Do System Gains Reflect Useful Predictive Variation?

Anonymous authors
Paper under double-blind review

Abstract

Scientific prediction systems can reduce error by correcting an overall offset, removing harmful baseline variation, or predicting useful differences between samples. How much of an observed gain comes from predicting sample differences? We compare fixed drug-response and histology predictions with controls that remove sample variation while retaining the reference adjustment. In four of 28 drug comparisons, the system improves on the adjusted baseline even though its added predictive variation increases error; removing the old variation accounts for the gain. In histology, source-fitted amplitude changes contribution without changing gene-wise correlation, and risk weighting changes the judgment of an identical output. Downstream Hallmark scoring introduces a further distinction. Across five models, 92 patients, and 50 gene sets, moving removal from expression to final scores reduces the number of positive simultaneous intervals from 39 to 14 out of 250 comparisons. The gap arises because the two operations produce constant controls with different means, changing the loss against which the same prediction is judged. These findings separate the value of a complete system improvement from the contribution of the sample differences it predicts.

1 Introduction

Predicting gene expression from histology requires recovering differences between patients. A lower average error, however, can also result from correcting an overall offset or removing an adverse pattern in an existing baseline. The same distinction arises in drug-response prediction and forecasting: a useful system replacement need not owe its gain to the sample differences predicted by the replacement model.

Recent scientific benchmarks show how strongly conclusions depend on the reference prediction. Simple linear models remain competitive in gene-perturbation prediction (Ahlmann-Eltze et al., 2025), and evaluations of single-cell foundation models find that richer representations do not consistently improve downstream performance (Kedzierska et al., 2025; Wenteler et al., 2025). These comparisons establish the value of testing against informative baselines. They also motivate a more specific question: when a system improves, how much of that improvement comes from predicting differences between samples?

We address this question by comparing fixed model outputs with outputs from which the variation of interest has been removed. The comparisons retain the reference adjustment and use the same observations and loss. This follows the logic of removal-based evaluation (Covert et al., 2021; 2020; Hessel & Lee, 2020); the object evaluated here is variation in the scientific prediction itself. For a replacement, we distinguish the benefit of removing the old baseline variation from the contribution of adding the new model variation.

Three empirical findings develop this distinction. In drug response, four of 28 model–database comparisons improve on the reference-adjusted baseline even though the added model variation increases error. Their gains come from removing harmful baseline variation. In histology prediction, both risk weighting and source-fitted amplitude change the contribution assigned to a prediction, while positive amplitude scaling leaves gene-wise correlation unchanged. In downstream Hallmark scoring, removing variation before scoring and removing it after scoring give different contributions to the same complete prediction. Moving removal to the score level changes 25 of the original 39 positive simultaneous intervals to intervals crossing zero.

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.

Loss decompositions explain these observations: which operation supplies a system gain, when useful alignment is outweighed by amplitude, and why changing the removal location shifts the downstream comparison. A precipitation example extends the system-gain comparison to RMSE. The resulting account connects observed performance improvements to the particular changes in output that produce them.

2 Research Question and Comparison Design

2.1 Separating system gains from predictive contributions

Let F be a complete prediction and C the output after removing the variation to be evaluated. Both are assessed by the same loss L, using common observations, sample support, and weights. Define

$$P = L(C) - L(F). \tag{1}$$

A positive P means that retaining this variation lowers loss. In drug and histology prediction, C removes differences between evaluation samples while preserving a reference-adjusted mean. In precipitation, it is the historical mean used by the selected forecasting configuration.

For an original baseline b and a reference-adjusted baseline B, the loss differences satisfy

$$\begin{gathered} S = L(b) - L(F) = R + H + P, \\ R = L(b) - L(B), \qquad H = L(B) - L(C), \qquad J = H + P. \end{gathered} \tag{2}$$

Here S is the total system gain, R the reference-adjustment gain, and $J = L(B) - L(F)$ the gain from replacing B with F. When C removes variation from B while preserving its reference adjustment, H is the gain from removing baseline variation. Equation 2 is a telescoping identity for any common scalar loss. Normalized risks use the same positive denominator within each block and aggregate all differences with the same weights.

2.2 Data and evaluation units

The drug task uses fixed DrEval predictions trained on CTRPv2 and transferred to CCLE, CTRPv1, GDSC1, and GDSC2, yielding 28 model–database comparisons (Bernett et al., 2026; 2025). Histology models are trained on TCGA and evaluated on CPTAC breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), and glioblastoma (GBM), using the public cohort resources (The Cancer Genome Atlas Research Network et al., 2013; Heath et al., 2021; Edwards et al., 2015). The evaluation sets contain 103, 92, and 157 patients, with eight disjoint reference patients per cancer. We use UNI features (Chen et al., 2024) to train five models locally: MORSA (geogr0n, 2026), SEQUOIA (Pizurica et al., 2024), HE2RNA (Schmauch et al., 2020), DeepPT (Hoang et al., 2024), and Path2Omics (Hoang et al., 2025). Appendix I specifies their architectures and training settings. The Hallmark analysis uses the same 92 COAD evaluation patients.

Drug and histology risks are normalized by baseline risk and reported in percentage points (pp). Gene-set scores use MSE; precipitation uses date-average spatial RMSE in millimeters. Centering uses the unlabeled evaluation batch, whereas reference adjustment uses labels from a separate reference set. Bootstrap draws are paired across methods and resample cells, patients, or consecutive date blocks, as appropriate (Efron, 1979; Kunsch, 1989). Approximate 95% simultaneous intervals cover the specified comparison family; positive, negative, and crossing zero describe the location of the entire interval.

Model parameters and source-fitted amplitudes remain fixed during evaluation. The intervals describe variation across evaluation samples, with batch-estimated controls recomputed in each draw. Appendix A–B gives the loss definitions, resampling families, and sensitivity analyses.

3 Sources of System Improvement

3.1 Complete replacement and predictive variation

The drug baseline b_i contains cell and drug mean effects and thus varies across cells. For each drug and fixed fit, let δ be the reference adjustment, p_i the original model output, and α the source-fitted

amplitude. We construct

$$B_i = b_i + \delta, \qquad C_i = \bar{b} + \delta, \qquad F_i = C_i + \alpha(p_i - \bar{p}). \tag{3}$$

Replacing B with F removes the baseline's cell variation and adds the model's cell variation. The replacement gain is therefore

$$J = L(B) - L(F) = \underbrace{L(B) - L(C)}_{H} + \underbrace{L(C) - L(F)}_{P}. \tag{4}$$

All three outputs share δ. Here H measures the benefit of removing old variation and P the benefit of adding new variation. Means use the same evaluation samples. Within each cell–drug pair, squared errors are averaged over repeated measurements.

Table 1: All 28 drug comparisons. Cells give J/P in pp; superscripts $+, -, 0$ indicate positive, negative, or zero-crossing 95% simultaneous intervals. ENet, GBDT, RF, SNN, and SF denote ElasticNet, GradientBoosting, RandomForest, SimpleNeuralNetwork, and SuperFELTR. Full intervals appear in Appendix F.

Model	CCLE	CTRPv1	GDSC1	GDSC2
DIPK	-2.43^{0} / -15.71^{0}	26.25^{+} / 0.66^{0}	5.82^{+} / 0.08^{+}	6.33^{+} / 0.13^{+}
ENet	13.87^{0} / 0.59^{0}	26.90^{+} / 1.31^{0}	5.83^{+} / 0.09^{0}	6.76^{+} / 0.56^{0}
GBDT	19.32^{+} / 6.04^{0}	29.81^{+} / 4.22^{0}	5.93^{+} / 0.20^{+}	6.62^{+} / 0.43^{+}
RF	22.45^{+} / 9.18^{0}	32.01^{+} / 6.42^{+}	5.40^{+} / -0.34^{-}	5.65^{+} / -0.55^{-}
SRMF	-76.49^{-} / -89.76^{-}	-5.83^{0} / -31.42^{-}	-641.80^{-} / -647.54^{-}	-1001.16^{-} / -1007.35^{-}
SNN	16.43^{+} / 3.15^{0}	28.24^{+} / 2.65^{0}	4.25^{+} / -1.49^{-}	5.10^{+} / -1.10^{-}
SF	13.45^{0} / 0.17^{0}	25.66^{+} / 0.07^{+}	5.79^{+} / 0.05^{+}	6.20^{+} / $1.16\times10^{-8+}$

GDSC2/SF has $P = 1.16 \times 10^{-8}$ pp at source amplitude 10^{-8}; its positive classification corresponds to an extremely small effect.

Four comparisons—RandomForest and SimpleNeuralNetwork on GDSC1 and GDSC2—have positive J intervals and negative P intervals. Across all 28 comparisons, P has eight positive intervals, eight negative intervals, and 12 crossing zero (Table 1). Thus a favorable replacement can include a harmful contribution from the variation it adds.

