0.4902	0.9702	0.9634	0.9932	supported
Ridge	0.8989	0.0033	0.8965	0.4584	0.9724	0.9665	0.9951	supported
GEARS	0.6819	0.0122	0.6826	0.2512	0.9078	0.8915	0.9843	5/5
h_0	0.8960	0.0037	0.8999	0.4288	0.9664	0.9618	0.9945	5/5
CellScientist	0.9020	0.0039	0.9098	0.4345	0.9750	0.9697	0.9943	5/5
Fold 5 replication								
Matching mean	0.8982	0.0043	0.8799	0.4681	0.9614	0.9469	0.9938	supported
Additive singles	0.8982	0.0033	0.8799	0.4681	0.9614	0.9469	0.9952	supported
Ridge	0.8959	0.0029	0.8813	0.4375	0.9672	0.9546	0.9957	supported
GEARS	0.6676	0.0115	0.6156	0.2267	0.8021	0.8096	0.9853	5/5
h_0	0.9051	0.0027	0.8844	0.4065	0.9542	0.9474	0.9961	5/5
CellScientist	0.9096	0.0028	0.8944	0.4071	0.9650	0.9554	0.9960	5/5

Table 31: **Matched agentic model discovery on the Norman transfer.** Each method receives the same initial model, trainer, Fold-3 feedback, ten trajectories, and ten candidate evaluations. AIDE, CellForge, and HarmonyCell are implemented within the same runnable model space. Arrows mark the preferred direction. Executed slots count candidate proposals that trained; duplicate design cards consume slots but are not executable failures.

Policy	Best@3 ↑	Best@5 ↑	Best@10 ↑	Frontier AUC ↑	Δh_0 [95% CI] ↑	Positive ↑	Executed slots ↑	Calls ↓	Tokens ↓
CellScientist	**0.9192**	<u>0.9201</u>	**0.9210**	**0.9194**	**+0.0062** [+0.0036,+0.0087]	10/10	77/100	154	403,762
AIDE	0.9164	0.9168	0.9181	0.9170	+0.0033 [+0.0011,+0.0054]	9/10	91/100	120	335,610
CellForge	0.9151	0.9152	0.9152	0.9151	+0.0003 [−0.0004,+0.0010]	1/10	97/100	108	309,742
HarmonyCell	<u>0.9191</u>	**0.9203**	<u>0.9204</u>	<u>0.9193</u>	<u>+0.0056</u> [+0.0028,+0.0084]	9/10	88/100	117	328,705

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 32: **Models selected on Fold 3 before audit.** One model per search method is chosen by the common rule: highest Fold-3 Global PCC, then lower MSE and fewer parameters.

A. Selected architecture

Policy	Architecture	Width	Depth	Dropout	PCC–MSE weight
CellScientist	Gated MLP	512	4	0.1	0.3
AIDE	Gated MLP	512	4	0.2	0.2
CellForge	Residual MLP	256	3	0.1	0.1
HarmonyCell	Residual MLP	384	4	0.2	0.1

B. Optimization and selection

Policy	Optimizer / learning rate	Parameters	Fold-3 PCC
CellScientist	AdamW / 0.001	5,850,037	0.9237
AIDE	AdamW / 0.0005	5,850,037	0.9234
CellForge	AdamW / 0.001	1,850,549	0.9195
HarmonyCell	AdamW / 0.001	3,758,261	0.9247

Table 33: **Matched input-use test and final replication of selected models.** Neural models are refit with five paired seeds. Brackets are 95% intervals over those seeds; bundle drop averages 32 fixed interventions that jointly replace the pair identity and matching observed-single anchor. All sources are selected before Fold 4 and are refit at Fold 5 without source reselection.

Policy	**Global PCC ↑** [95% CI]	MSE ↓	**Condition** PCC ↑	**DEG-20** PCC ↑	**DEG-50** PCC ↑	**Bundle** drop	**Bundle-use** support
Fold 4 audit							
CellScientist	**0.9020** [0.8988,0.9051]	0.00388	**0.9098**	**0.9750**	**0.9697**	0.5578	5/5
AIDE	0.8981 [0.8920,0.9042]	0.00407	0.9038	0.9682	0.9637	0.5458	5/5
CellForge	0.8960 [0.8909,0.9011]	0.00372	0.8999	0.9664	0.9618	0.5425	5/5
HarmonyCell	0.8960 [0.8920,0.9000]	**0.00369**	0.9019	0.9693	0.9641	0.5506	5/5
Fold 5 replication							
CellScientist	0.9096 [0.9051,0.9142]	0.00277	**0.8944**	**0.9650**	**0.9554**	0.6280	5/5
AIDE	**0.9125** [0.9088,0.9161]	0.00285	0.8912	0.9567	0.9488	0.6270	5/5
CellForge	0.9051 [0.9023,0.9079]	0.00267	0.8844	0.9542	0.9474	0.6193	5/5
HarmonyCell	0.9088 [0.9029,0.9148]	**0.00255**	0.8867	0.9597	0.9505	0.6294	5/5

Bold/underline: best/second distinct displayed predictive mean within each task or fold. Ties share a rank; interval bounds and diagnostic/resource columns are unranked.

Table 34: **Paired CellScientist contrasts in the Norman held-out study.** Effects are CellScientist minus the comparator over five paired fitting seeds. Global-PCC wins count seeds with a positive effect.

Comparator	**Δ Global PCC** [95% CI]	PCC wins	**Δ bundle drop** [95% CI]
Fold 4 audit			
AIDE	+0.0039 [−0.0044,+0.0122]	4/5	+0.0121 [−0.0004,+0.0245]
CellForge	+0.0060 [−0.0008,+0.0128]	4/5	+0.0154 [+0.0030,+0.0277]
HarmonyCell	+0.0060 [−0.0009,+0.0129]	4/5	+0.0072 [−0.0061,+0.0205]
Fold 5 replication			
AIDE	−0.0028 [−0.0103,+0.0046]	2/5	+0.0010 [−0.0059,+0.0078]
CellForge	+0.0045 [−0.0004,+0.0094]	5/5	+0.0087 [+0.0037,+0.0137]
HarmonyCell	+0.0008 [−0.0056,+0.0072]	4/5	−0.0015 [−0.0131,+0.0101]

E INDEPENDENT-ACQUISITION TEST OF SMALL-MOLECULE CLAIM GENERALIZATION

After the final falsification-guided procedure is fixed, LINCS–Pilot1 provides an out-of-development-cohort test. It uses an observed control profile, compound identity, and dose to predict a joint 2,719-dimensional morphology–transcription response across 4,155 matched conditions and 1,108 Murcko groups. The ten trajectory-selected design instances are then fixed for transfer to `cpg0004/LKCP/2017_12_05_Batch2`, an independently acquired Cell Painting release (Weisbart et al., 2024). Its 134 files contain 11,904 rows; 8,958 treatment and 2,946 control rows yield 1,728 retained conditions and 64 compound-held-out groups. Response- and fold-blind interface quality control removes 110 non-finite or extreme release features from 1,781 common morphology columns, retaining a 1,671-dimensional same-plate control profile and response.

The scientific question and model designs precede LKCP transfer. The final dataset adapter is fixed after interface quality control and before LKCP training or held-out-fold access. The transferred designs, input-permutation construction, map seeds, threshold-construction rule, and decision rule

are registered and hash-bound at that point. Each design is then refit under five paired seeds with no further source edit or model reselection. Compound permutations hold dose fixed; dose permutations use only observed doses of the same compound, require at least a 0.1 log-dose contrast, and retain the same eligibility rule. Across 2 cohorts $\times$ 2 folds $\times$ 10 designs $\times$ 5 seeds, the analysis contains 200 cohort–fold–design–seed refit boundaries. Each boundary reuses the 32 registered ordered compound–dose permutation pairs, yielding 6,400 matched within-refit permutation units; these maps are repeated measurements rather than independent statistical units. For U_{11} with both inputs correct, U_{01} with compound shuffled, U_{10} with dose shuffled, and U_{00} with both shuffled, the two-player allocation is $\phi_{\mathrm{chem}} = \frac{1}{2}[(U_{11}-U_{01})+(U_{10}-U_{00})]$ and $\phi_{\mathrm{dose}} = \frac{1}{2}[(U_{11}-U_{10})+(U_{01}-U_{00})]$. These allocate the contrast $U_{11} - U_{00}$; adding the residual $U_{00} - U_{\mathrm{anchor}}$ gives Full−control-only to numerical precision. The component allocations thus refer to the fixed replacement baseline, while Full−control-only measures the separately fitted predictors' performance difference.

Table 35: **Compound identity and dose yield distinct replacement effects.** LINCS evaluates the final small-molecule procedure, and LKCP transfers the ten trajectory-selected LINCS design instances to an independent acquisition. These instances contain six unique configurations and retain repeated selections as part of the trajectory-level policy distribution. Compound identity and dose are two-player Shapley allocations of the Global-PCC change from both inputs replaced to both correct. Adding the both-shuffled residual gives Full−control-only, the separate full-model versus control-only comparison. The interaction is reported separately because it is already allocated across the two Shapley terms. Intervals use the ten endpoint instances after averaging five paired refits within each instance.

A. LINCS–Pilot1

Contrast	Fold 4 [95% CI]	Fold 5 [95% CI]
Full−control-only	+0.0293 [+0.0281, +0.0305]	+0.0233 [+0.0226, +0.0241]
Compound identity	+0.0175 [+0.0164, +0.0186]	+0.0126 [+0.0117, +0.0134]
Dose	+0.0486 [+0.0473, +0.0498]	+0.0467 [+0.0456, +0.0478]
Both-shuffled residual	-0.0368 [-0.0378, -0.0358]	-0.0359 [-0.0370, -0.0349]
Interaction	+0.0050 [+0.0033, +0.0067]	+0.0042 [+0.0027, +0.0058]

B. LKCP Batch 2

Contrast	Fold 4 [95% CI]	Fold 5 [95% CI]
Full−control-only	+0.0104 [+0.0103, +0.0106]	+0.0211 [+0.0205, +0.0217]
Compound identity	-0.00007000 [-0.00028998, +0.00014998]	+0.0033 [+0.0027, +0.0039]
Dose	+0.0237 [+0.0228, +0.0247]	+0.0341 [+0.0331, +0.0350]
Both-shuffled residual	-0.0132 [-0.0140, -0.0124]	-0.0163 [-0.0170, -0.0157]
Interaction	+0.00004049 [-0.00001503, +0.00009601]	+0.0002 [+0.0001, +0.0003]

Table 36: **Dose exceeds the reference-relative criterion consistently; compound status varies by cohort and fold.** The criterion compares each refit's lower effect quantile with its fold-specific reference threshold. Counts summarize stability relative to that reference scale, separately from the sign of the mean contribution. The mean-effect column is the mean of the 32 map-level coordinate effects used in this threshold test and is distinct from the two-player Shapley allocations in the preceding table. The 50 refits at each boundary are nested within ten trajectory-selected endpoint instances; the endpoint instance is the unit for continuous-effect intervals.

Cohort	Fold	Mean effect	Lower-effect quantile	Threshold	Reference margin	Refits above
Chemical identity						
LINCS–Pilot1	4	0.0200	0.0139	0.0112	+0.0028	35/50
LINCS–Pilot1	5	0.0147	0.0090	0.0100	-0.0010	10/50
LKCP Batch 2	4	0.00003872	-0.0016	0.0303	-0.0319	0/50
LKCP Batch 2	5	0.0032	0.0013	0.0307	-0.0294	0/50
Dose						
LINCS–Pilot1	4	0.0510	0.0467	0.0068	+0.0398	50/50
LINCS–Pilot1	5	0.0488	0.0452	0.0059	+0.0393	50/50
LKCP Batch 2	4	0.0238	0.0205	0.0060	+0.0145	50/50
LKCP Batch 2	5	0.0342	0.0267	0.0125	+0.0142	50/50

F FULL SELECTED-MODEL AUDIT PANEL AND CALIBRATION

The full Fold-4 panel contains the Fold-3-selected source from each of the four agents together with three distinct diagnostic sources: control-only, compound-only, and the shared h_0/simple-concatenation source shown under both reference roles. Predictive estimates use a separate five-seed refit panel from Table 21; model source or design, data fingerprint, population, and every non-seed fitting setting are identical. Each within-policy source is fixed before Fold 4; the pre-specified across-policy score rule then uses mean Fold-4 Global PCC to choose the single prediction-score source carried to Fold 5, which includes its paired control-only comparator. Controlled fixtures check recognized input consumption, the one-key attention contradiction, missing declarations, abstention, and known input dependence. A fixed compound-strength sweep characterizes threshold crossing in those fixtures; Appendix L explains the criterion's reference-relative interpretation.

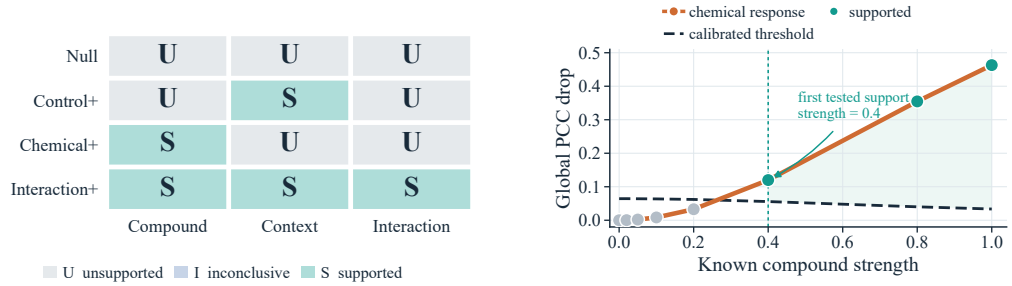


Figure 5: **Controlled checks of input-use measurements.** Null, control-only, perturbation-only, and gated-interaction fixtures produce distinct effects; the strength sweep locates threshold crossing in these fixtures. Source fixtures in Table 41 check implemented, implementation-contradicted, and unresolved source statuses.