3.2 Negative contributions within positive gains

For GDSC1/SimpleNeuralNetwork, removing baseline variation reduces risk by 5.738 pp, exceeding the 1.492 pp cost of adding model variation. Figure 1c follows this sequence to the positive net replacement gain. The other three cases show the same relationship. Comparing with B evaluates the complete replacement; comparing with C isolates the contribution of the newly added variation.

3.3 Overall improvement and variation in precipitation

We also examine CFSv2++ forecasts of weeks 3–4 accumulated precipitation (Mouatadid et al., 2023a;b), covering 102 weekly target dates in 2022–2023 and 862 contiguous-US grid points. Loss is the date-average spatial RMSE. Let f_t denote the raw forecast and $\mathcal{H}_t$ the historical samples used by the selected configuration. At each grid point, define

$$C_t = \frac{1}{|\mathcal{H}_t|} \sum_{r \in \mathcal{H}_t} Y_r, \qquad B_t = f_t + \frac{1}{|\mathcal{H}_t|} \sum_{r \in \mathcal{H}_t} (Y_r - f_r). \tag{5}$$

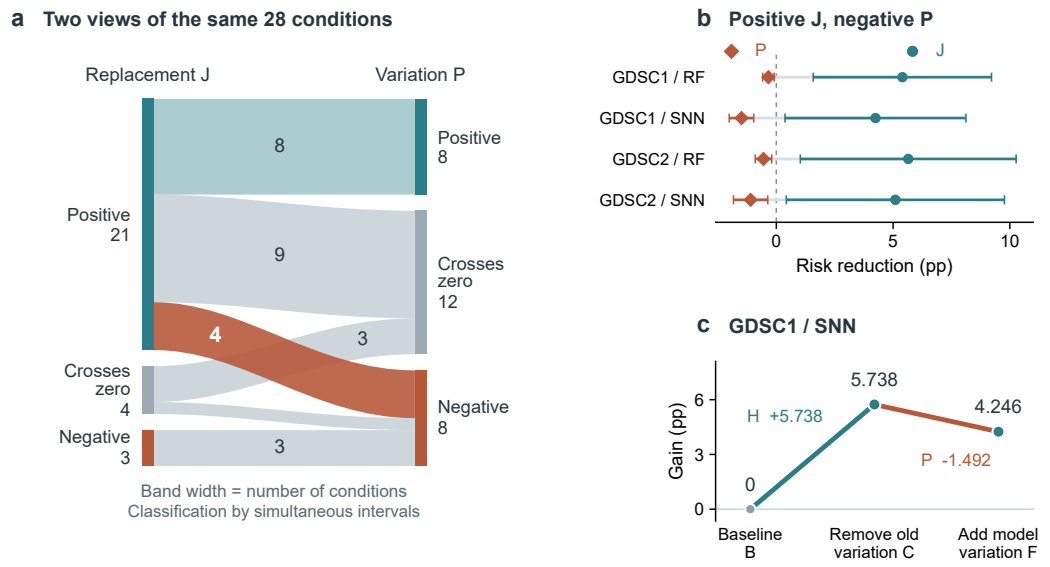

Figure 1: **A positive replacement gain can contain a negative contribution from predictive variation.** (a) Classification of all 28 model–database conditions under replacement gain J and variation contribution P. Bands count conditions; the rust band highlights the four positive-J, negative-P cases. (b) Their paired estimates and 95% simultaneous intervals. (c) Cumulative gain at three discrete outputs for GDSC1/SNN, relative to B. The sequence first removes baseline variation and then adds model variation. Panels (b,c) use normalized risk differences in pp. Intervals use 4,999 paired cell draws and a family of 88 contrasts. Conditions may share cells. RF and SNN denote RandomForest and SimpleNeuralNetwork.

The complete forecast F_t is the selected CFSv2++ output, and G_t is the native debiased CFSv2 baseline. The historical-mean control C_t varies with location and the selected historical support.

Table 2: Sources of precipitation improvement, in mm. Positive values indicate lower date-average spatial RMSE. Intervals jointly cover S, R, H, P, J, using 102 dates and 4,999 circular four-week block draws.

Contrast	Definition	Estimate	95% simultaneous CI
S	$L(G) - L(F)$	1.799	$[1.003, 2.595]$
J	$L(B) - L(F)$	0.683	$[0.007, 1.359]$
P	$L(C) - L(F)$	0.060	$[-0.427, 0.548]$

The complete system improves on G by 1.799 mm and on B by 0.683 mm. The decomposition attributes 1.116 mm to reference adjustment and 0.622 mm to removing the old forecast variation. These two operations account for most of the estimated system gain. Retaining the selected forecast variation contributes a further 0.060 mm, with interval $[-0.427, 0.548]$ mm (Table 2). Appendix H reports the component intervals and the eight-week block analysis.

4 Direction and Amplitude of Predictive Variation

4.1 Alignment benefit and amplitude cost

To explain why the added variation increases loss, we separate its direction from its size. Negative contribution can arise from misalignment or excessive amplitude. Under squared loss, let $v = F - C$.

With the same weights and normalization,

$$P = \underbrace{2\langle Y - C, v\rangle_w}_{A} - \underbrace{\langle v, v\rangle_w}_{Q}. \tag{6}$$

The term A measures alignment with the residual, while Q is the squared-amplitude cost. Even when $A > 0$, a larger Q makes $P < 0$. Three drug comparisons have positive A intervals and negative P intervals: CTRPv1/SRMF and SimpleNeuralNetwork on GDSC1 and GDSC2. In the GDSC1 example, the alignment benefit is 0.516 pp and the amplitude cost is 2.008 pp. These classifications use a joint family of 116 contrasts, including all 28 alignment terms. A location–amplitude construction isolates these effects with known population losses (Appendix D); the identities used in the comparisons are derived in Appendix C.

4.2 Amplitude control in histology prediction

For histology, the control removes differences between patients while retaining the gene baseline and reference adjustment: $B = C = b + \delta$ and $F = C + \alpha(p - \overline{p})$. Thus $H = 0$, and the replacement gain equals the predictive contribution. We evaluate three risks: equal weighting of per-gene relative improvements, pooling errors on the original scale, and pooling after standardization by source gene variance.

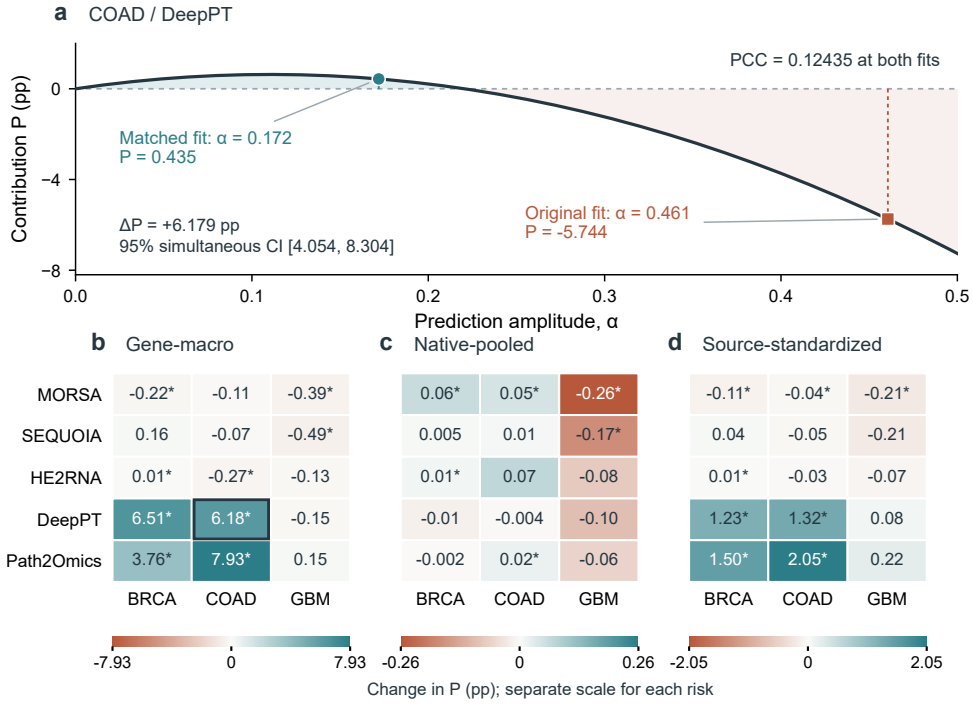

Figure 2: **Amplitude changes contribution while preserving correlation.** (a) Gene-macro contribution for COAD/DeepPT as amplitude varies. The curve uses target predictions and observations; both marked amplitudes are fitted at the source. Squares and circles mark the original-scale and risk-matched amplitude fits. Their common PCC and paired contribution difference are shown in the panel. (b–d) All 45 paired differences, matched minus original-scale fit, in pp. The outlined cell corresponds to (a). Asterisks indicate 95% simultaneous intervals excluding zero, based on 1,999 patient draws and 405 contrasts. Each risk uses its own normalization and color scale. Full intervals appear in Table 19.

Table 3: Three risks for the same COAD/DeepPT prediction, with amplitude 0.4606. Contributions are in pp; intervals use 270 contrasts.

Risk	P (95% simultaneous CI)
Gene-macro	$-5.744\ [-8.506, -2.981]$
Native-pooled	$1.896\ [0.238, 3.554]$
Standardized	$-0.699\ [-2.263, 0.866]$

The same COAD/DeepPT prediction has negative, positive, and zero-crossing contribution intervals under these three risks (Table 3). Only the risk weights and normalization change; the output is identical. We use this prediction as a worked example for the gene-macro amplitude analysis.

The original-scale fit estimates amplitude around the source baseline b, whereas evaluation defines variation around the prediction mean. A risk-matched amplitude fit instead centers the source out-of-fold predictions and minimizes the evaluation risk on those source patients. For COAD/DeepPT, this changes contribution by $6.179\ [4.054, 8.304]$ pp without changing any gene's across-patient PCC. The adjusted contribution is $0.435\ [-0.473, 1.343]$ pp: the paired improvement is supported even though the adjusted contribution interval crosses zero. Figure 2a shows both source-fitted amplitudes.

The target-data curve in Figure 2a is $P(\alpha) = 11.4651\alpha - 51.9650\alpha^2$ pp. Alignment dominates at small amplitudes, before the quadratic cost overtakes it. Its descriptive maximum occurs near $\alpha = 0.1103$; the evaluated prediction uses the source-fitted value 0.1719.

Across three cancers, five models, and three risks, the 45 risk-matched fits produce 31 positive P intervals and 14 crossing zero. The paired comparison asks a different question: relative to the original-scale fit, 14 contributions increase, nine decrease, and 22 differences cross zero (Figure 2b–d). Matching the source fitting risk therefore changes transferred contribution, with the direction depending on the model and risk. Appendix G reports all paired results.

5 Predictive Contributions to Downstream Scores

5.1 Defining removal in the downstream task

Downstream use requires specifying where variation is removed. Expression predictions are often used to compute gene-set scores. We evaluate 50 Hallmark sets using the five models' original-scale fitted outputs on the same 92 CPTAC-COAD patients and 20,717 genes. The singscore readout g uses within-sample gene ranks (Foroutan et al., 2018; Liberzon et al., 2015) and is applied identically to observed RNA and predictions.

One control averages expression predictions before scoring, $g(\overline{F})$; the other scores predictions before averaging, $\overline{g(F)}$. Both means use evaluation patients. The nonlinear readout permits these constant scores to differ. Holding the complete scores $g(F)$ and their MSE L_g fixed, define

$$
\begin{aligned}
P_{\text{gene}} &= L_g\big(g(\overline{F})\big) - L_g(g(F))\,, \\
P_{\text{score}} &= L_g\Big(\overline{g(F)}\Big) - L_g(g(F))\,, \\
\Delta = P_{\text{gene}} - P_{\text{score}} &= L_g\big(g(\overline{F})\big) - L_g\Big(\overline{g(F)}\Big)\,. \qquad (7)
\end{aligned}
$$

Both contributions are evaluated on downstream score MSE; the subscript identifies the removal location. P_{score} measures the benefit of patient-to-patient score variation. Δ is the loss difference between the two constant controls.

5.2 How removal location changes contribution estimates

Moving removal from expression to scores changes 25 of the original 39 positive classifications to intervals crossing zero, leaving 14 positive intervals among the 250 comparisons. In 24 of those 25 cases, the paired Δ interval is positive. Thus the difference between the two contribution estimates is itself supported, rather than inferred only from a change in their separate interval classifications.

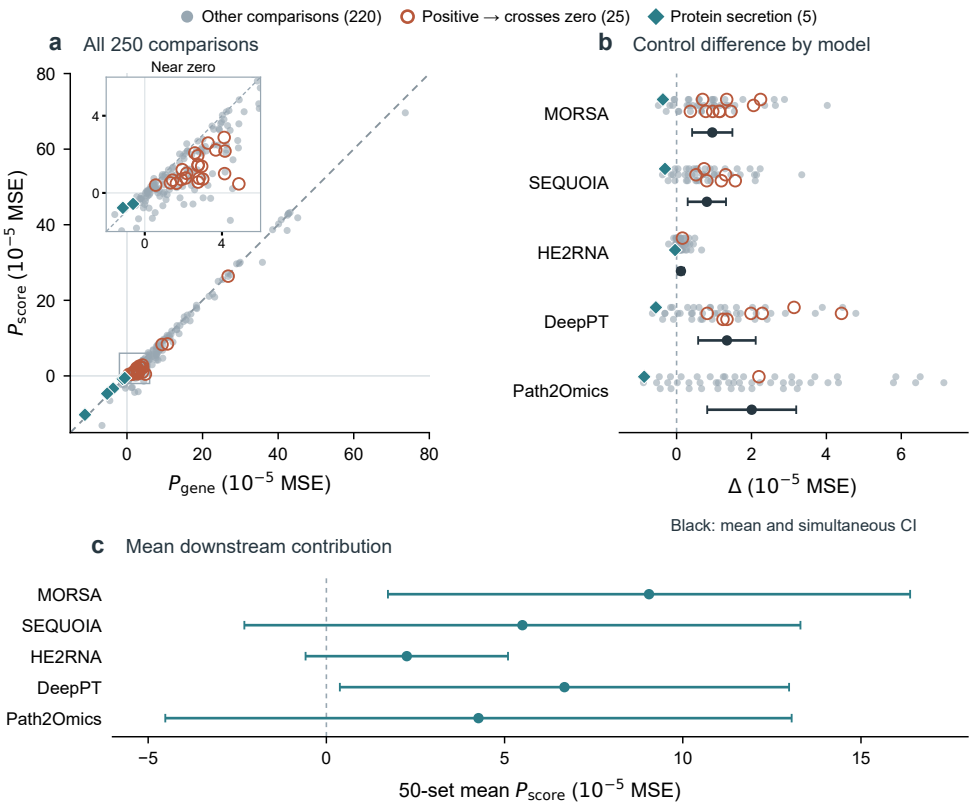

Figure 3: **Removal location changes the contribution assigned to the same downstream prediction.** (a) Expression-level contribution P_{gene} versus score-level contribution P_{score} for all 250 model–gene-set conditions. The dashed line marks equality; the inset expands $[-2, 6]$ on both axes. Rust rings mark the 25 positive-to-zero-crossing comparisons; teal diamonds mark the five protein-secretion conditions with negative Δ intervals. (b) The 50 Δ estimates for each model, with the mean and its 95% simultaneous interval in black. (c) Mean P_{score} and its interval for each model. Units are 10^{-5} score MSE. All conditions share 92 patients; intervals use 1,999 paired patient draws and a family of 765 contrasts.

Across all 250 comparisons, the control gap Δ has 141 positive intervals and five negative intervals; the remaining 104 cross zero. All five negative gaps concern protein secretion, one per model. For this gene set, expression-level removal assigns a smaller contribution than score-level removal. Full classifications and intervals appear in Appendix E.

Averaging over the 50 gene sets gives positive P_{score} point estimates for all five models. The intervals are positive for MORSA and DeepPT and cross zero for SEQUOIA, HE2RNA, and Path2Omics (Figure 3c). All five mean Δ intervals are positive (Figure 3b).

The direction of the comparison persists across 256 alternative reference partitions: each has more positive P_{gene} point estimates than positive P_{score} point estimates. These are signs of point estimates, rather than repeated simultaneous-interval classifications; Appendix B reports the ranges and rank correlations.

An affine readout gives the predicted zero removal-location gap across 750 cancer–model–gene-set conditions (Appendix D).

5.3 Where the control difference comes from

Write $u_i = g(Y_i)$, $s_i = g(F_i)$, $a = g(\overline{F})$, $b = \overline{s}$, and $m = \overline{u}$. Under equally weighted MSE,

$$P_{\text{score}} = 2\,\text{Cov}_n(u, s) - \text{Var}_n(s), \tag{8}$$

$$\Delta = (a - m)^2 - (b - m)^2 = d^2 + 2de, \qquad d = a - b, \quad e = b - m. \tag{9}$$

Covariance and variance use denominator n. Score variation reduces MSE when twice its covariance with observations exceeds its variance. The control difference has an equally direct interpretation: moving the constant from b to a changes its squared distance from the observed mean m. Nonlinearity can make the constants differ; their positions relative to m determine the sign of the loss difference.