Table 37: Post-search Fold-4 code and behavioral audit, held out from discovery. Global, CP, and L1000 PCC describe each selected model, while pre-specified input permutations test whether the trained predictor uses compound identity. A separate paired five-seed panel refits every model under the same training and evaluation setup used in Table 21. Values are mean ± 95% t-interval half-width over five refits. Each seed-level mean aggregates 32 fixed permutations, and the compound-identity test uses the pre-specified threshold rule. Q/U/I gives the qualified/unsupported/inconclusive seed counts; aggregate status applies the registered five-seed rule, with an inconclusive aggregate when no category reaches four of five refits. The across-policy prediction-score selection for Fold 5 is performed only after this panel is complete.

A. Predictive performance

Model	Global PCC	MSE	CP PCC	L1000 PCC
BBBC036				
CELLSCIENTIST	0.3096 ± 0.0029	0.0586 ± 0.0001	0.3113 ± 0.0014	0.3089 ± 0.0041
AIDE	0.2970 ± 0.0075	0.0593 ± 0.0004	0.3052 ± 0.0031	0.2944 ± 0.0102
CELLFORGE	0.2960 ± 0.0036	0.0592 ± 0.0002	0.3056 ± 0.0043	0.2925 ± 0.0037
HARMONYCELL	0.3110 ± 0.0016	0.0586 ± 0.0001	0.3109 ± 0.0020	0.3110 ± 0.0017
h_0	0.2879 ± 0.0040	0.0598 ± 0.0004	0.2981 ± 0.0016	0.2839 ± 0.0056
Control-only	0.3098 ± 0.0024	0.0586 ± 0.0001	0.3094 ± 0.0013	0.3099 ± 0.0028
Chemical-only	0.2239 ± 0.0002	0.0616 ± 0.0000	0.2998 ± 0.0003	0.1832 ± 0.0002
Simple concat	0.2879 ± 0.0040	0.0598 ± 0.0004	0.2981 ± 0.0016	0.2839 ± 0.0056
BBBC047				
CELLSCIENTIST	0.3035 ± 0.0008	0.0457 ± 0.0000	0.3260 ± 0.0016	0.2784 ± 0.0017
AIDE	0.2697 ± 0.0048	0.0469 ± 0.0002	0.3148 ± 0.0047	0.2152 ± 0.0116
CELLFORGE	0.2727 ± 0.0019	0.0467 ± 0.0001	0.3175 ± 0.0035	0.2190 ± 0.0059
HARMONYCELL	0.2983 ± 0.0110	0.0459 ± 0.0003	0.3271 ± 0.0030	0.2654 ± 0.0231
h_0	0.2697 ± 0.0048	0.0469 ± 0.0002	0.3148 ± 0.0047	0.2152 ± 0.0116
Control-only	0.3019 ± 0.0015	0.0458 ± 0.0001	0.3223 ± 0.0011	0.2789 ± 0.0023
Chemical-only	0.2159 ± 0.0012	0.0480 ± 0.0000	0.2935 ± 0.0022	0.0924 ± 0.0028
Simple concat	0.2697 ± 0.0048	0.0469 ± 0.0002	0.3148 ± 0.0047	0.2152 ± 0.0116

B. Input-use evidence

Model	E_{chem}	E_{control}	$I_{\text{chem,control}}$	Q/U/I	Status
BBBC036					
CELLSCIENTIST	0.0000 ± 0.0000	0.1622 ± 0.0095	0.0000 ± 0.0000	0/5/0	Unsup.
AIDE	0.0029 ± 0.0017	0.1577 ± 0.0118	-0.0021 ± 0.0016	0/5/0	Unsup.
CELLFORGE	0.0089 ± 0.0052	0.1464 ± 0.0063	-0.0000 ± 0.0007	0/5/0	Unsup.
HARMONYCELL	0.0000 ± 0.0000	0.1693 ± 0.0033	-0.0000 ± 0.0000	0/5/0	Unsup.
h_0	0.0193 ± 0.0087	0.1359 ± 0.0057	-0.0003 ± 0.0007	1/2/2	Inc.
Control-only	0.0000 ± 0.0000	0.1719 ± 0.0067	0.0000 ± 0.0000	–	N/C
Chemical-only	0.0011 ± 0.0002	0.0000 ± 0.0000	0.0000 ± 0.0000	0/0/5	Inc.
Simple concat	0.0193 ± 0.0087	0.1359 ± 0.0057	-0.0003 ± 0.0007	1/2/2	Inc.
BBBC047					
CELLSCIENTIST	0.0000 ± 0.0000	0.1932 ± 0.0044	0.0000 ± 0.0000	0/5/0	Unsup.
AIDE	0.0118 ± 0.0052	0.1119 ± 0.0121	0.0013 ± 0.0009	0/1/4	Inc.
CELLFORGE	0.0100 ± 0.0045	0.1114 ± 0.0059	0.0012 ± 0.0008	0/2/3	Inc.
HARMONYCELL	0.0006 ± 0.0010	0.1866 ± 0.0225	0.0009 ± 0.0015	0/5/0	Unsup.
h_0	0.0118 ± 0.0052	0.1119 ± 0.0121	0.0013 ± 0.0009	0/1/4	Inc.
Control-only	0.0000 ± 0.0000	0.2030 ± 0.0054	0.0000 ± 0.0000	–	N/C
Chemical-only	0.0111 ± 0.0010	0.0000 ± 0.0000	0.0000 ± 0.0000	5/0/0	Sup.
Simple concat	0.0118 ± 0.0052	0.1119 ± 0.0121	0.0013 ± 0.0009	0/1/4	Inc.

Aggregate status: Sup., claim-supported; Unsup., unsupported; Inc., inconclusive; N/C, not claimed.

Table 38: Fold-5 evaluation after automatic model selection on Fold 4. Prediction-score selection carries forward the source with the highest mean Fold-4 Global PCC; none of the four agent-selected sources meets the pre-specified aggregate compound-use criterion on either task. Values are mean ± 95% t-interval half-width over five paired Fold-5 refits of the frozen source under the same seed schedule. Q/U/I gives the qualified/unsupported/inconclusive seed counts; aggregate status applies the registered five-seed rule, with an inconclusive aggregate when no category reaches four of five refits. "Status replicated" records whether this aggregate status agrees between Fold 4 and Fold 5.

A. Predictive performance

Policy	Model	Global PCC ↑	Global MSE ↓	CP PCC ↑	L1000 PCC ↑
BBBC036					
h_0 reference	h_0	0.2588 ± 0.0059	0.0608 ± 0.0004	0.2596 ± 0.0100	0.2586 ± 0.0094
Score-selected	HARMONYCELL	0.2932 ± 0.0020	0.0590 ± 0.0001	0.2909 ± 0.0018	0.2942 ± 0.0023
BBBC047					
h_0 reference	h_0	0.2786 ± 0.0031	0.0465 ± 0.0003	0.3220 ± 0.0091	0.2243 ± 0.0105
Score-selected	CELLSCIENTIST	0.3153 ± 0.0014	0.0452 ± 0.0000	0.3393 ± 0.0007	0.2872 ± 0.0025

B. Input-use evidence and replication

Model	E_{chem}	E_{control}	$I_{\text{chem,control}}$	Q/U/I	Status	Status replicated?
BBBC036						
h_0	0.0047 ± 0.0031	0.1202 ± 0.0085	0.0042 ± 0.0011	0/5/0	Unsup.	no
HARMONYCELL	0.0000 ± 0.0000	0.1516 ± 0.0035	-0.0000 ± 0.0000	0/5/0	Unsup.	yes
BBBC047						
h_0	0.0096 ± 0.0037	0.1199 ± 0.0109	0.0012 ± 0.0007	0/1/4	Inc.	yes
CELLSCIENTIST	0.0000 ± 0.0000	0.2004 ± 0.0061	0.0000 ± 0.0000	0/5/0	Unsup.	yes

Score-selected denotes prediction-score selection. Aggregate status: Unsup., unsupported; Inc., inconclusive.

Table 39: Fold-5 comparison with a control-only reference. The score-selected model and control-only predictor are refit with the same five seeds and evaluated without reselection. Values are mean ± 95% t-interval half-width; differences are paired across seeds.

Task	Score-selected model	Selected-model PCC ↑	Control-only PCC ↑	Paired Δ [95% CI]
BBBC036	HARMONYCELL	0.2932 ± 0.0020	0.2921 ± 0.0017	+0.0011 [−0.0008, +0.0030]
BBBC047	CELLSCIENTIST	0.3153 ± 0.0014	0.3142 ± 0.0012	+0.0011 [−0.0005, +0.0027]

Table 40: Control-adjusted Fold-4 response evaluation. Target and prediction are residualized by the same paired control-only prediction g before Global PCC is computed. Chemical effect is the mean residual-PCC reduction after a target-blind chemical shuffle. Control-only self-comparison has exactly zero increment, so residual PCC and its shuffle contrast are undefined ($--$). MSE is unchanged: $\mathrm{MSE}(f-g, y-g) = \mathrm{MSE}(f, y)$; the displayed MSE remains on the original response scale. Residual PCC is mean [95% t interval] over five paired fitting seeds.

A. BBBC036

Model	**Residual PCC ↑** Mean [95% CI]	**MSE ↓**	**Pert. effect ↑**
h_0	0.0474 [0.0371, 0.0577]	0.0598	0.0381
CELLSCIENTIST	0.0294 [0.0052, 0.0535]	0.0586	0.0000
AIDE	0.0301 [0.0236, 0.0366]	0.0593	0.0050
CELLFORGE	0.0336 [0.0234, 0.0438]	0.0592	0.0223
HARMONYCELL	0.0341 [0.0226, 0.0455]	0.0586	0.0000
Control-only	Zero increment	0.0586	–
Chemical-only	0.0052 [-0.0044, 0.0148]	0.0616	0.0012
Simple concat	0.0474 [0.0371, 0.0577]	0.0598	0.0381

B. BBBC047

Model	**Residual PCC ↑** Mean [95% CI]	**MSE ↓**	**Pert. effect ↑**
h_0	0.0558 [0.0510, 0.0606]	0.0469	0.0205
CELLSCIENTIST	0.0538 [0.0422, 0.0654]	0.0457	0.0000
AIDE	0.0558 [0.0510, 0.0606]	0.0469	0.0205
CELLFORGE	0.0553 [0.0510, 0.0596]	0.0467	0.0174
HARMONYCELL	0.0567 [0.0470, 0.0663]	0.0459	0.0020
Control-only	Zero increment	0.0458	–
Chemical-only	0.0457 [0.0382, 0.0532]	0.0480	0.0104
Simple concat	0.0558 [0.0510, 0.0606]	0.0469	0.0205

Table 41: Calibration of the fixed audit rule. **Top:** controlled code examples test whether source inspection identifies known implementation errors. **Bottom:** synthetic responses with known input dependence test the held-out permutation rule. These fixtures verify the fixed adjudication rule independently of the biological-model results.

A. Code inspection

Test case	Expected code pattern	Decision
Compound input used	Direct compound input path	Implemented
Control input used	Direct control input path	Implemented
Gated interaction	Joint compound–control path	Implemented
Singleton attention	Query cancellation in the cited path	Implementation-contradicted
Absent source symbol	Manifest anchor missing from source	Implementation-contradicted
Dynamic function	Input use cannot be resolved statically	Unresolved

B. Held-out input test

Fixture	Supported input-use claims	Other tested claims
Null	None	Chemical, control, interaction unsupported
Chemical positive	Chemical	Control and interaction unsupported
Control positive	Control	Chemical and interaction unsupported
Interaction positive	Chemical, control, interaction	–

F.1 KNOWN-TRUTH MEASUREMENT AND UNCERTAINTY CHECKS

Registered independent-dataset experiment. The fixed-context simulator uses $f(z, c) = (az, c)$ and $Y = (bz + e, c)$. Matched replacement draws an independent normal z' while fixing c, Y. The loss is the *sum* of squared errors over both channels, so the population paired loss increment is $\tau = 2ab$; the mean squared Euclidean prediction distance is $\eta = 2a^2$, not RMS. Groups receive equal weight and rows equal weight within group. The experiment evaluates equal-group-weighted population contrasts; real-data row-weighted resampling is reported separately. Four classes are invariant $(a, b) = (0, .2)$, sensitive without benefit $(.2, 0)$, beneficial $(.2, .2)$, and harmful $(-.2, .2)$. Their loss truths are $0, 0, .08, -.08$ and distance truths $0, .08, .08, .08$.

The frozen $2^4 = 16$ scenarios cross $G \in \{8, 32\}$ independent source groups, balanced size 8 or independently sampled sizes $\{2, 4, 8, 32\}$ with equal probabilities, within-group correlation $\rho \in \{0, .6\}$, and noise SD $\sigma \in \{.25, 1\}$. Each scenario has 400 independent datasets (6400 total). Maps 32/128 are nested on the same datasets, not independent cohorts. Group-percentile intervals use 999 bootstrap draws; group-t uses $G - 1$ degrees of freedom. Source/donor, bootstrap and context seeds are 2026091401, 2026091402 and 2026091403.