For all 25 positive-to-zero-crossing comparisons, d and e have the same sign. Expression-level removal moves the constant farther from the observed mean, increasing control loss and the apparent contribution. All five protein-secretion comparisons satisfy $b < a < m$: the movement instead brings the constant closer to the observed mean, producing negative Δ. Table 7 gives the complete decomposition for these cases.

Using $b = \overline{s}$ preserves the predicted score mean when patient differences are removed. Reporting P_{score} alongside Δ separates the gain from predicting patient differences from the loss difference caused by moving the constant control.

6 Related Work

Removal and the object being evaluated. Removal-based explanations distinguish what is removed, what replaces it, and which behavior is measured (Covert et al., 2021). SAGE assigns feature importance through predictive loss reduction (Covert et al., 2020); EMAP evaluates cross-modal interaction by removing it from a fixed predictor (Hessel & Lee, 2020). Sanity checks test whether attribution maps respond to changes in a model or its training labels (Adebayo et al., 2018), while ROAR evaluates feature rankings through removal and retraining (Hooker et al., 2019). Our comparisons act on fixed scientific outputs: the removed object is their sample-to-sample variation, and the retained reference adjustment makes its contribution distinguishable from other gains in the replacement.

Baselines, calibration, and scientific performance. Systema separates systematic responses from perturbation-specific prediction through a training-perturbation reference centroid (Viñas Torné et al., 2026). Linear-baseline studies and recent foundation-model benchmarks expose how shared structure and evaluation settings affect apparent performance (Ahlmann-Eltze et al., 2025; Kedzierska et al., 2025; Wenteler et al., 2025). Forecast verification provides complementary tools: MSE skill decompositions separate correlation, bias, and scale (Murphy, 1988; Gupta et al., 2009); benchmark choice determines the interpretation of normalized skill (Knoben et al., 2019); and the loss defines which point prediction is optimal (Gneiting, 2011). We connect these issues within matched output comparisons. The drug decomposition distinguishes removal of an adverse baseline pattern from addition of new variation, and the histology analysis measures how source amplitude fitting and gene weighting alter that variation's contribution.

Readout and mean replacement. Gene-set methods turn expression profiles into sample-level molecular summaries (Hänzelmann et al., 2013; Foroutan et al., 2018). Optimal Ablation studies replacement constants for internal components and explains why nonlinear downstream computation complicates mean replacement (Li & Janson, 2024). Our contribution is an empirical account of the resulting loss gap in scientific outputs: we measure its size, explain both signs through the two controls' positions relative to the observed score mean, and test whether it changes contribution judgments. The complete prediction and downstream loss stay the same throughout this comparison.

7 Discussion and Conclusion

A system can be worth replacing without the replacement predicting useful sample differences. This distinction changes how an improvement should be interpreted. A positive total gain supports the

combined operation, while the matched contribution identifies what the added variation supplies. The drug cases show that removing an adverse baseline pattern can account for the gain even when the new variation increases error. Reporting the two components makes that finding visible and identifies a part of the output that can be improved.

Correlation alone cannot settle this question. It detects co-variation but does not determine whether the predicted amplitude is suitable for the loss or whether the errors receiving most weight have improved. The direction–amplitude decomposition separates these possibilities: positive alignment with a negative contribution points to excessive amplitude, whereas negative alignment cannot be repaired by positive rescaling. Source-only amplitude fitting provides one practical response. Its mixed paired effects across the histology comparisons also show why a calibration should be evaluated under the risk for which it will be used.

Downstream scoring makes the evaluation object more specific. To measure the benefit of differences between patient scores, removing those differences should retain the predicted score mean. Averaging expression first can move that mean, so its loss contrast includes a second effect. Reporting P_{score} and Δ distinguishes the contribution of patient differences from this change in the control. Across the empirical analyses, baseline removal, amplitude, and readout location each change the account of where a system's gain comes from. Evaluating the output variation used by the task makes that account explicit.

AI Use Statement

Generative AI tools assisted manuscript drafting, editing, and translation; literature search and discovery; data formatting; and preparation of analysis code and figures. Numerical claims were checked against archived predictions and reproducible computations, and references were checked against primary sources. The authors take responsibility for the final content and conclusions.

Ethics Statement

This study reanalyzes existing drug-response, TCGA/CPTAC expression, and precipitation resources without new participant recruitment or animal experiments. The supplementary data contain contribution estimates, saved resampling results, patient-level Hallmark scores without study identifiers, and weather observations and predictions. Raw histology slides and controlled-access sequencing reads are not redistributed. Data sources and reuse conditions are documented in the supplementary package. The analyses concern model evaluation and are not intended for individual clinical decisions.

Reproducibility Statement

The supplementary ZIP contains code, archived estimates and paired resamples, Hallmark patient scores, weather inputs, and figure source tables. It supports reconstruction of the reported simultaneous intervals, direct verification of Hallmark point contributions and decomposition identities, recalculation of summaries from saved deletion and reference-partition results, regeneration of the weather analyses and analytic example, and redrawing of all figures. Drug record-moment matrices and gene-expression matrices are omitted to meet the supplementary file-size limit, so the package does not regenerate the drug, histology, or Hallmark bootstrap draws from those matrices; affine-readout checks are supplied as archived summaries. Appendices A and B specify output construction, losses, and inference; Appendix C derives the identities, and Appendix I documents model architectures and training. Source-model training and upstream forecast selection are not rerun.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

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: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Judith Bernett, Pascal Iversen, and Mario Picciani. Dataset for drevalpy, may 2025.

Judith Bernett, Pascal Iversen, Mario Picciani, Mathias Wilhelm, Katharina Baum, and Markus List. Critical evaluation of drug response prediction models with DrEval. *Nature Communications*, 17: 4238, 2026. URL https://www.nature.com/articles/s41467-026-72903-w.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c7bf0b7c1a86d5eb3be2c722cf2cf746-Abstract.html.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Nathan J. Edwards, Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, Peter B. McGarvey, Shine Jacob, Subha Madhavan, and Karen A. Ketchum. The CPTAC data portal: A resource for cancer proteomics research. *Journal of Proteome Research*, 14(6):2707–2713, 2015. doi: 10.1021/pr501254j.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552.

Momeneh Foroutan, Dharmesh D. Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19:404, 2018. doi: 10.1186/s12859-018-2435-4.

geogr0n. MORSA: Morphology-omics representation via structural alignment. Software repository, 2026. URL https://github.com/geogr0n/MORSA.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009. doi: 10.1016/j.jhydrol.2009.08.003.

Allison P. Heath, Vincent Ferretti, Stuti Agrawal, Maksim An, James C. Angelakos, Renuka Arya, Rosita Bajari, Bilal Baqar, Justin H. B. Barnowski, Jeffrey Burt, et al. The NCI genomic data commons. *Nature Genetics*, 53(3):257–262, 2021. doi: 10.1038/s41588-021-00791-5.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 861–877, 2020. doi: 10.18653/v1/2020.emnlp-main.62.

Danh-Tai Hoang, Gal Dinstag, Eldad D. Shulman, et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nature Cancer*, 5(9):1305–1317, 2024. doi: 10.1038/s43018-024-00793-2.

Danh-Tai Hoang, Eldad D. Shulman, Saugato Rahman Dhruba, et al. Path2Omics enhances transcriptomic and methylation prediction accuracy from tumor histopathology. *Cancer Research*, 85(24):5098–5112, 2025. doi: 10.1158/0008-5472.CAN-25-4350.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1), 2013. doi: 10.1186/1471-2105-14-7.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1), 2025. doi: 10.1186/s13059-025-03574-x.

Wouter J. M. Knoben, Jim E. Freer, and Ross A. Woods. Technical note: Inherent benchmark or not? comparing nash–sutcliffe and kling–gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331, 2019. doi: 10.5194/hess-23-4323-2019.

Hans R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3), 1989. doi: 10.1214/aos/1176347265.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c55e6792923cc16fd6ed5c3f672420a5-Abstract-Conference.html.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Judah Cohen, Miruna Oprescu, Ernest Fraenkel, and Lester Mackey. Adaptive bias correction for improved subseasonal forecasting. *Nature Communications*, 14:3482, 2023a. doi: 10.1038/s41467-023-38874-y.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester Mackey. SubseasonalClimateUSA: A dataset for subseasonal forecasting and benchmarking. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/18ef499ee57c4822e1e3ea9b9948af18-Abstract-Datasets_and_Benchmarks.html.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988. URL https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml.

Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15:9886, 2024. doi: 10.1038/s41467-024-54182-5.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*, 11:3877, 2020. doi: 10.1038/s41467-020-17678-4.

The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.

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*, 44:1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66633–66677. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wenteler25a.html.

A Output Construction and Loss Definitions

A.1 Histology predictions, reference adjustment, and amplitude

BRCA, COAD, and GBM contain 20,790, 20,717, and 20,806 common genes and 103, 92, and 157 evaluation patients, respectively. Each cancer has eight additional reference patients. All models share the patient and gene axes within each cancer. A deterministic ordering of patient identifiers, independent of expression labels and predictions, assigns eight patients to the reference set and the remainder to evaluation. The supplementary package provides the resulting contribution estimates and paired resampling results; patient identifiers and expression matrices are not included. Source gene means b_g and standard deviations σ_g are fixed. The reference adjustment is

$$\delta_g = \frac{\sum_{r\in\mathcal{R}}(y_{rg} - b_g)}{8+8},$$

where the second 8 is the prior weight. Evaluation outputs are $F_{ig} = C_g + \alpha(p_{ig} - \overline{p}_g)$, with $B_g = C_g = b_g + \delta_g$ and $\overline{p}_g$ the unlabeled evaluation-batch mean.

Unit amplitude sets $\alpha = 1$. Original-scale and standardized fits use source out-of-fold predictions:

$$\alpha = \text{clip}_{[0,1]} \frac{\sum_{i,g} w_g(p_{ig} - b_g)(y_{ig} - b_g)}{\sum_{i,g} w_g(p_{ig} - b_g)^2},$$

with $w_g = 1$ and $w_g = 1/\sigma_g^2$, respectively. A zero denominator gives $\alpha = 0$. The fitted coefficient is then fixed for target evaluation.

For the centered, risk-matched fit, define source variation $v_{ig} = p_{ig} - \overline{p}_g$ and source baseline risk $D_g^{\text{src}} = \text{mean}_i(y_{ig} - b_g)^2$. Gene-macro, native-pooled, and source-standardized risks use weights proportional to $1/D_g^{\text{src}}$, 1, and $1/\sigma_g^2$, respectively:

$$\alpha = \text{clip}_{[0,1]} \frac{\sum_g w_g \,\text{mean}_i[(y_{ig} - b_g)v_{ig}]}{\sum_g w_g \,\text{mean}_i[v_{ig}^2]}.$$

Centering y in the numerator gives the same result because v is centered. This amplitude fit changes both the centering and the fitting risk. Table 18 gives all coefficients and source sample sizes.

A.2 Three gene-level risks

Let $L_g(h) = n^{-1}\sum_i(y_{ig} - h_{ig})^2$, $D_g = L_g(b)$, and $d_g = L_g(C) - L_g(F)$. The three contribution summaries are

$$P_{\text{macro}} = \frac{1}{G}\sum_g \frac{d_g}{D_g}, \qquad P_{\text{native}} = \frac{\sum_g d_g}{\sum_g D_g}, \qquad P_{\text{std}} = \frac{\sum_g d_g/\sigma_g^2}{\sum_g D_g/\sigma_g^2}. \tag{10}$$

Genes must have finite, strictly positive source variance and positive baseline risk on the original evaluation sample. The latter condition excludes no genes in the three cancers. The common gene set is fixed across draws. A nonpositive denominator or nonfinite statistic invalidates the entire draw vector and stops interval classification; all 1,999 draws are valid. Inclusion does not depend on correlation or contribution sign. Gene-macro risk weights relative improvements equally, whereas native pooling weights them by D_g, explaining how the same prediction can have different contribution signs.

A.3 Hallmark readout

We use the 50 Hallmark gene sets from MSigDB 2026.1. For each patient, all $N = 20{,}717$ genes are ranked using the minimum rank for ties. For a set with m genes intersecting the common gene axis,

$$g(x) = \frac{\text{meanrank}_{\text{set}}(x) - (m+1)/2}{N - m} - \frac{1}{2}.$$

The same rule is applied to observed RNA, complete expression predictions, and constant expression controls. Score loss is $L_g(s) = n^{-1}\sum_i(g(Y_i) - s_i)^2$. The macro-average gives equal weight to the 50 set-specific loss differences. Patients are the resampling unit. The downstream experiment uses original-scale fitted outputs; matched amplitudes are evaluated separately under gene-level risks.

A.4 Drug record risk and reference information

Each fixed drug configuration has eight reference cells. For a drug with k available reference observations, $\delta = \sum_{\mathcal{R}_{\text{obs}}}(y-b)/(k+8)$; it is zero if there are no valid reference observations. The source amplitude is fitted on out-of-fold pair means standardized using source drug scales:

$$\alpha = \text{clip}_{[10^{-8},1]} \frac{\sum_{\text{source}} yp}{\sum_{\text{source}} p^2}.$$

A database–model pair shares one coefficient across drugs and ten target fits. Source fitting excludes cell identifiers from the target database, while source standardization and baseline parameters remain fixed. The source denominator must be finite and strictly positive. Centering in Equation (3) uses common support and is recomputed with the weights of each cell draw.

Each observed cell–drug pair has unit weight. Repeated records within a pair receive equal weight in its squared error, retaining their second moments. Each drug uses the same original-baseline risk as denominator. Drug-level quantities are macro-averaged over the predetermined drug axis and then averaged over ten fits. Models and databases can share drugs, cells, and fits. Writing $S = L(b) - L(F)$ and $R = L(b) - L(B)$ gives $S = R + J = R + H + P$; this identity is preserved by the averaging.

The seven models are DIPK, ElasticNet, GradientBoosting, RandomForest, SRMF, SimpleNeuralNetwork, and SuperFELTR. Tables abbreviate the last two as SNN and SF, and GradientBoosting and RandomForest as GBDT and RF. All 28 model–database pairs are retained. GDSC2/SF has source amplitude 10^{-8}, fitted from five nonoverlapping source cells and 1,211 pairs. Its tiny contribution is computed directly as $A - Q$ to avoid cancellation between larger losses; independent high-precision checks confirm the result.

Reference cells are selected through a deterministic ordering independent of response labels and predictions. Sixteen cells per database are reserved, with the first eight used for the primary adjustment. The union of these pools contains 62 distinct cells, excluded from all evaluations, leaving a canonical universe of 875 evaluation cells. Drugs use common observed support across ten fits, require at least 20 evaluation cells, and have positive original-baseline risk.

A.5 Precipitation support and configuration

Evaluation covers weekly targets from January 19, 2022 to December 27, 2023: 102 dates and 862 contiguous-US grid points at one-degree resolution. The target is weeks 3–4 accumulated precipitation over 14 days. Configuration selection uses a 12-year fitting window, a three-year tuning window, and 80 candidate configurations. Target labels become available after 29 days. The historical mean C and raw-forecast residual adjustment B use the same historical samples as the selected CFSv2++ configuration.

With native debiased CFSv2 denoted by G, define $S = L(G) - L(F)$, $R = L(G) - L(B)$, $H = L(B) - L(C)$, $P = L(C) - L(F)$, and $J = H + P$. Thus $S = R + H + P$. Mean losses are $L(G) = 25.3443$, $L(B) = 24.2281$, $L(C) = 23.6057$, and $L(F) = 23.5453$ mm. Each date first yields a spatial RMSE, then dates are averaged. Contributions are direct differences of these RMSE losses, without applying the squared-loss decomposition. Table 21 reports both block lengths.

B Intervals, Statistical Families, and Sensitivity

All comparisons within a data domain share sample indices in each draw. Let $\hat{\theta}_j$ be a point estimate, $\hat{\theta}_j^{(r)}$ its estimate in draw r, and s_j its bootstrap standard deviation. Approximate simultaneous intervals are

$$\hat{\theta}_j \pm q_{0.95} s_j, \qquad q_{0.95} = \text{Quantile}_{0.95}\left\{ \max_{j: s_j > 0} \left| \frac{\hat{\theta}_j^{(r)} - \hat{\theta}_j}{s_j} \right| \right\}.$$

The drug analysis uses linear interpolation for the empirical quantile; histology, Hallmark, and precipitation use the upper empirical quantile. Standard deviations use denominator $R-1$. Numerically constant coordinates receive zero width. The release code preserves each analysis's numerical tolerance.

Table 4: Simultaneous-interval families. S/P denotes two separate contrasts per condition.

Analysis	Family	Draws
Drug gains	28 each of S, J, P, plus four database-level H: 88	4,999
Drug alignment	The 88 gain contrasts plus 28 A: 116	4,999
Histology, original	S/P for 135 conditions: 270	1,999
Matched amplitudes	Original 270 plus S/P for 45 fits: 360	1,999
Paired amplitudes	The 360 contrasts plus 45 paired P differences: 405	1,999
Hallmark readout	Five models $\times$ 51 summaries $\times$ three contrasts: 765	1,999
Precipitation	S, R, H, P, J: five	4,999

Drug resampling uses the canonical 875-cell universe and each drug's available observations within a database. The critical values are 3.4319074002 for 88 contrasts and 3.5820669514 for 116. The histology 270- and 360-contrast families both use 3.5761578528; the 405-contrast paired family uses 3.5864534578. The Hallmark critical value is 3.9651985233. Original and matched amplitudes share patient weights, and their within-draw differences form the paired contrasts. Intervals quantify evaluation-sample uncertainty conditional on the trained systems, coefficients, and primary reference sets. This is a retrospective analysis: the choice of comparisons and illustrative examples was informed by target-cohort results. The intervals do not account for this analysis selection or for retraining the models. Source-fitted amplitudes use source patients only.