Conditional and population objectives. Map-only intervals target the donor-integrated contrast of one fixed dataset; group-aware intervals target the population expectation over independent groups. Their coverage is therefore reported against both truths. Positive evidence means a strictly positive lower bound of a two-sided 95% CI: the positive one-tail nominal Type-I error is .025, not .05. Power is positive-direction evidence for benefit and negative-direction evidence for harm. Prediction distance does not test target benefit. Invariant data and intervals are identically zero and are reported as a deterministic check, not "100% calibrated" stochastic coverage.

Table 42: Calibration ranges across all 16 scenarios under the two nested donor-map budgets.

Endpoint / class	CI method	**Population** coverage	**Conditional** coverage	**Directed** evidence rate
32 donor maps				
Null loss	Map only	[0.1500, 0.3000]	[0.9275, 0.9675]	[0.2850, 0.5000]
Null loss	Group t	[0.9200, 0.9775]	[1.0000, 1.0000]	[0.0150, 0.0725]
Null loss	Group bootstrap	[0.8175, 0.9400]	[1.0000, 1.0000]	[0.0350, 0.1475]
Beneficial loss	Map only	[0.1425, 0.3350]	[0.9250, 0.9650]	[0.7100, 1.0000]
Beneficial loss	Group t	[0.9100, 0.9650]	[0.9975, 1.0000]	[0.1200, 1.0000]
Beneficial loss	Group bootstrap	[0.8350, 0.9425]	[0.9975, 1.0000]	[0.2650, 1.0000]
Harmful loss	Map only	[0.1050, 0.2750]	[0.9375, 0.9650]	[0.6775, 1.0000]
Harmful loss	Group t	[0.8625, 0.9725]	[1.0000, 1.0000]	[0.0550, 1.0000]
Harmful loss	Group bootstrap	[0.8300, 0.9400]	[1.0000, 1.0000]	[0.1900, 1.0000]
Squared distance	Map only	[0.2175, 0.5250]	[0.9400, 0.9700]	[1.0000, 1.0000]
Squared distance	Group t	[0.8375, 0.9525]	[1.0000, 1.0000]	[0.9975, 1.0000]
Squared distance	Group bootstrap	[0.7800, 0.9375]	[0.9975, 1.0000]	[1.0000, 1.0000]
128 donor maps				
Null loss	Map only	[0.0675, 0.1625]	[0.9225, 0.9725]	[0.3600, 0.5375]
Null loss	Group t	[0.9125, 0.9775]	[1.0000, 1.0000]	[0.0125, 0.0775]
Null loss	Group bootstrap	[0.8200, 0.9350]	[1.0000, 1.0000]	[0.0350, 0.1450]
Beneficial loss	Map only	[0.0750, 0.1750]	[0.9175, 0.9650]	[0.7675, 1.0000]
Beneficial loss	Group t	[0.9075, 0.9725]	[1.0000, 1.0000]	[0.1200, 1.0000]
Beneficial loss	Group bootstrap	[0.8400, 0.9450]	[1.0000, 1.0000]	[0.2475, 1.0000]
Harmful loss	Map only	[0.0400, 0.1450]	[0.9125, 0.9600]	[0.7025, 1.0000]
Harmful loss	Group t	[0.8600, 0.9700]	[1.0000, 1.0000]	[0.0500, 1.0000]
Harmful loss	Group bootstrap	[0.8200, 0.9425]	[1.0000, 1.0000]	[0.1875, 1.0000]
Squared distance	Map only	[0.0925, 0.2900]	[0.9350, 0.9600]	[1.0000, 1.0000]
Squared distance	Group t	[0.8425, 0.9500]	[1.0000, 1.0000]	[0.9975, 1.0000]
Squared distance	Group bootstrap	[0.8025, 0.9350]	[1.0000, 1.0000]	[1.0000, 1.0000]

Each rate uses 400 independent datasets per scenario; ranges are minima/maxima over scenarios, not pooled coverage. The last column is positive Type I for null loss (nominal .025), correctly directed power for beneficial/harmful loss, and positive-distance evidence for distance. Conditional and population targets differ. Invariant cases are excluded from stochastic coverage.

Limits retained in interpretation. The group-aware methods still under-cover in some scenarios; the bootstrap is particularly weak with few groups, and positive null declarations can exceed the nominal one-tail rate. Independent normal donors, independent groups, and correct context match-

Table 43: All 16 null-loss scenarios at 128 maps: population coverage / positive Type I (each cell, top / bottom).

G/size/ρ/σ	Map only	Group t	**Group** bootstrap	t **Type-I** MCSE	t **Type-I** Wilson 95% CI
8/B/0/0.25	0.1275 0.4600	0.9500 0.0400	0.8675 0.0950	0.0098	[0.0248, 0.0640]
8/B/0/1	0.1175 0.4925	0.9550 0.0325	0.8950 0.0700	0.0089	[0.0191, 0.0548]
8/B/0.6/0.25	0.0875 0.5075	0.9325 0.0600	0.8200 0.1450	0.0119	[0.0406, 0.0877]
8/B/0.6/1	0.0675 0.4925	0.9650 0.0225	0.8525 0.0850	0.0074	[0.0119, 0.0422]
8/I/0/0.25	0.1300 0.4675	0.9325 0.0650	0.8500 0.1275	0.0123	[0.0447, 0.0935]
8/I/0/1	0.1550 0.3600	0.9775 0.0125	0.8825 0.0650	0.0056	[0.0054, 0.0289]
8/I/0.6/0.25	0.0725 0.5375	0.9350 0.0650	0.8400 0.1375	0.0123	[0.0447, 0.0935]
8/I/0.6/1	0.0825 0.4625	0.9600 0.0250	0.8300 0.1000	0.0078	[0.0136, 0.0454]
32/B/0/0.25	0.1125 0.4275	0.9325 0.0300	0.9225 0.0375	0.0085	[0.0172, 0.0517]
32/B/0/1	0.1350 0.4500	0.9625 0.0225	0.9350 0.0350	0.0074	[0.0119, 0.0422]
32/B/0.6/0.25	0.0850 0.4650	0.9325 0.0575	0.9175 0.0675	0.0116	[0.0386, 0.0848]
32/B/0.6/1	0.0875 0.4600	0.9600 0.0250	0.9350 0.0350	0.0078	[0.0136, 0.0454]
32/I/0/0.25	0.1450 0.3975	0.9450 0.0425	0.9225 0.0500	0.0101	[0.0267, 0.0670]
32/I/0/1	0.1625 0.4175	0.9450 0.0375	0.9225 0.0425	0.0095	[0.0229, 0.0609]
32/I/0.6/0.25	0.0975 0.4900	0.9125 0.0775	0.8925 0.0850	0.0134	[0.0551, 0.1079]
32/I/0.6/1	0.0750 0.4450	0.9375 0.0450	0.9025 0.0625	0.0104	[0.0287, 0.0700]

B/I: balanced/imbalanced. Type-I nominal .025. MCSE is $\sqrt{p(1-p)/400}$ (worst possible .025); Wilson intervals are pointwise Monte Carlo intervals, not simultaneous bounds over 16 scenarios.

Table 44: Estimator bias and Monte Carlo uncertainty across the 16 scenarios.

Endpoint / class	Maps	Truth	Bias range	Bias MCSE range
Null loss	32	0	$[-8.41\times10^{-3},\ 4.95\times10^{-3}]$	$[3.84\times10^{-4},\ 5.21\times10^{-3}]$
Null loss	128	0	$[-8.42\times10^{-3},\ 5.48\times10^{-3}]$	$[3.82\times10^{-4},\ 5.14\times10^{-3}]$
Beneficial loss	32	0.08	$[-7.32\times10^{-3},\ 4.58\times10^{-3}]$	$[3.50\times10^{-4},\ 5.14\times10^{-3}]$
Beneficial loss	128	0.08	$[-7.47\times10^{-3},\ 5.06\times10^{-3}]$	$[3.47\times10^{-4},\ 5.06\times10^{-3}]$
Harmful loss	32	−0.08	$[-4.17\times10^{-3},\ 6.55\times10^{-3}]$	$[5.70\times10^{-4},\ 5.48\times10^{-3}]$
Harmful loss	128	−0.08	$[-4.68\times10^{-3},\ 6.51\times10^{-3}]$	$[5.65\times10^{-4},\ 5.40\times10^{-3}]$
Squared distance	32	0.08	$[-4.19\times10^{-4},\ 7.09\times10^{-4}]$	$[1.69\times10^{-4},\ 7.60\times10^{-4}]$
Squared distance	128	0.08	$[-3.08\times10^{-4},\ 5.67\times10^{-4}]$	$[1.68\times10^{-4},\ 7.49\times10^{-4}]$

All CI methods share the same point estimator. Bias MCSE is the empirical SD of estimation error divided by $\sqrt{400}$.

ing are explicit simulation assumptions, not demonstrations of real biological conditional exchangeability or support. Finite reused donor pools and dependent biological groups require additional dependence handling. More maps reduce conditional Monte Carlo error but do not create more biological replicates.

Table 45: Historical reference-relative gate replay at 128 maps with unrelated context variation.

Class	Amplitude	Mean effect	Mean Q	Support-rate range
Invariant	0	0	0	[0, 0]
Invariant	0.2	0	0.08792	[0, 0]
Invariant	2.0	0	8.79230	[0, 0]
Sensitive/no benefit	0	−0.00067	0.07744	[0, 0.007]
Sensitive/no benefit	0.2	−0.00067	0.12034	[0, 0.003]
Sensitive/no benefit	2.0	−0.00067	8.79274	[0, 0]
Beneficial	0	0.07946	0.08226	[0, 1.000]
Beneficial	0.2	0.07946	0.12260	[0, 0.110]
Beneficial	2.0	0.07946	8.79279	[0, 0]
Harmful	0	−0.07961	0.08218	[0, 0]
Harmful	0.2	−0.07961	0.12276	[0, 0]
Harmful	2.0	−0.07961	8.79248	[0, 0]

Means average all 16 scenarios; support ranges remain scenario-specific. This replays the historical formula on negative squared-loss scores, not historical PCC. Changing orthogonal context changes its reference threshold while keeping the chemical contrast fixed. Non-support is not a false negative for the different proposition $\tau > 0$.

G Complete input-intervention results

The complete intervention matrix reports search performance, held-out predictive performance, compound-identity effect, control-profile effect, target-loss gain, and input-use status in separate columns. Rows summarize the full pre-specified model sets, and −− marks contrasts that cannot be identified under the protocol. Response-block decompositions are diagnostic views of the same selected models and fitted checkpoints. Figure 6 follows the compound and control effects of the falsification-guided checkpoints across both held-out folds.

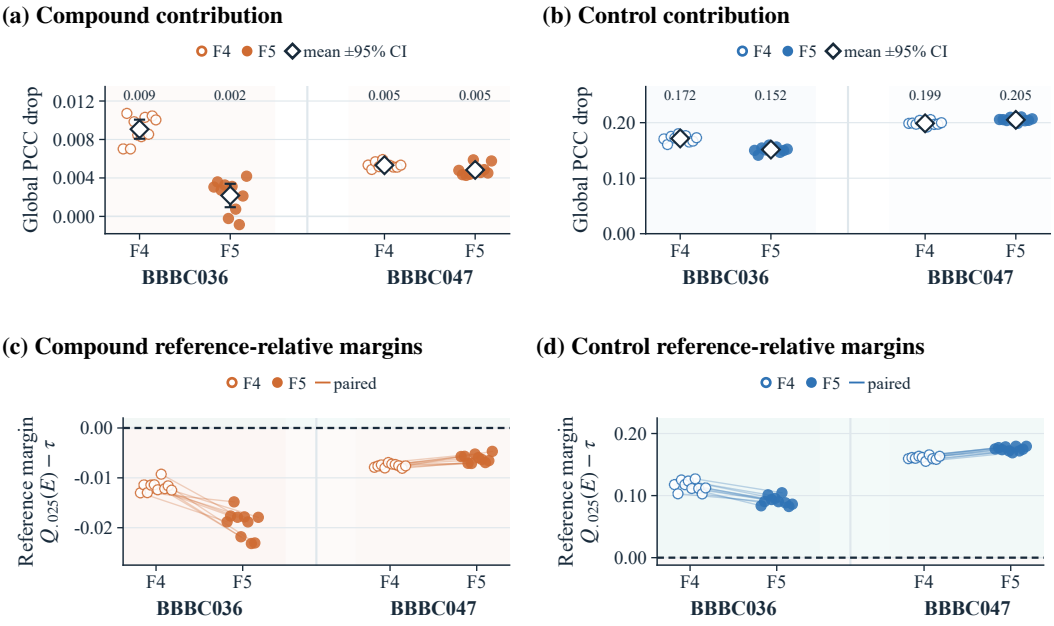


Figure 6: **Compound sensitivity and control dependence across the BBBC cohorts.** Each cohort contributes ten trajectory-selected checkpoints, evaluated unchanged on both folds. (a,b) Checkpoint-level factorial effects E_{cmpd} and E_{ctrl}; diamonds and bars give ten-endpoint means and 95% t-intervals. (c,d) Lower effect quantile minus the checkpoint-local threshold; lines connect the same checkpoint across folds. Inputs use distinct vertical scales. Dose contrasts are non-identifiable under the registered eligibility rule (Table 50). Table 47 gives predictive comparisons.