Sensitivity analyses change reference partitions or delete evaluation cells one at a time. There are 15,085 drug deletion records. The sign of the H point estimate remains stable in all 28 conditions; the signs of P and J each remain stable in 27. Table 15 gives all resulting ranges. These are deletion ranges, not confidence intervals.

The Hallmark experiment uses 256 alternative partitions with eight reference patients and prior weight eight. Across each partition's 250 conditions, the number of positive P_{gene} point estimates ranges from 214 to 233 (median 225); the corresponding P_{score} count is 190–215 (median 201). The difference is 15–32, with median 24. Spearman correlation of P_{score} with the primary partition is 0.934177–0.994400 (median 0.985229). These sign counts differ from the main-text counts of 39 and 14 positive simultaneous intervals. A further nine reference-count/prior-weight combinations use 64 partitions each.

Precipitation uses circular moving blocks over the ordered weekly dates, with four-week blocks for the primary analysis and eight-week blocks for sensitivity. The 862 spatial grid points jointly determine a date's RMSE; date blocks, rather than individual grid points, are resampled. Effects and intervals are reported in millimeters.

C Identities Used in the Comparisons

C.1 System gains

Equation 2 follows by adding and subtracting $L(B)$ and $L(C)$ from $L(b) - L(F)$. It holds for any common scalar loss, including the date-average spatial RMSE used for precipitation. A common positive normalizer and common outer averaging preserve the identity. Its terms acquire their interpretation from the outputs being compared.

C.2 Alignment, amplitude, and correlation

For a weighted squared loss, put $v = F - C$. Expanding $(Y - C - v)^2$ coordinatewise gives

$$P = 2\langle Y - C, v\rangle_w - \langle v, v\rangle_w = A - Q.$$

The identity allows a sample-varying C. If C is constant across samples and v has weighted mean zero, the alignment term is twice the weighted covariance of outcomes and predictions. Common positive normalizers rescale all three terms equally.

For $F_i = c + \alpha(p_i - \overline{p})$ with sample-constant c and $\alpha > 0$,

$$\text{Corr}(Y, F) = \text{Corr}(Y, p).$$

Centering removes c, covariance scales by α, and prediction standard deviation scales by $|\alpha| = \alpha$. The factors cancel. This holds separately for each gene whenever both variances are nonzero. It explains why positive amplitude changes in the histology analysis leave PCC unchanged.

C.3 Removal location

For equally weighted downstream MSE, let $u_i = g(Y_i)$, $s_i = g(F_i)$, $a = g(\overline{F})$, $b = \overline{s}$, and $m = \overline{u}$. Expanding around the sample means gives

$$P_{\text{score}} = 2\,\text{Cov}_n(u, s) - \text{Var}_n(s), \qquad P_{\text{gene}} - P_{\text{score}} = (a - m)^2 - (b - m)^2.$$

In the second identity, the common complete-prediction loss cancels. Each constant-control loss is the outcome variance plus its squared distance from m, so the variance also cancels. Substituting $d = a - b$ and $e = b - m$ gives $\Delta = d^2 + 2de$. These are finite-sample identities; they require no distributional model for the outcomes. An affine g commutes with averaging and makes the gap zero. A nonlinear g permits a gap, whose sign depends on the two constants' positions relative to m.

D Location and Amplitude Example and a Linear Readout Check

D.1 Location and amplitude with known population losses

Let Z, ϵ be independent, zero-mean, unit-variance variables, $Y = 1 + Z$, and $F = 1 + \beta + \alpha Z + \sigma\epsilon$. With $C = \mathbb{E}F$, baseline zero, and denominator $\mathbb{E}Y^2 = 2$,

$$R = (1 - \beta^2)/2, \quad A = \alpha, \quad Q = (\alpha^2 + \sigma^2)/2,$$
$$P = A - Q, \quad S = R + P, \quad \text{Corr}(Y, F) = \frac{\alpha}{\sqrt{\alpha^2 + \sigma^2}}.$$

Here R is the gain from the zero baseline to the mean control, and $H = 0$. With $\sigma = 0.2$, the 3,939-point grid computes these terms from population losses. Changing location affects R and S without changing P or correlation. Increasing amplitude eventually makes P negative despite increasing correlation (Figure 4). This construction isolates the two operations behind the empirical comparisons.

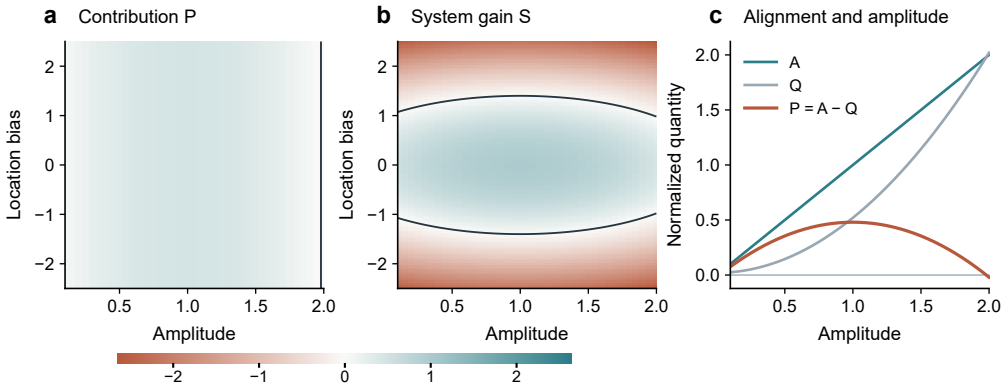

Figure 4: **Location and amplitude have different effects on contribution.** (a) Population P over prediction amplitude and location bias; (b) population S on the same grid and scale, with the zero contours drawn separately. Each grid has 3,939 settings. (c) At zero location bias, alignment grows linearly and amplitude cost quadratically; their difference determines P. Noise standard deviation is 0.2 and the common baseline risk is 2. These are analytic population values.

D.2 Affine readout on the observed cohorts

Replace singscore by the mean expression over each fixed Hallmark gene set. This affine readout satisfies $g(\overline{F}) = \overline{g(F)}$ exactly in real arithmetic. We apply it to the original-scale fitted outputs for

all three cancers, five models and 50 sets: 750 conditions. The maximum absolute control difference is 4.00×10^{-15} and the maximum loss difference is 1.48×10^{-15}. These differences are at numerical precision, as predicted by the affine identity.

E Supplementary Downstream Results

Table 5: Means across 50 gene sets and 95% simultaneous intervals, in 10^{-5} score MSE.

Model	P_{gene}	P_{score}	Δ
MORSA	10.01 [2.63, 17.38]	9.05 [1.73, 16.38]	0.95 [0.41, 1.49]
SEQUOIA	6.31 [−1.50, 14.12]	5.50 [−2.30, 13.30]	0.81 [0.30, 1.32]
HE2RNA	2.37 [−0.47, 5.21]	2.26 [−0.58, 5.10]	0.12 [0.03, 0.20]
DeepPT	8.02 [1.68, 14.37]	6.68 [0.38, 12.98]	1.34 [0.58, 2.11]
Path2Omics	6.27 [−2.59, 15.14]	4.27 [−4.52, 13.05]	2.01 [0.82, 3.19]

Table 5 reports the five model means. Table 6 lists all 25 positive-to-zero-crossing changes with their paired differences. Table 9 retains all 250 conditions and three contrasts. Table 8 maps set identifiers.

Table 6: All 25 positive-to-zero-crossing changes. Set identifiers appear in Table 8. Values are in 10^{-5} score MSE; 24 paired Δ intervals are positive.

Model	Set	P_{gene}	P_{score}	Δ
MORSA	01	2.17 [0.56, 3.77]	1.01 [−0.63, 2.66]	1.15 [0.62, 1.69]
MORSA	16	2.81 [0.56, 5.07]	0.76 [−1.51, 3.03]	2.06 [1.17, 2.94]
MORSA	19	3.68 [0.84, 6.52]	2.23 [−0.65, 5.10]	1.45 [0.78, 2.12]
MORSA	22	10.70 [0.66, 20.75]	8.46 [−1.62, 18.55]	2.24 [0.86, 3.62]
MORSA	26	26.77 [0.22, 53.32]	26.40 [−0.20, 53.01]	0.37 [−1.24, 1.97]
MORSA	39	3.28 [0.04, 6.51]	2.58 [−0.55, 5.72]	0.69 [0.29, 1.10]
MORSA	42	2.75 [0.70, 4.81]	1.42 [−0.55, 3.38]	1.34 [0.94, 1.74]
MORSA	46	1.44 [0.03, 2.85]	0.66 [−0.70, 2.02]	0.78 [0.45, 1.11]
MORSA	48	1.64 [0.28, 2.99]	0.50 [−0.72, 1.72]	1.14 [0.63, 1.64]
MORSA	49	9.28 [0.16, 18.40]	8.31 [−0.77, 17.40]	0.97 [0.09, 1.84]
SEQUOIA	01	1.89 [0.20, 3.59]	0.70 [−0.83, 2.23]	1.20 [0.57, 1.82]
SEQUOIA	19	2.77 [0.30, 5.23]	1.46 [−1.03, 3.95]	1.31 [0.54, 2.08]
SEQUOIA	21	2.59 [0.47, 4.71]	2.07 [−0.02, 4.17]	0.52 [0.26, 0.78]
SEQUOIA	31	1.94 [0.43, 3.45]	1.21 [−0.16, 2.57]	0.73 [0.22, 1.25]
SEQUOIA	42	2.95 [0.44, 5.46]	1.38 [−1.04, 3.80]	1.57 [0.99, 2.15]
SEQUOIA	46	1.33 [0.28, 2.37]	0.52 [−0.43, 1.48]	0.81 [0.41, 1.20]
HE2RNA	19	0.56 [0.01, 1.11]	0.41 [−0.15, 0.97]	0.16 [0.04, 0.27]
DeepPT	15	4.87 [0.55, 9.20]	0.47 [−3.40, 4.35]	4.40 [2.48, 6.32]
DeepPT	16	4.14 [0.75, 7.52]	1.00 [−1.80, 3.80]	3.14 [1.80, 4.47]
DeepPT	19	4.16 [0.86, 7.45]	2.17 [−1.03, 5.37]	1.99 [1.02, 2.95]
DeepPT	21	2.74 [0.01, 5.48]	1.93 [−0.71, 4.56]	0.82 [0.39, 1.24]
DeepPT	31	4.11 [0.12, 8.10]	2.87 [−0.82, 6.57]	1.24 [0.40, 2.08]
DeepPT	42	3.01 [0.15, 5.86]	0.72 [−2.05, 3.49]	2.29 [1.51, 3.07]
DeepPT	46	2.12 [0.02, 4.22]	0.77 [−1.21, 2.76]	1.35 [0.74, 1.95]
Path2Omics	46	2.76 [0.17, 5.35]	0.56 [−1.75, 2.87]	2.20 [1.10, 3.29]

E.1 Patient-level reconstruction and mean decomposition

All models use the same observed RNA, reference set, source means, patient and gene axes, and Hallmark definitions. Two independent ranking implementations cross-check the minimum-rank score, assigning tied expression values their group's minimum rank. Missing genes are not imputed. Patient-level scores reconstruct absolute losses, contributions, and mean decompositions. All 250 conditions satisfy Equations (8) and (9).

Table 7 includes all 25 classification changes and the five reverse differences for protein secretion. In the former, d and e have the same sign. In the latter, $d > 0$, $e < 0$, and $a < m$: expression-

level removal moves the constant toward the observed mean. These decompositions explain point estimates; contribution and difference intervals appear in Table 9.

Table 7: Mean decomposition for the 25 classification changes and five reverse protein-secretion cases. Set identifiers appear in Table 8. The displayed d, e are multiplied by 10^3; the last three columns by 10^5. In the first 25 cases d, e have the same sign; the last five satisfy $b < a < m$.

Model	Set	d	e	d^2	$2de$	Δ
MORSA	01	0.466	12.158	0.022	1.132	1.154
MORSA	16	0.544	18.604	0.030	2.025	2.055
MORSA	19	0.428	16.756	0.018	1.433	1.451
MORSA	22	0.629	17.495	0.040	2.202	2.241
MORSA	26	0.557	3.018	0.031	0.337	0.368
MORSA	39	0.316	10.820	0.010	0.684	0.694
MORSA	42	0.345	19.190	0.012	1.326	1.338
MORSA	46	0.331	11.620	0.011	0.770	0.781
MORSA	48	0.390	14.370	0.015	1.122	1.137
MORSA	49	0.121	39.692	0.001	0.964	0.965
SEQUOIA	01	0.483	12.141	0.023	1.172	1.195
SEQUOIA	19	0.385	16.799	0.015	1.294	1.309
SEQUOIA	21	0.275	9.304	0.008	0.511	0.518
SEQUOIA	31	0.390	9.206	0.015	0.719	0.734
SEQUOIA	42	0.405	19.130	0.016	1.551	1.567
SEQUOIA	46	0.342	11.610	0.012	0.793	0.805
HE2RNA	19	0.046	17.138	0.0002	0.156	0.156
DeepPT	15	0.828	26.183	0.069	4.335	4.403
DeepPT	16	0.837	18.311	0.070	3.065	3.136
DeepPT	19	0.588	16.596	0.035	1.951	1.986
DeepPT	21	0.437	9.142	0.019	0.799	0.818
DeepPT	31	0.669	8.927	0.045	1.194	1.239
DeepPT	42	0.594	18.941	0.035	2.252	2.287
DeepPT	46	0.577	11.374	0.033	1.313	1.346
Path2Omics	46	0.958	10.993	0.092	2.106	2.198
MORSA	41	0.279	-6.670	0.008	-0.373	-0.365
SEQUOIA	41	0.238	-6.629	0.006	-0.315	-0.310
HE2RNA	41	0.031	-6.421	0.0001	-0.039	-0.039
DeepPT	41	0.419	-6.810	0.018	-0.571	-0.553
Path2Omics	41	0.644	-7.035	0.042	-0.906	-0.865

Table 8: Hallmark gene-set identifiers. Names omit the common HALLMARK prefix.

ID	Set	ID	Set
01	ADIPOGENESIS	26	INTERFERON ALPHA RESPONSE
02	ALLOGRAFT REJECTION	27	INTERFERON GAMMA RESPONSE
03	ANDROGEN RESPONSE	28	KRAS SIGNALING DN
04	ANGIOGENESIS	29	KRAS SIGNALING UP
05	APICAL JUNCTION	30	MITOTIC SPINDLE
06	APICAL SURFACE	31	MTORC1 SIGNALING
07	APOPTOSIS	32	MYC TARGETS V1
08	BILE ACID METABOLISM	33	MYC TARGETS V2
09	CHOLESTEROL HOMEOSTASIS	34	MYOGENESIS
10	COAGULATION	35	NOTCH SIGNALING
11	COMPLEMENT	36	OXIDATIVE PHOSPHORYLATION
12	DNA REPAIR	37	P53 PATHWAY
13	E2F TARGETS	38	PANCREAS BETA CELLS
14	EPITHELIAL MESENCHYMAL TRANSITION	39	PEROXISOME
15	ESTROGEN RESPONSE EARLY	40	PI3K AKT MTOR SIGNALING

Table 8 (continued)

ID	Set	ID	Set
16	ESTROGEN RESPONSE LATE	41	PROTEIN SECRETION
17	FATTY ACID METABOLISM	42	REACTIVE OXYGEN SPECIES PATHWAY
18	G2M CHECKPOINT	43	SPERMATOGENESIS
19	GLYCOLYSIS	44	TGF BETA SIGNALING
20	HEDGEHOG SIGNALING	45	TNFA SIGNALING VIA NFKB
21	HEME METABOLISM	46	UNFOLDED PROTEIN RESPONSE
22	HYPOXIA	47	UV RESPONSE DN
23	IL2 STAT5 SIGNALING	48	UV RESPONSE UP
24	IL6 JAK STAT3 SIGNALING	49	WNT BETA CATENIN SIGNALING
25	INFLAMMATORY RESPONSE	50	XENOBIOTIC METABOLISM

Table 9: All 250 gene-set comparisons and 95% simultaneous intervals, in 10^{-5} score MSE. M/S/H/D/P denote MORSA, SEQUOIA, HE2RNA, DeepPT, and Path2Omics. Set identifiers appear in Table 8. Classifications use unrounded values.