Table 46: Prediction and input-use evidence across three model-search settings. The prediction-score source is chosen by the Fold-4 score rule and summarized across five paired refits; constrained settings retain all ten trajectory endpoints. Values are mean ± 95% interval half-width. Factorial effects E_{cmpd} and E_{ctrl} average PCC drops over the other input's two states; $E^{\mathcal{L}}_{\mathrm{cmpd}}$ is the corresponding loss effect. Q/U/I reports contribution relative to the registered reference variation.

A. Selection record

Task	Search setting	Selected source	Fold-3 PCC
BBBC036	Prediction-score	HarmonyCell	0.2985
BBBC036	Path-constrained	10-endpoint cohort	0.2925 ± 0.0019
BBBC036	Falsification-guided	10-endpoint cohort	0.3006 ± 0.0006
BBBC047	Prediction-score	CellScientist	0.3231
BBBC047	Path-constrained	10-endpoint cohort	0.3008 ± 0.0012
BBBC047	Falsification-guided	10-endpoint cohort	0.3238 ± 0.0002

B. Fold 4: held-out input-use evidence

Search setting	PCC	E_{cmpd}	$E^{\mathcal{L}}_{\mathrm{cmpd}}$	E_{ctrl}	Q/U/I
BBBC036					
Prediction-score	0.3110 ± 0.0016	0.0000 ± 0.0000	0.00000 ± 0.00000	0.1693 ± 0.0033	0/5/0
Path-constrained	0.3034 ± 0.0029	0.0007 ± 0.0006	0.00003 ± 0.00003	0.1592 ± 0.0050	0/10/0
Falsification-guided	0.3167 ± 0.0013	0.0091 ± 0.0010	0.00039 ± 0.00004	0.1716 ± 0.0042	0/10/0
BBBC047					
Prediction-score	0.3035 ± 0.0008	0.0000 ± 0.0000	0.00000 ± 0.00000	0.1932 ± 0.0044	0/5/0
Path-constrained	0.2822 ± 0.0017	0.0083 ± 0.0016	0.00024 ± 0.00005	0.1388 ± 0.0052	0/6/4
Falsification-guided	0.3060 ± 0.0004	0.0053 ± 0.0002	0.00018 ± 0.00001	0.1988 ± 0.0027	0/10/0

C. Fold 5: held-out input-use evidence

Search setting	PCC	E_{cmpd}	$E^{\mathcal{L}}_{\mathrm{cmpd}}$	E_{ctrl}	Q/U/I
BBBC036					
Prediction-score	0.2932 ± 0.0020	0.0000 ± 0.0000	0.00000 ± 0.00000	0.1516 ± 0.0035	0/5/0
Path-constrained	0.2886 ± 0.0018	0.0039 ± 0.0022	0.00016 ± 0.00009	0.1423 ± 0.0050	0/10/0
Falsification-guided	0.2897 ± 0.0008	0.0022 ± 0.0012	0.00009 ± 0.00005	0.1515 ± 0.0037	0/10/0
BBBC047					
Prediction-score	0.3153 ± 0.0014	0.0000 ± 0.0000	0.00000 ± 0.00000	0.2004 ± 0.0061	0/5/0
Path-constrained	0.2922 ± 0.0018	0.0074 ± 0.0017	0.00022 ± 0.00005	0.1477 ± 0.0057	0/7/3
Falsification-guided	0.3173 ± 0.0004	0.0048 ± 0.0004	0.00017 ± 0.00001	0.2051 ± 0.0024	0/10/0

Table 47: **Falsification-guided models increase average held-out compound contribution, while predictive increments vary by task and fold.** Δ compares the full and control-only predictors on the same fold. Factorial effects E_{cmpd} and E_{ctrl} average PCC drops over the other input's two states; $E^{\mathcal{L}}_{\mathrm{cmpd}}$ averages the corresponding loss increases. Q/U/I summarizes per-model contribution relative to the registered reference variation.

A. Predictive performance

Task	Fold	Control-only PCC	Full PCC	Δ Full−Control
BBBC036	Fold 4	0.3121 ± 0.0011	0.3167 ± 0.0013	0.0046 ± 0.0006
BBBC036	Fold 5	0.2917 ± 0.0006	0.2897 ± 0.0008	-0.0020 ± 0.0009
BBBC047	Fold 4	0.3023 ± 0.0004	0.3060 ± 0.0004	0.0037 ± 0.0003
BBBC047	Fold 5	0.3145 ± 0.0005	0.3173 ± 0.0004	0.0028 ± 0.0004

B. Input-use evidence

Task	Fold	E_{cmpd}	$E^{\mathcal{L}}_{\mathrm{cmpd}}$	E_{ctrl}	Model Q/U/I
BBBC036	Fold 4	0.0091 ± 0.0010	0.00039 ± 0.00004	0.1716 ± 0.0042	0/10/0
BBBC036	Fold 5	0.0022 ± 0.0012	0.00009 ± 0.00005	0.1515 ± 0.0037	0/10/0
BBBC047	Fold 4	0.0053 ± 0.0002	0.00018 ± 0.00001	0.1988 ± 0.0027	0/10/0
BBBC047	Fold 5	0.0048 ± 0.0004	0.00017 ± 0.00001	0.2051 ± 0.0024	0/10/0

Table 48: **Predictive performance and perturbation evidence define a measurable frontier.** Falsification-guided entries summarize ten independently selected models; fixed residual learners summarize five paired refits. Global PCC, compound-shuffle PCC drop, and compound-shuffle loss increase are evaluated on the same responses, data splits, and fixed input permutations.

Predictor	**Global PCC** [95% CI]	**Chemical PCC drop** [95% CI]	**Target-loss gain** [95% CI]
BBBC036 / Fold 4			
Falsification-guided	0.3167 [0.3154, 0.3181]	0.0091 [0.0081, 0.0101]	0.00039 [0.00034, 0.00043]
Residual MLP	0.3056 [0.3027, 0.3085]	0.0036 [0.0028, 0.0044]	0.00015 [0.00012, 0.00019]
Joint Ridge	0.0892 [0.0867, 0.0917]	0.0153 [0.0147, 0.0159]	0.00299 [0.00290, 0.00308]
BBBC036 / Fold 5			
Falsification-guided	0.2897 [0.2890, 0.2905]	0.0022 [0.0010, 0.0034]	0.00009 [0.00004, 0.00014]
Residual MLP	0.2861 [0.2829, 0.2892]	0.0007 [0.0001, 0.0013]	0.00003 [0.00000, 0.00005]
Joint Ridge	0.1114 [0.1092, 0.1136]	0.0336 [0.0328, 0.0344]	0.00569 [0.00557, 0.00582]
BBBC047 / Fold 4			
Falsification-guided	0.3060 [0.3056, 0.3064]	0.0053 [0.0051, 0.0055]	0.00018 [0.00017, 0.00019]
Residual MLP	0.3031 [0.3007, 0.3055]	0.0119 [0.0094, 0.0145]	0.00042 [0.00033, 0.00050]
Joint Ridge	0.0774 [0.0765, 0.0783]	0.0082 [0.0079, 0.0085]	0.00420 [0.00416, 0.00423]
BBBC047 / Fold 5			
Falsification-guided	0.3173 [0.3169, 0.3177]	0.0048 [0.0044, 0.0052]	0.00017 [0.00015, 0.00018]
Residual MLP	0.3154 [0.3135, 0.3174]	0.0113 [0.0095, 0.0132]	0.00040 [0.00033, 0.00047]
Joint Ridge	0.0704 [0.0693, 0.0715]	0.0052 [0.0050, 0.0053]	0.00394 [0.00393, 0.00395]

Table 49: Response-block decomposition of the pre-specified 2×2 input audit. Perturbation and context columns are factorial PCC effects E_{cmpd} and E_{ctrl}; interaction is their score nonadditivity. The blocks locate input contribution within the joint response. Values are mean $\pm$ 95% interval half-width across selected models.

Fold	Perturbation		Context		Interaction	
	CP	L1000	CP	L1000	CP	L1000
Path-constrained / BBBC036						
F4	0.0012 $\pm$ 0.0009	0.0004 $\pm$ 0.0005	0.0250 $\pm$ 0.0023	0.2164 $\pm$ 0.0057	0.0003 $\pm$ 0.0004	-0.0003 $\pm$ 0.0006
F5	0.0008 $\pm$ 0.0004	0.0053 $\pm$ 0.0031	0.0290 $\pm$ 0.0024	0.1935 $\pm$ 0.0056	0.0005 $\pm$ 0.0002	0.0013 $\pm$ 0.0008
Path-constrained / BBBC047						
F4	0.0130 $\pm$ 0.0026	0.0020 $\pm$ 0.0002	0.0911 $\pm$ 0.0037	0.2178 $\pm$ 0.0054	0.0026 $\pm$ 0.0012	0.0007 $\pm$ 0.0002
F5	0.0118 $\pm$ 0.0026	0.0013 $\pm$ 0.0005	0.1005 $\pm$ 0.0054	0.2251 $\pm$ 0.0054	0.0011 $\pm$ 0.0012	-0.0001 $\pm$ 0.0003
Falsification-guided / BBBC036						
F4	0.0141 $\pm$ 0.0023	0.0069 $\pm$ 0.0007	0.0248 $\pm$ 0.0015	0.2363 $\pm$ 0.0041	0.0000 $\pm$ 0.0000	0.0001 $\pm$ 0.0000
F5	-0.0014 $\pm$ 0.0016	0.0039 $\pm$ 0.0013	0.0284 $\pm$ 0.0016	0.2088 $\pm$ 0.0035	0.0001 $\pm$ 0.0000	0.0002 $\pm$ 0.0001
Falsification-guided / BBBC047						
F4	0.0088 $\pm$ 0.0003	0.0014 $\pm$ 0.0001	0.1338 $\pm$ 0.0031	0.2759 $\pm$ 0.0013	0.0002 $\pm$ 0.0001	0.0001 $\pm$ 0.0000
F5	0.0079 $\pm$ 0.0007	0.0013 $\pm$ 0.0002	0.1421 $\pm$ 0.0025	0.2806 $\pm$ 0.0014	-0.0000 $\pm$ 0.0001	0.0001 $\pm$ 0.0000

Table 50: Input-use outcomes for the two constrained search settings. Counts are qualified/unsupported/inconclusive/not-identifiable (Q/U/I/N) under the reference-variation criterion across ten selected models. Path-constrained dose eligibility accepts any observed dose change within the released BRD compound identity; falsification-guided eligibility additionally requires a 0.1 log-dose contrast and at least eight eligible compounds. Identity validation gives 0/4,380 and 14/3,878 eligible BBBC047 rows on Folds 4 and 5, below the 64-row and 10% requirements. $--$ with N=10 denotes a non-identifiable contrast, not missing execution.

Task	Fold	Δ_{dose}	Compound	Control	Dose	**Compound**$\times$ control
Path-constrained						
BBBC036	4	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC036	5	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC047	4	–	0/6/4/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC047	5	–	0/7/3/0	10/0/0/0	0/0/0/10	0/10/0/0
Falsification-guided						
BBBC036	4	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC036	5	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC047	4	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0
BBBC047	5	–	0/10/0/0	10/0/0/0	0/0/0/10	0/10/0/0

H ALL-CANDIDATE SOURCE CHECKING

Table 51 applies the deterministic checker to every retained prediction-score candidate and separates execution failure from checker abstention. Resolution coverage is the fraction of checked candidates whose cited source locations are all resolved. For candidates without a structured manifest declaring input use, the checker reports findings supported by pre-specified source locations. The one-key attention certificate is a code-level finding; complete-model replacement tests provide the corresponding behavioral evidence.

A deterministic hash selects model sources within each checker category for source-interface inspection. The inspection is restricted to pre-specified source locations and public fusion interfaces; fixed behavioral replacements remain applicable across all code-check categories.

Table 51: Deterministic code-checker results over all prediction-score candidate slots. Resolved candidates have every registered source location linked to input-consuming code or a specific contradiction; abstention leaves at least one location unresolved. Coverage reports resolved/checked candidates and checked/total slots, rather than classification accuracy. A singleton-attention certificate concerns the cited pathway.

	Candidate slots			Code checker			Coverage	
Policy	Total	Exec.	Checked	Resolved	Singleton	Abstain	**Resolved/** checked	**Checked/** slots
BBBC036								
CellScientist	100	99	99	60	15	39	60.6%	99.0%
AIDE	100	100	100	100	0	0	100.0%	100.0%
CellForge	100	91	91	91	0	0	100.0%	91.0%
HarmonyCell	100	95	95	94	0	1	98.9%	95.0%
BBBC047								
CellScientist	100	94	94	48	17	46	51.1%	94.0%
AIDE	100	99	99	99	0	0	100.0%	99.0%
CellForge	100	94	94	94	0	0	100.0%	94.0%
HarmonyCell	100	95	95	94	0	1	98.9%	95.0%

Table 52: **Deterministic compilation links each declared input use to code.** "All statements resolved" means that the checker or compiler finds supporting code or a specific contradiction for every emitted static statement. After selection, the independent behavioral audit tests the registered input-use claim; singleton-attention findings remain narrow code-level certificates.

Search	Slots	Exec.	Resolved	Singleton	Abstain	Resolve %	Exec. %
BBBC036							
Prediction-score	100	99	60	15	39	60.6%	99.0%
Path-constrained	100	100	100	0	0	100.0%	100.0%
BBBC047							
Prediction-score	100	94	48	17	46	51.1%	94.0%
Path-constrained	100	100	100	0	0	100.0%	100.0%

Table 53: **Checker abstention and source-interface inspection.** A deterministic hash selects sources within each category for inspection of metadata anchors and exposed fusion interfaces. The taxonomy describes static-analysis coverage; complete-model input tests provide behavioral evidence.

Type	Sources	Inspected	Deterministic disposition
Missing fusion source location	75	12	No cited fusion symbol is available; the checker returns unresolved and the behavioral input tests remain applicable.
Partial input scope	11	11	The cited fusion exposes only part of the registered input-use claim; the unchecked role receives an unresolved source status.
Generic unbound aliases	1	1	No declared binding identifies the generic argument names; the checker assigns an unresolved source status.

A linked sample of 48 real candidate sources extends testing beyond the one-key attention case. Of these, 47 change their predictions after compound replacement at both boundaries; 20 have positive target-loss scaffold intervals at both. An implemented multi-tower source is exactly invariant. Source consumption, fitted dependence, and target benefit thus separate in diverse generated code (Appendix M).

I PREDICTION-SPACE DEPENDENCE

The prediction-space analysis measures normalized RMS change and cosine change between $f(c, p, a)$ and predictions obtained after fixed input permutations. Table 54 reports prediction-change RMS divided by the target RMS after centering each output feature across evaluation rows, with a denominator floor of 10^{-12}. Both RMS quantities average over rows and features within the reported response block. This complements target-based effects: near-zero distance indicates negligible prediction change on the tested replacements, whereas nonzero distance with near-zero target-loss gain indicates sensitivity with little measured benefit on the observed response. The calculation requires inference only over fixed checkpoints.

Table 54: Prediction-space dependence under target-blind interventions. Normalized RMS changes quantify whether the fitted outputs change under replacement; target-loss contrasts, reported separately, determine whether that change improves predictive utility.

Model	Compound replacement			Control replacement
	Global	CP	L1000	Global
BBBC036				
CellScientist	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.3112 ± 0.0246
Chemical-only	0.0560 ± 0.0013	0.0580 ± 0.0014	0.0552 ± 0.0013	0.0000 ± 0.0000
Control-only	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.3390 ± 0.0199
BBBC047				
CellScientist	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.3593 ± 0.0207
Chemical-only	0.1253 ± 0.0066	0.1467 ± 0.0093	0.1025 ± 0.0055	0.0000 ± 0.0000
Control-only	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.4011 ± 0.0199

J REPRESENTATIVE DISCOVERY RECORDS

The main text follows one model from selection through source checking and held-out replication. This section complements that input-use analysis with two preselected trajectory-level run records: a representative successful search and a search containing code repair. Each record preserves the initial hypothesis, decisive score changes, component edits, failures, repair actions, prompt and source hashes, and final selection. The complete ten-slot traces show both successful and failed execution paths.

The two cases are selected using Fold-3 run records only: a lower-median BBBC036 trajectory and the highest-scoring BBBC047 trajectory among those with at least one recorded candidate-code repair. Their records include the initial program, complete candidate sequence, parent hashes, structured decision context, candidate source, execution and repair outcomes, provider request/response hashes, and the versioned prompt constructor that defines the generation input.

J.1 REPRESENTATIVE MULTIMODAL DISCOVERY TRACE

The initial candidate is a reproducible concatenation–MLP joint predictor with joint input fusion, a shared response program, a joint readout, response loss, and AdamW optimization; it obtains Fold-3 Global PCC 0.2668. The recorded decision context for the retained final revision is: *"Make one final diagnosis-driven revision of the best incumbent; prioritize a complete, executable perturbation-response hypothesis over a hyperparameter-only change."* The selected multi-head cross-modal-attention candidate uses separate control, compound, and dose encoders, cross-modal attention fusion, a shared response program, joint readout, block-weighted MSE, AdamW, and cosine scheduling, reaching 0.2928. The trace contains 11 documented provider calls (53,096 reported tokens), each with a request hash, response hash, status, and token record.

Table 55: Complete candidate record for the preselected representative search trajectory. Selection-fold PCC is shown for every completed candidate; repairs are candidate-code repair attempts. The selected candidate is bold.

Slot	Recorded revision design	Global PCC ↑	Repairs
h_0	Initial concatenation–MLP joint predictor	0.2668	0
1	Chemical-query/control-key-value cross-modal attention with dose residual	0.2903	0
2	FiLM-style chemical-and-dose modulation	0.2742	0
3	Learned chemical–control gating	0.2865	0
4	Low-rank bilinear chemical–control fusion	0.2355	0
5	Multi-tower residual fusion with correlation-balanced MSE	0.2812	0
6	Conditional low-rank transformation of control representation	0.2733	0
7	Gated bilinear bottleneck fusion	0.2682	1
8	Chemical-and-dose hypernetwork fusion	0.2787	1
9	**Multi-head cross-modal attention revision**	**0.2928**	0

J.2 Repair-containing discovery trace

The second trace begins from the same initial program at Fold-3 Global PCC 0.2881. Its structured agenda specifies separate control, compound, and dose encoders; compound-query/control-key-value interaction with a dose residual; a shared response program; joint response prediction; CP/L1000 block-weighted MSE (0.4/0.6); AdamW; and cosine scheduling. Its selected first revision implements this design, resolves an output-shape mismatch by returning a $[\mathrm{batch}, \mathrm{target}]$ tensor, and reaches 0.3230. The complete trace contains 13 provider calls (63,572 reported tokens); a later slot exhausts all three repair attempts and remains recorded as a failed slot.

Table 56: Complete candidate record for the preselected search trajectory containing a repair failure. The failed eighth slot has no evaluation score because all three allowed repair attempts were exhausted.

Slot	Recorded revision design	Global PCC ↑	Repairs
h_0	Initial concatenation–MLP joint predictor	0.2881	0
1	**Cross-modal attention, dose residual, block-weighted MSE**	**0.3230**	1
2	FiLM-style gated perturbation modulation	0.3065	0
3	Factorized response heads with cross-modal gate	0.2894	0
4	Low-rank bilinear fusion with dose gate	0.2837	0
5	Multi-tower residual response program	0.3223	0
6	Low-rank joint readout bottleneck	0.3165	0
7	Multi-tower residual response program	0.3227	0
8	Failed candidate-code attempt	–	3
9	Multi-tower residual response program	0.3217	0

K Extended related work

Cellular perturbation data and virtual-cell predictors. Cell Painting, L1000, Perturb-seq, and sci-Plex measure morphological and transcriptional responses across chemical and genetic interventions (Bray et al., 2016; Caicedo et al., 2017; Subramanian et al., 2017; Dixit et al., 2016; Srivatsan et al., 2020). Matched resources and harmonized collections broaden this coverage (Haghighi et al., 2022; Chandrasekaran et al., 2024; Peidli et al., 2024). Predictors use latent-state transfer, compositional models, neural transport, genetic graphs, and foundation models (Lotfollahi et al., 2019; 2023; Bunne et al., 2023; Roohani et al., 2024; Cui et al., 2024); cycleCDR adds cycle-consistency constraints to learn transferable perturbation representations (Huang & Liu, 2024). CIPHER combines unperturbed-cell covariance with the specified perturbation to predict responses, illustrating the predictive value of baseline cellular structure (Kuznets-Speck et al., 2025). CellAudit asks whether a selected predictor uses the particular perturbation input invoked by its design.

Evaluation of perturbation prediction. PerturbBench and PertEval-scFM standardize predictive comparisons across splits and baselines (Wu et al., 2025; Wenteler et al., 2025). Deep predictors need not outperform linear references, and common metrics can reward systematic variation shared across perturbations (Ahlmann-Eltze et al., 2025; Viñas Torné et al., 2026). Other studies show that well-calibrated predictive metrics and informative foundation-model representations reveal gains

over simple baselines (Miller et al., 2025; Hasanaj et al., 2025; Cole et al., 2026). In-the-wild evaluation further examines context and perturbation shifts (Mao et al., 2026). Together, these studies establish the importance of metrics, representations, and transfer conditions. CELLAUDIT tests a complementary property: whether registered input-use claims are supported from cited source computation through fitted dependence to target-relevant predictive contribution.

Agentic model discovery. Scientific agents span laboratory workflows, program search, and iterative machine-learning experimentation (Boiko et al., 2023; Romera-Paredes et al., 2024; Huang et al., 2024; Li et al., 2024; Lu et al., 2026; Jiang et al., 2025). CellScientist, CellForge, and HarmonyCell apply related search procedures to cellular prediction (Li et al., 2026; Tang et al., 2025; Huang et al., 2026). The existing CELLSCIENTIST policy supplies candidate models in these experiments; CELLAUDIT links their registered input-use claims to deterministic tests of source consumption, fitted dependence, and target-relevant predictive contribution.

Reliance, shortcuts, and automated falsification. Permutation reliance measures performance changes after input disruption; conditional permutations account for dependence among inputs (Fisher et al., 2019; Chamma et al., 2023). Shortcut and underspecification studies show that benchmark success can coexist with unintended or unstable decision rules (D'Amour et al., 2022; Geirhos et al., 2020; Lapuschkin et al., 2019; DeGrave et al., 2021). Attribution sanity checks, removal-based evaluation, and behavioral suites test explanations and capabilities under targeted interventions (Adebayo et al., 2018; Hooker et al., 2019; Ribeiro et al., 2020). ConceptSMILE audits concept-explanation reliability through input perturbations and local surrogate modeling (Mollapour et al., 2026); POPPER tests free-form hypotheses through agentic sequential falsification with statistical error control (Huang et al., 2025). CELLAUDIT follows one selected cellular-response model and its registered input-use claim from the cited code to complete fitted predictions and later-data evaluation. This links source implementation, fitted dependence, and target-loss changes under registered replacements, distinguishing a blocked pathway, sensitivity without predictive benefit, and a contribution that recurs on new data.

L FACTORIAL CONTRASTS AND STABILITY SUMMARIES

An explicit input route establishes a possible computation; its fitted contribution is tested with the checkpoint and observed targets held fixed. We adapt permutation reliance and behavioral testing to the registered biological inputs (Fisher et al., 2019; Ribeiro et al., 2020). For a chemical-response model, the 2×2 test combines the correct or shuffled perturbation with the correct or an alternative control profile. Let m_{pc} denote the response score with both correct inputs, $m_{\tilde{p}c}$ the score after a matched perturbation shuffle, $m_{p\tilde{c}}$ the score after a control-profile swap, and $m_{\tilde{p}\tilde{c}}$ the score after both. For a higher-is-better metric m,

$$E_{\text{pert}} = \tfrac{1}{2}[(m_{pc} - m_{\tilde{p}c}) + (m_{p\tilde{c}} - m_{\tilde{p}\tilde{c}})], \tag{4}$$

$$E_{\text{context}} = \tfrac{1}{2}[(m_{pc} - m_{p\tilde{c}}) + (m_{\tilde{p}c} - m_{\tilde{p}\tilde{c}})], \tag{5}$$

$$I_{\text{pert,context}} = m_{pc} - m_{\tilde{p}c} - m_{p\tilde{c}} + m_{\tilde{p}\tilde{c}}. \tag{6}$$

These are factorial contrasts in predictive score under the registered replacement distribution: E_{pert} averages perturbation-replacement effects across the two tested contexts, while $I_{\text{pert,context}}$ measures their nonadditivity in score. For compound inputs we write E_{cmpd} and E_{ctrl} in tables. The loss contrast $E^{\mathcal{L}}_{\text{cmpd}}$ similarly averages the increase after compound replacement over both context states. Each checkpoint uses 32 fixed permutations constructed without response values or model scores. Continuous effects and their intervals are the primary quantitative evidence. For a per-model stability summary, let τ be the 95th percentile of absolute pairwise differences among the corresponding reference scores. A contribution is *qualified* if its effect's empirical 2.5th percentile exceeds τ, *unsupported* if its 97.5th percentile does not exceed τ, and *inconclusive* otherwise. These labels summarize effect stability relative to map-induced score variation; an unsupported effect may still be positive. The empirical quantiles describe replacement variability within a checkpoint; Appendix B specifies the reference scores and separate across-model uncertainty.

Interpreting the reference-relative criterion. Reference-score variation can include variation induced by the other input. Holding the compound effect fixed while increasing this variation can

therefore change the status. An existing synthetic check separates invariant, sensitive-without-benefit, target-relevant, and harmful predictors. All 96 target-relevant instances have positive mean effects, while 32 of 96 exceed the reference-relative criterion, identically with 32, 128, or 512 maps. The continuous measurements recover predictive benefit, whereas the categorical rule asks whether that benefit exceeds the specified reference variation. Its threshold is an operational effect scale. The additional 6,400-dataset study in Appendix F.1 tests estimator bias, interval coverage, and sensitivity to the reference scale; it distinguishes conditional replacement uncertainty from population uncertainty and retains the observed finite-group undercoverage.

M Observed-support replay and matched reference checks

M.1 Single-coordinate replay on observed support

The frozen replay changes one encoded coordinate while holding the source row's other inputs fixed. Eligible donors satisfy the recorded metadata constraints and produce an actual input change. Table 57 reports the supported populations separately from the original context-averaged factorial summaries. Score-selected BBBC047 remains exactly invariant; the path-constrained and guided panels retain positive compound effects on both boundaries. On LKCP, dose intervals are positive and compound intervals include zero. For the replay and real-source audit below, 95% intervals resample source groups while conditioning on the selected checkpoint set, donor pools, and 128 replacement maps; BBBC and LINCS use Murcko source scaffolds, and LKCP uses registered BRD identifier groups. The matched-reference and control-bank intervals instead describe paired training-seed variation at the fixed task split. Individual intervals and positive-interval counts are descriptive and unadjusted.

Table 57: Observed-support, fixed-context single-coordinate effects. Positive PCC drop and target-loss increase indicate benefit from the original input relative to its replacement. Cells report mean and 95% source-group interval.

Panel	Fold	Input	**Supported/all rows** source groups	**PCC drop** Δm_q	**Target-loss increase** $\Delta \mathcal{L}_q$
BBBC047: score-selected	4	Compound	3341/4380	0	0
			838 groups	$[0, 0]$	$[0, 0]$
BBBC047: score-selected	5	Compound	2927/3878	0	0
			855 groups	$[0, 0]$	$[0, 0]$
BBBC047: path-constrained	4	Compound	3341/4380	0.01025	3.13×10^{-4}
			838 groups	$[0.00654, 0.01425]$	$[1.95, 4.59] \times 10^{-4}$
BBBC047: path-constrained	5	Compound	2927/3878	0.00654	2.08×10^{-4}
			855 groups	$[0.00236, 0.01097]$	$[0.75, 3.51] \times 10^{-4}$
BBBC047: guided	4	Compound	3341/4380	0.00534	1.84×10^{-4}
			838 groups	$[0.00336, 0.00770]$	$[1.14, 2.73] \times 10^{-4}$
BBBC047: guided	5	Compound	2927/3878	0.00404	1.44×10^{-4}
			855 groups	$[0.00179, 0.00634]$	$[0.64, 2.28] \times 10^{-4}$
LINCS: guided	4	Compound	843/847	0.01996	5.95×10^{-3}
			219 groups	$[0.01276, 0.02863]$	$[3.71, 8.80] \times 10^{-3}$
LINCS: guided	4	Dose	847/847	0.05096	1.52×10^{-2}
			221 groups	$[0.04419, 0.05778]$	$[1.28, 1.77] \times 10^{-2}$
LINCS: guided	5	Compound	864/868	0.01446	3.83×10^{-3}
			235 groups	$[0.00816, 0.02113]$	$[2.22, 5.64] \times 10^{-3}$
LINCS: guided	5	Dose	867/868	0.04853	1.28×10^{-2}
			236 groups	$[0.04158, 0.05586]$	$[1.08, 1.50] \times 10^{-2}$
LKCP: guided	4	Compound	351/405	0.00043	1.19×10^{-3}
			13 groups	$[-0.00149, 0.00239]$	$[-4.18, 6.24] \times 10^{-3}$
LKCP: guided	4	Dose	405/405	0.02407	6.61×10^{-2}
			15 groups	$[0.01154, 0.04369]$	$[3.14, 11.41] \times 10^{-2}$
LKCP: guided	5	Compound	324/432	0.00321	1.19×10^{-2}
			12 groups	$[-0.00049, 0.00728]$	$[-0.40, 3.27] \times 10^{-2}$
LKCP: guided	5	Dose	432/432	0.03537	1.21×10^{-1}
			16 groups	$[0.02124, 0.04906]$	$[0.50, 2.14] \times 10^{-1}$

Original-entry-weighted means: score-selected BBBC047 has 5 entries per fold; path-constrained and guided BBBC047 have 10 each; LINCS and LKCP have 50 each. LKCP's 50 entries contain 30 unique weight hashes. Path-constrained and guided checkpoints are fixed across the two boundary evaluations. BBBC047 has no registered observed-support dose contrast.

M.2 COMPLETE MATCHED INPUT-REFERENCE FAMILY

The four input subsets use the same masked-MLP family, loss, fit budget, and five paired training seeds. Each selected checkpoint is evaluated on both held-out folds. Here c denotes context, p compound representation, and a dose. Dose improves the LINCS reference; its increment is small on BBBC and LKCP. On LKCP Fold 5, the full reference has a positive PCC increment over $g(c, a)$, while its MSE interval includes zero. These trained-subset comparisons describe the fixed reference family; within-checkpoint replacement measures use in the separately selected agent models.

Table 58: Complete four-input masked-MLP references. Each cell gives mean Global PCC (top; higher is better) and MSE (bottom; lower is better) across five paired seeds.

Cohort	Fold	$g(c)$	$g(c, a)$	$g(c, p)$	$g(c, p, a)$
BBBC036	4	0.3112 0.058572	0.3108 0.058582	0.2981 0.059153	0.2981 0.059154
BBBC036	5	0.2928 0.059025	0.2922 0.059035	0.2708 0.059966	0.2707 0.059968
BBBC047	4	0.3033 0.045777	0.3031 0.045781	0.2736 0.046620	0.2737 0.046619
BBBC047	5	0.3145 0.045252	0.3144 0.045257	0.2840 0.046161	0.2840 0.046159
LINCS	4	0.1702 0.731582	0.1912 0.725847	0.1636 0.733309	0.1643 0.733138
LINCS	5	0.1721 0.590465	0.1917 0.586218	0.1631 0.593010	0.1638 0.592862
LKCP	4	0.1855 2.849319	0.1856 2.850595	0.2031 2.998213	0.2081 3.020705
LKCP	5	0.2050 4.550455	0.2056 4.548941	0.2525 4.510473	0.2420 4.536487

Paired increments and their 95% intervals appear in Table 59.

Table 59: Paired input-reference increments (mean and 95% training-seed interval). Differences are the first predictor minus the second; positive PCC and negative MSE favor the first predictor.

Cohort	Fold	$g(c, a) - g(c)$ PCC	$g(c, a) - g(c)$ MSE	$g(c, p, a) - g(c, a)$ PCC	$g(c, p, a) - g(c, a)$ MSE
BBBC036	4	-0.00035 $[-0.00127, 0.00058]$	1.02×10^{-5} $[-3.25, 5.30] \times 10^{-5}$	-0.01279 $[-0.01490, -0.01068]$	5.71×10^{-4} $[4.74, 6.69] \times 10^{-4}$
BBBC036	5	-0.00052 $[-0.00115, 0.00011]$	9.87×10^{-6} $[-1.47, 21.21] \times 10^{-6}$	-0.02150 $[-0.02258, -0.02041]$	9.33×10^{-4} $[8.51, 10.15] \times 10^{-4}$
BBBC047	4	-0.00011 $[-0.00016, -0.00006]$	3.86×10^{-6} $[2.48, 5.23] \times 10^{-6}$	-0.02948 $[-0.03342, -0.02553]$	8.38×10^{-4} $[7.45, 9.31] \times 10^{-4}$
BBBC047	5	-0.00012 $[-0.00016, -0.00008]$	4.89×10^{-6} $[3.16, 6.62] \times 10^{-6}$	-0.03039 $[-0.03351, -0.02726]$	9.03×10^{-4} $[8.03, 10.02] \times 10^{-4}$
LINCS	4	0.02095 $[0.02026, 0.02164]$	-5.73×10^{-3} $[-5.95, -5.52] \times 10^{-3}$	-0.02687 $[-0.02779, -0.02595]$	7.29×10^{-3} $[7.05, 7.54] \times 10^{-3}$
LINCS	5	0.01954 $[0.01899, 0.02009]$	-4.25×10^{-3} $[-4.33, -4.16] \times 10^{-3}$	-0.02790 $[-0.02887, -0.02693]$	6.64×10^{-3} $[6.35, 6.93] \times 10^{-3}$
LKCP	4	0.00004 $[-0.00050, 0.00057]$	1.28×10^{-3} $[-1.88, 4.44] \times 10^{-3}$	0.02256 $[-0.00440, 0.04952]$	1.70×10^{-1} $[-0.70, 4.10] \times 10^{-1}$
LKCP	5	0.00061 $[0.00016, 0.00107]$	-1.51×10^{-3} $[-2.53, -0.49] \times 10^{-3}$	0.03637 $[0.00366, 0.06908]$	-1.25×10^{-2} $[-10.17, 7.68] \times 10^{-2}$

M.3 LINKED SOURCE AND BEHAVIOR IN REAL CANDIDATES

A fixed, dataset–policy-balanced and source-structure-stratified sample links 48 source records to 96 same-checkpoint boundary evaluations. Implemented source consumption coexists with distinct fitted outcomes: an implemented `MultiTowerFusion` source is exactly invariant on both boundaries, whereas 47 sources change predictions and 20 have positive target-loss intervals on both boundaries. Unresolved source locations also admit complete-model behavioral tests.

M.4 CONTROL-REFERENCE REUSE ON SCI-PLEX

The existing control-bank comparison holds the response target $Y - R$ fixed and changes whether the context input reuses reference bank R or uses the other bank. Reference A/B assignment is

Table 60: Descriptive source–behavior counts in 48 real candidates. Each case contributes its paired boundary pattern once.

Cohort	Source status	Cases	**Invariant** both folds	**Sensitive; loss CI not** positive on ≥ 1 fold	**Positive loss CI** both folds
BBBC036	Implemented	22	0	22	0
BBBC036	Unresolved	2	0	2	0
BBBC047	Implemented	21	1	1	19
BBBC047	Unresolved	3	0	2	1
Total		48	1	27	20

The middle behavioral category has positive prediction distance but lacks a positive target-loss interval on at least one fold. Source status describes the cited computation. Counts describe this structurally sampled set and its scripted source classifications. Claim origin remains `task_contract` where no candidate-authored manifest was emitted.

fixed before fitting and counterbalanced within cell-line, replicate, and control-layout strata (24 plate groups per orientation). Three input sets, two bank assignments, and five paired seeds give 30 fitted checkpoints. All six shared-minus-disjoint PCC intervals include zero. This comparison measures control-reference reuse on sci-Plex.

Table 61: sci-Plex control-reference reuse: shared-minus-disjoint Global PCC (mean and paired-seed 95% interval), with identical centered targets in both arms.

Predictor	Fold 4	Fold 5
$g(c)$	0.00092 $[-0.00215, 0.00399]$	0.00053 $[-0.00136, 0.00241]$
$g(c, a)$	-0.00030 $[-0.00336, 0.00277]$	-0.00006 $[-0.00473, 0.00462]$
$g(c, p, a)$	0.00511 $[-0.00241, 0.01263]$	0.00402 $[-0.01467, 0.02271]$

N MATCHED FEEDBACK: PREDICTIVE OUTCOMES AND INPUT CONTRIBUTION

Table 62: Frozen predictors after score or audit feedback on sci-Plex. Values are mean $\pm$ trajectory SD ($n = 5$).

Predictor	**F4 PCC** $\uparrow$	**F4 MSE** $\downarrow$	**F5 PCC** $\uparrow$	**F5 MSE** $\downarrow$
Anchor $g(c, a)$	0.9664 ± 0.0001	0.0308 ± 0.0001	0.9728 ± 0.0002	0.0251 ± 0.0001
Score	0.9756 ± 0.0027	0.0225 ± 0.0025	0.9764 ± 0.0015	0.0219 ± 0.0013
Audit	0.9779 ± 0.0004	0.0204 ± 0.0004	0.9778 ± 0.0002	0.0205 ± 0.0002

The anchor is shared within each seed. F4 and F5 are held-out combination boundaries of the sci-Plex development source.

Table 63: Target-relevant input use after feedback. Mean $\pm$ SD across five endpoints.

Set	**Feedback**	**Chemical** PCC drop	**Chemical** loss gain	**Dose** PCC drop	**Dose** loss gain
F4	Score	0.0203 ± 0.0046	0.0186 ± 0.0042	0.0068 ± 0.0026	0.0062 ± 0.0024
F4	Audit	0.0240 ± 0.0016	0.0220 ± 0.0015	0.0090 ± 0.0006	0.0082 ± 0.0006
F5	Score	0.0070 ± 0.0016	0.0064 ± 0.0015	0.0039 ± 0.0017	0.0035 ± 0.0015
F5	Audit	0.0086 ± 0.0005	0.0079 ± 0.0005	0.0049 ± 0.0004	0.0045 ± 0.0004

Positive values favor the correct input over its legal replacement. Each displayed coordinate metric is positive in 5/5 endpoints in both arms at both boundaries.

N.1 SCORE–AUDIT FEEDBACK COMPARISON: COMPLETE RECORDS

Comparison and evaluation. We analyze Score and Audit feedback using all five paired discovery trajectories and their shared context-plus-dose anchors. The descriptive comparison holds model space, trainer, selection rule, and candidate budget fixed, and reports every selected endpoint, paired effect, and uncertainty interval. Within the sci-Plex development source, compound–cell-line roles were reassigned and frozen before response preprocessing. F4 and F5 evaluate the same selected checkpoints on held-out combinations.

Table 64: Audit minus Score: paired predictive effects and two uncertainty summaries. Effect and interval entries are in 10^{-3} units.

Set	Metric	Mean Δ	Paired t 95% CI	Joint bootstrap 95% CI	Wins
F4	PCC	+2.364	[-1.277, +6.005]	[+0.253, +5.619]	4/5
F4	MSE	-2.159	[-5.482, +1.164]	[-5.136, -0.234]	4/5
F5	PCC	+1.460	[-0.496, +3.416]	[+0.202, +3.021]	3/5
F5	MSE	-1.333	[-3.116, +0.450]	[-2.748, -0.187]	3/5

Descriptive paired comparison. Paired t: five trajectories, df=4. Joint bootstrap: 4,000 shared trajectory–compound draws. Wins favor higher PCC or lower MSE. Both uncertainty summaries retain all five trajectory pairs.

Table 65: Discovery outcomes and realized resource use under matched candidate and training opportunities.

Feedback	Selected PCC	Valid	Failed	Repairs	LLM calls	Reported tokens	Training wall s
Score	0.9757	25/25	0	11	36	$\geq$286,995	194.93
Audit	0.9781	25/25	0	13	38	592,347	217.41

Selected PCC averages the five discovery-selection scores. Calls include repairs; the Score token total is a lower bound with one missing usage receipt. Five shared anchors contribute a further 32.61 training seconds, counted once.

Data and splitting. The fit, selection, F4, and F5 packages contain 2,666, 891, 438, and 442 treatment wells. Complete compound $\times$ cell-line groups, including their doses and replicate wells, remain within one role; evaluated compounds and cell lines each appear among the fit marginals. The 2,000 target genes are selected by variance among fit treatment wells, with fit-only feature scaling. Control bank A supplies inputs and bank B the reference; they contain distinct wells matched on plate, cell line, and replicate, with control pools reused across these development roles. The target is the complete gene-expression profile.

Feedback and model revision. Score supplies selection metrics, sampled learning curves, model size, timing, and past designs. Audit adds grouped errors, component-source clues, and chemical/dose replacement effects. This comparison estimates the complete enhanced-feedback package. Both arms share $g(c, a)$ within seed, five new candidate slots, and three repair opportunities per slot. The language model may revise encoders, fusion, readout, capacity, regularization, and residual scale; a direct predictor or frozen-anchor residual is equally available. Each new component is trained from scratch. Completed selection diagnostics inform subsequent proposals; final checkpoints are frozen before F4/F5 evaluation.

Fixed training and selection. AdamW uses learning rate 10^{-3}, weight decay 10^{-5}, batch size 128, gradient clipping at 5, a 100-epoch maximum, and early-stopping patience 15. The shared loss is fit-standardized target MSE plus $0.02(1 - \mathrm{PCC})$, using flattened Pearson correlation. The common limit is five million trainable parameters. Epoch and endpoint selection prioritize selection Global PCC, then lower MSE; endpoint ties additionally prefer fewer parameters and earlier slots. The anchor stays frozen in residual models.

Uncertainty and continuous evidence. SD uses denominator $n - 1$ across five trajectories; paired two-sided Student-t intervals use df=4. Compound-only bootstrap fixes the fitted trajectories, trajectory-only bootstrap fixes the evaluated compounds, and joint bootstrap resamples both; each uses 4,000 shared-index draws. Appendix N.2 compares these intervals with complete leave-one-pair-out and exact sign-flip analyses. Maps are averaged within endpoints before trajectory inference. RMS measures sensitivity; PCC drop is correct-input minus replacement PCC, and loss gain is replacement minus correct-input MSE. Positive continuous effects are distinct from individual qualification or biological-mechanism evidence.

Model identifier and indexing. Recorded model identifier: `deepseek-v41`. All compared calls use this serving identity. Trajectory indices 1–5 correspond, in order, to seeds 2026091201, 2026091202, 2026091203, 2026091204, 2026091205.

Ratios and resources. Relative MSE reduction against the anchor is computed within seed as $1 - \mathrm{MSE}_{\mathrm{endpoint},s}/\mathrm{MSE}_{\mathrm{anchor},s}$ and then averaged. A ratio of arm mean MSEs is reported separately

when used. Costs retain failed attempts and recorded interruptions. Shared anchor training totals 32.611546 seconds and is counted once, separately from each condition's candidate-training costs. Trainer and worker wall times overlap; missing token usage produces a lower bound.

Table 66: Every frozen endpoint at both held-out boundaries: all five paired trajectories and their shared anchors.

Tr.	Model	Slot	PCC	MSE	C-R	C-P	C-L	D-R	D-P	D-L
Fold 4										
1	Anchor	0	0.96652	30.768	0.000	0.000	0.000	28.653	1.152	1.042
1	Score	5	0.97557	22.551	141.301	20.580	18.805	84.343	6.798	6.212
1	Audit	5	0.97806	20.278	142.171	23.162	21.173	94.176	8.838	8.081
2	Anchor	0	0.96646	30.836	0.000	0.000	0.000	25.007	0.985	0.896
2	Score	3	0.97740	20.889	136.127	21.774	19.882	84.750	7.580	6.923
2	Audit	4	0.97809	20.252	155.045	25.294	23.142	101.848	9.453	8.650
3	Anchor	0	0.96647	30.813	0.000	0.000	0.000	31.861	1.311	1.189
3	Score	3	0.97570	22.432	138.732	20.644	18.823	96.003	7.982	7.278
3	Audit	3	0.97817	20.181	138.099	22.698	20.763	91.844	8.613	7.880
4	Anchor	0	0.96646	30.820	0.000	0.000	0.000	26.818	1.066	0.960
4	Score	5	0.97807	20.274	157.142	25.704	23.535	96.377	9.070	8.307
4	Audit	5	0.97725	21.023	145.497	22.784	20.858	90.603	8.193	7.502
5	Anchor	0	0.96629	30.977	0.000	0.000	0.000	17.193	0.640	0.580
5	Score	4	0.97115	26.587	117.806	13.009	11.887	48.269	2.392	2.186
5	Audit	2	0.97815	20.203	159.745	26.108	23.941	102.759	9.697	8.893
Fold 5										
1	Anchor	0	0.97265	25.253	0.000	0.000	0.000	27.811	0.712	0.647
1	Score	5	0.97586	22.332	100.896	7.672	7.048	65.922	3.595	3.306
1	Audit	5	0.97785	20.495	90.349	8.720	7.978	71.763	5.036	4.612
2	Anchor	0	0.97297	24.955	0.000	0.000	0.000	24.562	0.606	0.553
2	Score	3	0.97777	20.567	80.270	7.558	6.905	71.301	4.866	4.452
2	Audit	4	0.97772	20.619	90.797	8.556	7.835	74.695	4.693	4.303
3	Anchor	0	0.97297	24.957	0.000	0.000	0.000	31.737	0.839	0.764
3	Score	3	0.97626	21.947	92.154	7.588	6.919	75.540	4.604	4.202
3	Audit	3	0.97815	20.230	86.705	8.379	7.691	71.783	4.883	4.487
4	Anchor	0	0.97270	25.230	0.000	0.000	0.000	26.502	0.688	0.621
4	Score	5	0.97771	20.623	89.824	8.217	7.526	77.179	5.209	4.776
4	Audit	5	0.97758	20.744	88.060	7.995	7.320	70.291	4.508	4.131
5	Anchor	0	0.97292	24.999	0.000	0.000	0.000	17.108	0.399	0.363
5	Score	4	0.97425	23.789	75.832	4.122	3.769	40.870	1.079	0.988
5	Audit	2	0.97786	20.504	99.006	9.405	8.635	83.693	5.505	5.060

PCC is unscaled; all other numeric outcomes are in 10^{-3} units. C/D: chemical/dose; R: prediction RMS; P: PCC drop; L: target-loss gain. Slot 0 is the shared anchor. RMS has no preferred direction.

Table 67: All ten discovery trajectories and their realized costs.

Tr.	Arm	Slot	Sel. PCC	Repair	Calls	HTTP	Tokens	Train s	Work s	Diag. s
1	Score	5	0.97491	1	6	7	$\geq$50,685	48.09	67.26	7.10
1	Audit	5	0.97828	2	7	7	106,239	41.24	60.03	12.10
2	Score	3	0.97757	3	8	8	62,703	41.96	60.07	7.50
2	Audit	4	0.97783	3	8	8	125,047	42.32	61.01	12.94
3	Score	3	0.97628	4	9	9	70,709	40.04	58.06	7.36
3	Audit	3	0.97825	3	8	8	136,386	44.25	61.58	12.15
4	Score	5	0.97852	3	8	8	68,757	35.40	52.75	8.73
4	Audit	5	0.97798	3	8	8	129,227	39.60	55.93	12.24
5	Score	4	0.97145	0	5	5	34,141	29.44	46.61	6.75
5	Audit	2	0.97830	2	7	7	95,448	50.00	67.66	11.96

Every trajectory completed 5/5 new candidates and had zero fully failed slots. Train is known completed trainer wall time; Work is candidate worker wall time; Diag. is postfit diagnostic wall time. Logical calls and physical HTTP attempts are distinct.

Table 68: All eight Audit–Score outcome contrasts at both boundaries. Effects and intervals are in 10^{-3} units.

Set	Metric	Mean Δ	Paired 95% CI	+/–/0
F4	PCC	+2.364	[-1.277, +6.005]	4/1/0
F4	MSE	-2.159	[-5.482, +1.164]	1/4/0
F4	Chem. RMS	+9.890	[-16.189, +35.969]	3/2/0
F4	Chem. PCC drop	+3.667	[-3.577, +10.912]	4/1/0
F4	Chem. loss gain	+3.389	[-3.269, +10.048]	4/1/0
F4	Dose RMS	+14.298	[-16.033, +44.628]	3/2/0
F4	Dose PCC drop	+2.194	[-1.639, +6.028]	4/1/0
F4	Dose loss gain	+2.020	[-1.496, +5.536]	4/1/0
F5	PCC	+1.460	[-0.496, +3.416]	3/2/0
F5	MSE	-1.333	[-3.116, +0.450]	2/3/0
F5	Chem. RMS	+3.189	[-13.712, +20.089]	2/3/0
F5	Chem. PCC drop	+1.580	[-1.069, +4.228]	4/1/0
F5	Chem. loss gain	+1.458	[-0.979, +3.896]	4/1/0
F5	Dose RMS	+8.283	[-16.535, +33.100]	3/2/0
F5	Dose PCC drop	+1.054	[-1.483, +3.592]	3/2/0
F5	Dose loss gain	+0.974	[-1.355, +3.303]	3/2/0

Descriptive effect estimates over all five paired trajectories. Signs count positive, negative, and exactly zero differences, not wins. Lower MSE is favorable; RMS has no favorable direction.

Table 69: Predictive uncertainty under the three original sampling schemes; interval entries in 10^{-3} units.

Set	Metric	Paired t CI	Compound-only CI	Joint CI
F4	PCC	[-1.277, +6.005]	[+1.300, +3.704]	[+0.253, +5.619]
F4	MSE	[-5.482, +1.164]	[-3.389, -1.197]	[-5.136, -0.234]
F5	PCC	[-0.496, +3.416]	[+1.006, +1.975]	[+0.202, +3.021]
F5	MSE	[-3.116, +0.450]	[-1.800, -0.924]	[-2.748, -0.187]

Bootstrap intervals use 4,000 draws. No coordinate-effect bootstrap interval is substituted from predictive statistics, and no new confirmatory tests are introduced.

Table 70: Legal replacement support shared by all compared endpoints.

Set	Coordinate	Eligible wells	Maps	Unique	First 32 unique
F4	Chemical	438/438	128	128	32
F4	Dose	428/438	128	1	1
F5	Chemical	434/442	128	128	32
F5	Dose	436/442	128	1	1

Replacements match plate, cell line, replicate, and treatment time. Chemical replacement holds dose fixed; dose replacement holds compound fixed and requires at least 0.1 log-dose difference. No cross-group fallback is used. Discovery used 32 maps; final evaluation used 128, with only one distinct dose map. Unsupported wells are excluded, not assigned zero effect.

N.2 Robustness across the complete five trajectory pairs

All five matched feedback trajectories. This retrospective sensitivity analysis retains seeds 2026091201–2026091205 and all eight metrics at both boundaries. Audit minus Score is computed within each selected-trajectory pair; F4 and F5 reuse the same selected checkpoints and do not make $N = 10$. Means and descriptive two-sided 95% t_4 intervals use all five pairs. All 32 sign assignments are enumerated for the absolute unstudentized mean under a pair-exchangeability/sign-symmetry null, not a randomized-assignment guarantee. The smallest attainable two-sided tail probability is $2/32 = .0625$; no multiplicity-adjusted confirmatory claim is made.

Why inference methods disagree. Trajectory-only percentile bootstrap and paired-t inference target the same conditional mean paired contrast; their different conclusions reflect the five-point empirical bootstrap distribution versus Student-t sampling assumptions, not different estimands. Compound-only and joint intervals additionally have different resampling scopes. All predictive bootstrap intervals exclude zero, whereas the paired-t intervals and exact sign-flip tails do not provide the same evidence. Increasing bootstrap draws does not add independent trajectories or establish broad generalization.

Table 71: All five-pair Audit-minus-Score endpoints. Numerical effects and intervals are $\times 10^3$.

Fold	Metric	Mean	Paired-t 95% CI	Median	$+/-/0$	Tail
F4	Global PCC	2.364	$[-1.277, 6.005]$	2.465	4/1/0	6/32
F4	MSE	-2.159	$[-5.482, 1.164]$	-2.251	1/4/0	6/32
F4	Chemical RMS	9.890	$[-16.189, 35.969]$	0.870	3/2/0	12/32
F4	Chemical loss gain	3.389	$[-3.269, 10.048]$	2.368	4/1/0	8/32
F4	Chemical PCC drop	3.667	$[-3.577, 10.912]$	2.582	4/1/0	8/32
F4	Dose RMS	14.298	$[-16.033, 44.628]$	9.833	3/2/0	10/32
F4	Dose loss gain	2.020	$[-1.496, 5.536]$	1.727	4/1/0	6/32
F4	Dose PCC drop	2.194	$[-1.639, 6.028]$	1.873	4/1/0	6/32
F5	Global PCC	1.460	$[-0.496, 3.416]$	1.888	3/2/0	8/32
F5	MSE	-1.333	$[-3.116, 0.450]$	-1.716	2/3/0	8/32
F5	Chemical RMS	3.189	$[-13.712, 20.089]$	-1.763	2/3/0	24/32
F5	Chemical loss gain	1.458	$[-0.979, 3.896]$	0.930	4/1/0	4/32
F5	Chemical PCC drop	1.580	$[-1.069, 4.228]$	0.999	4/1/0	4/32
F5	Dose RMS	8.283	$[-16.535, 33.100]$	3.394	3/2/0	20/32
F5	Dose loss gain	0.974	$[-1.355, 3.303]$	0.285	3/2/0	12/32
F5	Dose PCC drop	1.054	$[-1.483, 3.592]$	0.279	3/2/0	12/32

Tail is the exact sign-flip count out of 32. All 16 paired-t intervals contain zero and all exact tails are at least .125. RMS measures sensitivity; target-loss gain and PCC drop measure changes in loss and correlation.

Table 72: All percentile-bootstrap 95% intervals ($\times 10^3$).

Fold	Metric	Compound only	Trajectory only	Joint
F4	Global PCC	$[1.300, 3.704]$	$[0.144, 4.828]$	$[0.253, 5.619]$
F4	MSE	$[-3.389, -1.197]$	$[-4.408, -0.132]$	$[-5.136, -0.234]$
F4	Chemical RMS	$[6.425, 13.160]$	$[-4.738, 28.126]$	$[-5.652, 27.694]$
F4	Chemical loss gain	$[2.032, 5.060]$	$[-0.482, 8.094]$	$[-0.456, 8.402]$
F4	Chemical PCC drop	$[2.201, 5.452]$	$[-0.534, 8.787]$	$[-0.508, 9.067]$
F4	Dose RMS	$[4.676, 22.279]$	$[-2.007, 34.959]$	$[-3.228, 38.642]$
F4	Dose loss gain	$[0.689, 3.753]$	$[0.039, 4.490]$	$[0.066, 5.631]$
F4	Dose PCC drop	$[0.750, 4.083]$	$[0.028, 4.884]$	$[0.068, 6.187]$
F5	Global PCC	$[1.006, 1.975]$	$[0.302, 2.618]$	$[0.202, 3.021]$
F5	MSE	$[-1.800, -0.924]$	$[-2.392, -0.274]$	$[-2.748, -0.187]$
F5	Chemical RMS	$[-1.073, 7.471]$	$[-6.751, 14.920]$	$[-7.875, 15.416]$
F5	Chemical loss gain	$[0.824, 2.180]$	$[0.248, 3.260]$	$[0.143, 3.537]$
F5	Chemical PCC drop	$[0.892, 2.364]$	$[0.266, 3.528]$	$[0.146, 3.843]$
F5	Dose RMS	$[1.196, 15.103]$	$[-3.716, 25.621]$	$[-5.151, 28.995]$
F5	Dose loss gain	$[0.324, 1.866]$	$[-0.261, 2.557]$	$[-0.253, 3.122]$
F5	Dose PCC drop	$[0.349, 2.021]$	$[-0.294, 2.768]$	$[-0.284, 3.397]$

4000 draws, seed 69313. Compound weights are shared across all paired arms and trajectories; trajectory draws resample the five paired indices. Joint draws resample both units. The 128 donor maps are fixed. Sufficient statistics are pooled before PCC; compound PCCs are not averaged. Compound-only intervals condition on five selected trajectories; trajectory-only intervals condition on the observed compound set, while joint intervals vary both empirical units.

Verification boundary. Retained full predictions independently verify global PCC/MSE and eligible-source correct statistics. Full counterfactual prediction arrays were not retained for this legacy feedback package, so coordinate reconstruction verifies stored sufficient statistics but is not a new independent full-array counterfactual check. These additional descriptions do not reselect checkpoints, retrain models, or revise historical qualification labels.

Table 73: Fifth-pair contribution and complete leave-one-pair-out mean ranges ($\times 10^3$). No pair is removed from the primary summary.

Fold	Metric	All-five mean	**Fifth-pair** $d_5/5$	**Omit-fifth** mean	All five LOO means
F4	Global PCC	2.364	1.399	1.206	[1.206, 3.160]
F4	MSE	−2.159	−1.277	−1.103	[−2.886, −1.103]
F4	Chemical RMS	9.890	8.388	1.877	[1.877, 15.273]
F4	Chemical loss gain	3.389	2.411	1.223	[1.223, 4.906]
F4	Chemical PCC drop	3.667	2.620	1.309	[1.309, 5.314]
F4	Dose RMS	14.298	10.898	4.249	[4.249, 19.315]
F4	Dose loss gain	2.020	1.341	0.848	[0.848, 2.726]
F4	Dose PCC drop	2.194	1.461	0.916	[0.916, 2.962]
F5	Global PCC	1.460	0.722	0.922	[0.922, 1.859]
F5	MSE	−1.333	−0.657	−0.845	[−1.697, −0.845]
F5	Chemical RMS	3.189	4.635	−1.808	[−1.808, 6.622]
F5	Chemical loss gain	1.458	0.973	0.606	[0.606, 1.875]
F5	Chemical PCC drop	1.580	1.057	0.654	[0.654, 2.030]
F5	Dose RMS	8.283	8.565	−0.352	[−0.352, 12.075]
F5	Dose loss gain	0.974	0.815	0.199	[0.199, 1.379]
F5	Dose PCC drop	1.054	0.885	0.211	[0.211, 1.493]

The fifth-pair column is its contribution to the five-pair mean, not its whole paired difference. F5 chemical and dose RMS have positive all-five means but negative omit-fifth means; the F5 chemical-RMS median is negative, whereas the dose-RMS median is positive. All five pairs remain in the primary estimate.

O RECORDED FEEDBACK, SOURCE REVISIONS, AND FROZEN ENDPOINTS

The two cases are the first two registered trajectory seeds (2026091201 and 2026091202), selected by registration order. Each retains the common initial predictor and all five new audit-feedback proposals. Parent links denote design ancestry: every new candidate is trained from scratch, while the shared context-plus-dose predictor $g(c, a)$ is frozen. This anchor concatenates 2,000 context features with dose and uses a 2001–256–256–2000 MLP with 1,092,816 trainable parameters. Local metrics describe the development-selection partition.

What the first revision actually changes. In the first trajectory, slot 2 is the incumbent after its child slot 3 fails to improve selection PCC. Before slot 4, the actual prompt contains slot 2's best epoch (5), sampled learning curve, group-error diagnostics, source-check findings, and positive compound and dose replacement effects. The model hypothesizes overfitting and proposes reducing encoder capacity; compound memorization is a proposed explanation, not a separately measured biological or learning mechanism. Inspection of the emitted source verifies that each chemical/context encoder changes from two Linear layers to one, with dropout 0.1 added in the encoders and fusion. Dose-conditioned FiLM and the rank-128 bilinear interaction remain. Trainable parameters decrease from 4,056,400 to 2,754,896; total parameters including the fixed anchor are 5,149,216 and 3,847,712. Slot 2 versus slot 4 selection PCC is 0.97808647 versus 0.97818047; MSE is 0.02028296 versus 0.02019429; compound target-loss gain is 0.01846193 versus 0.01949683. The static checker reports syntactic return-dependency hints while leaving functional usefulness unresolved; fixed-model input replacement supplies the separate behavioral evidence.

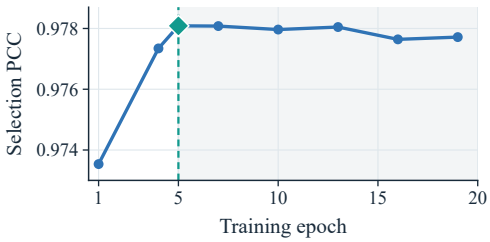
Initialization and the final endpoint. Both first-case sources zero-initialize their output readout. Under the fixed residual wrapper this initializes the newly fitted predictor at $g(c, a)$. Slot 5 adds a dose-conditioned scalar interaction gate and an extra projected chemical feature before fusion, then wins by selection PCC. The emitted `chem_direct` feature enters the nonlinear fusion, whereas its proposal calls it a direct additive target-space path. Slot 5 has higher selection PCC but lower compound target-loss gain than slot 4; selection and input contribution therefore remain distinct. The same frozen slot-5 checkpoint is subsequently evaluated on the two held-out sci-Plex partitions.

The second registered trajectory retains non-improvements. For seed 2026091202, slot 2 decreases selection PCC from its slot-1 parent (0.977817 to 0.977703). Slot 3 improves on slot 2 but remains below slot 1. Slot 4 explicitly returns to slot 1 as its design parent and is selected at PCC 0.977826; its MSE (0.02053288) is slightly worse than slot 1's (0.02052894), consistent with PCC-

(a) Recorded design ancestry



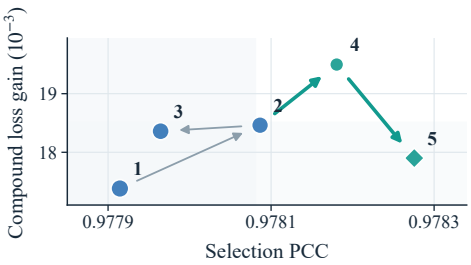
(b) Training feedback received



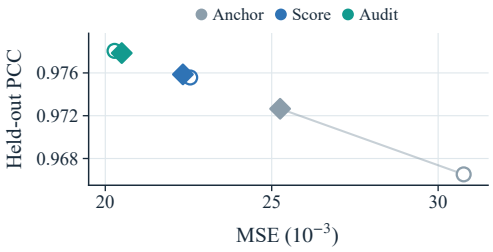
(c) Verified revision: slot 2 to 4

	Slot 2	Slot 4
Compound layers	2	**1**
Context layers	2	**1**
Dropout	0	**0.1**
Bilinear rank	128	128
Dose FiLM	Yes	Yes
Trainable (M)	4.056	**2.755**
Selection MSE	0.020283	**0.020194**

(d) Score and compound contribution



(e) Frozen held-out predictions



(f) Selected checkpoint contributions

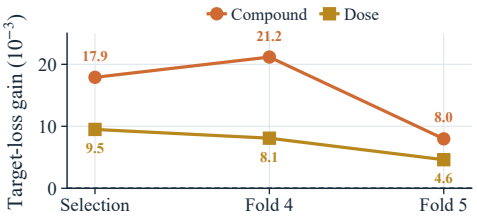


Figure 7: **A recorded revision from received feedback to frozen-model evaluation.** The first registered trajectory supplies (a) all parent links and selection scores, (b) the slot-2 curve and best epoch received before slot 4, and (c) verified source edits, parameter counts and selection MSE; shading marks changed entries. (d) All five new candidates: arrows show ancestry, marker area scales with trainable parameters, and the diamond denotes selected slot 5. (e) The matched anchor, Score and Audit checkpoints on Fold 4 (open circles) and Fold 5 (filled diamonds). (f) Compound and dose loss gains of the same selected slot-5 checkpoint. These are individual recorded measurements; both held-out partitions belong to the sci-Plex development source.

first selection. Slot 5 adds explicit pairwise projections but falls to PCC 0.977506 and is not selected. Its rationale claims that zero-initialized new terms preserve the parent, but the shared from-scratch training contract does not inherit fitted parent weights: this claim is preserved as an agent statement in the recorded proposal, not endorsed as a measured guarantee.

Recorded execution repairs. The first case uses one JSON-contract repair at each of slots 4 and 5; the second uses one at slot 4 and two at slot 5. These logged failures are completion-token-limit/JSON-contract failures. Lower-scoring, successfully trained candidates are retained without score-driven repair.

Table 74: Complete records of the first two registered audit-feedback trajectories. PCC and MSE use the selection partition; chemical gain is the selection-set MSE increase after compound replacement. Bold PCC identifies the selected endpoint within each trajectory.

Slot	Parent	Recorded design change	PCC	MSE	Chemical gain	Repairs
Trajectory 2026091201						
0	–	Common context-plus-dose predictor	0.968005	0.029461	0.000000	0
1	0	Chemical residual with dose FiLM and multiplicative fusion	0.977915	0.020443	0.017382	0
2	1	Rank-128 bilinear interaction; 256-dimensional encoders	0.978086	0.020283	0.018462	0
3	2	Four rank-64 interaction heads and dose gates	0.977964	0.020405	0.018360	0
4	2	Single-layer encoders and dropout 0.1	0.978180	0.020194	0.019497	1
5	4	Scalar interaction gate; extra chemical feature before fusion	**0.978276**	0.020115	0.017901	1
Trajectory 2026091202						
0	–	Common context-plus-dose predictor	0.967972	0.029506	0.000000	0
1	0	Gated multiplicative residual; scale 0.2	0.977817	0.020529	0.016928	0
2	1	Low-rank chemical FiLM and bilinear context interaction	0.977703	0.020634	0.018378	0
3	2	Triple-product fusion plus auxiliary FiLM/bilinear terms	0.977740	0.020608	0.017446	0
4	1	Return to slot 1; bottleneck and auxiliary FiLM	**0.977826**	0.020533	0.017235	1
5	4	Add three pairwise projections to slot 4	0.977506	0.020827	0.017108	2

Table 75: Frozen endpoints of the first two registered audit-feedback trajectories. The same selected checkpoint is used at both boundaries within the sci-Plex development source. These are individual observations, not policy-mean intervals or qualification counts.

Seed / slot	Boundary	PCC	MSE	Chemical loss gain	Dose loss gain
2026091201 / 5	Held-out audit	0.978061	0.020278	0.021173	0.008081
2026091201 / 5	Final held-out	0.977854	0.020495	0.007978	0.004612
2026091202 / 4	Held-out audit	0.978089	0.020252	0.023142	0.008650
2026091202 / 4	Final held-out	0.977718	0.020619	0.007835	0.004303