Model/set	P_{gene}	P_{score}	Δ
MORSA			
M/01	2.17 [0.56, 3.77]	1.01 [−0.63, 2.66]	1.15 [0.62, 1.69]
M/02	43.14 [7.02, 79.27]	43.24 [7.04, 79.45]	−0.10 [−1.34, 1.14]
M/03	1.97 [−2.01, 5.95]	1.91 [−2.05, 5.87]	0.06 [−0.25, 0.37]
M/04	22.66 [−8.18, 53.51]	21.66 [−9.07, 52.39]	1.00 [−0.22, 2.23]
M/05	9.41 [−6.99, 25.81]	6.77 [−9.46, 23.01]	2.64 [1.21, 4.07]
M/06	3.79 [−10.81, 18.39]	2.47 [−12.15, 17.10]	1.32 [0.60, 2.03]
M/07	5.15 [−0.97, 11.27]	4.21 [−1.87, 10.29]	0.94 [0.28, 1.60]
M/08	2.09 [−3.65, 7.83]	2.04 [−3.63, 7.72]	0.05 [−0.13, 0.22]
M/09	1.65 [−0.81, 4.12]	0.27 [−2.03, 2.58]	1.38 [0.55, 2.21]
M/10	13.78 [−1.05, 28.62]	11.18 [−3.39, 25.74]	2.61 [0.71, 4.50]
M/11	14.91 [1.69, 28.13]	14.35 [1.11, 27.59]	0.56 [−0.66, 1.78]
M/12	1.01 [−0.42, 2.43]	0.04 [−1.42, 1.50]	0.97 [0.57, 1.37]
M/13	10.62 [1.04, 20.20]	10.30 [0.88, 19.71]	0.33 [−0.29, 0.95]
M/14	73.64 [−3.93, 151.21]	69.62 [−6.35, 145.58]	4.02 [−0.90, 8.95]
M/15	3.27 [−1.42, 7.96]	0.39 [−4.42, 5.20]	2.88 [1.56, 4.20]
M/16	2.81 [0.56, 5.07]	0.76 [−1.51, 3.03]	2.06 [1.17, 2.94]
M/17	3.18 [−1.23, 7.59]	2.17 [−2.14, 6.49]	1.01 [0.47, 1.55]
M/18	8.47 [0.63, 16.32]	7.81 [0.16, 15.45]	0.67 [0.002, 1.33]
M/19	3.68 [0.84, 6.52]	2.23 [−0.65, 5.10]	1.45 [0.78, 2.12]
M/20	6.12 [−16.64, 28.88]	4.35 [−18.29, 26.99]	1.77 [0.54, 3.00]
M/21	2.96 [0.84, 5.09]	2.35 [0.25, 4.45]	0.61 [0.36, 0.87]
M/22	10.70 [0.66, 20.75]	8.46 [−1.62, 18.55]	2.24 [0.86, 3.62]
M/23	9.52 [−1.22, 20.26]	9.24 [−1.50, 19.99]	0.28 [−0.20, 0.76]
M/24	25.32 [−0.92, 51.55]	25.50 [−0.81, 51.80]	−0.18 [−1.42, 1.06]
M/25	40.81 [3.63, 78.00]	41.30 [3.87, 78.73]	−0.49 [−1.97, 0.99]
M/26	26.77 [0.22, 53.32]	26.40 [−0.20, 53.01]	0.37 [−1.24, 1.97]
M/27	42.47 [8.50, 76.44]	42.52 [8.40, 76.64]	−0.05 [−1.34, 1.24]
M/28	0.68 [−2.66, 4.02]	0.25 [−3.21, 3.72]	0.43 [0.09, 0.77]
M/29	13.24 [−5.38, 31.85]	13.31 [−5.23, 31.85]	−0.07 [−1.51, 1.37]
M/30	0.75 [−4.75, 6.25]	0.14 [−5.39, 5.68]	0.61 [0.21, 1.01]
M/31	3.14 [0.62, 5.65]	2.41 [0.05, 4.76]	0.73 [0.26, 1.20]
M/32	1.19 [−0.10, 2.48]	0.82 [−0.45, 2.09]	0.37 [0.14, 0.60]
M/33	6.74 [−0.59, 14.08]	5.20 [−2.02, 12.42]	1.54 [0.65, 2.44]
M/34	23.21 [−0.45, 46.87]	20.86 [−2.19, 43.91]	2.35 [0.53, 4.17]
M/35	−1.38 [−10.84, 8.08]	−2.32 [−11.77, 7.14]	0.93 [0.41, 1.45]
M/36	−0.002 [−3.98, 3.97]	−0.60 [−4.69, 3.50]	0.59 [0.16, 1.03]
M/37	2.21 [−0.61, 5.03]	1.04 [−1.81, 3.89]	1.17 [0.69, 1.64]
M/38	7.13 [−3.19, 17.46]	5.85 [−4.35, 16.04]	1.29 [−0.60, 3.18]
M/39	3.28 [0.04, 6.51]	2.58 [−0.55, 5.72]	0.69 [0.29, 1.10]
M/40	0.07 [−0.41, 0.54]	−0.26 [−0.71, 0.18]	0.33 [0.12, 0.54]

Table 9 (continued)

Model/set	P_{gene}	P_{score}	Δ
M/41	-1.13 [$-2.77, 0.50$]	-0.77 [$-2.39, 0.86$]	-0.37 [$-0.64, -0.09$]
M/42	2.75 [0.70, 4.81]	1.42 [$-0.55, 3.38$]	1.34 [0.94, 1.74]
M/43	0.47 [$-1.50, 2.44$]	0.75 [$-1.17, 2.66$]	-0.28 [$-0.60, 0.04$]
M/44	5.93 [$-5.82, 17.69$]	4.38 [$-7.41, 16.16$]	1.55 [0.38, 2.73]
M/45	21.72 [$-12.49, 55.93$]	21.07 [$-13.37, 55.50$]	0.65 [$-0.72, 2.03$]
M/46	1.44 [0.03, 2.85]	0.66 [$-0.70, 2.02$]	0.78 [0.45, 1.11]
M/47	3.98 [$-11.57, 19.53$]	3.67 [$-11.68, 19.02$]	0.31 [$-0.43, 1.06$]
M/48	1.64 [0.28, 2.99]	0.50 [$-0.72, 1.72$]	1.14 [0.63, 1.64]
M/49	9.28 [0.16, 18.40]	8.31 [$-0.77, 17.40$]	0.97 [0.09, 1.84]
M/50	1.84 [$-1.06, 4.73$]	0.83 [$-2.03, 3.68$]	1.01 [0.50, 1.52]
SEQUOIA			
S/01	1.89 [0.20, 3.59]	0.70 [$-0.83, 2.23$]	1.20 [0.57, 1.82]
S/02	29.97 [$-12.84, 72.78$]	30.07 [$-12.82, 72.95$]	-0.10 [$-1.30, 1.11$]
S/03	-0.26 [$-5.28, 4.75$]	-0.31 [$-5.34, 4.73$]	0.04 [$-0.17, 0.26$]
S/04	15.62 [$-12.94, 44.18$]	14.79 [$-13.87, 43.45$]	0.83 [$-0.19, 1.86$]
S/05	4.20 [$-10.08, 18.49$]	2.06 [$-12.16, 16.29$]	2.14 [0.79, 3.49]
S/06	0.03 [$-15.32, 15.37$]	-0.43 [$-15.76, 14.91$]	0.46 [$-0.07, 0.99$]
S/07	4.29 [$-2.37, 10.96$]	3.49 [$-3.15, 10.13$]	0.80 [0.17, 1.44]
S/08	2.20 [$-3.32, 7.71$]	2.16 [$-3.30, 7.61$]	0.04 [$-0.11, 0.18$]
S/09	1.10 [$-1.53, 3.73$]	-0.35 [$-2.82, 2.12$]	1.45 [0.43, 2.47]
S/10	7.82 [$-6.85, 22.50$]	5.59 [$-8.85, 20.03$]	2.23 [0.53, 3.93]
S/11	11.18 [$-4.67, 27.02$]	10.67 [$-5.20, 26.54$]	0.50 [$-0.60, 1.60$]
S/12	0.22 [$-1.97, 2.41$]	-0.79 [$-3.01, 1.44$]	1.01 [0.48, 1.54]
S/13	3.93 [$-2.63, 10.50$]	3.60 [$-2.96, 10.16$]	0.33 [$-0.31, 0.97$]
S/14	45.18 [$-20.02, 110.38$]	41.84 [$-22.51, 106.19$]	3.34 [$-0.70, 7.38$]
S/15	1.83 [$-2.64, 6.30$]	-0.25 [$-5.04, 4.53$]	2.08 [0.84, 3.32]
S/16	1.96 [$-0.20, 4.11$]	0.34 [$-1.70, 2.38$]	1.62 [0.67, 2.56]
S/17	2.98 [$-1.64, 7.59$]	2.06 [$-2.44, 6.57$]	0.91 [0.35, 1.48]
S/18	3.87 [$-1.39, 9.12$]	3.27 [$-1.87, 8.40$]	0.60 [$-0.02, 1.23$]
S/19	2.77 [0.30, 5.23]	1.46 [$-1.03, 3.95$]	1.31 [0.54, 2.08]
S/20	1.70 [$-20.90, 24.31$]	0.71 [$-21.95, 23.36$]	1.00 [0.18, 1.81]
S/21	2.59 [0.47, 4.71]	2.07 [$-0.02, 4.17$]	0.52 [0.26, 0.78]
S/22	8.33 [$-2.98, 19.64$]	6.37 [$-5.18, 17.92$]	1.96 [0.61, 3.32]
S/23	6.99 [$-5.88, 19.86$]	6.76 [$-6.11, 19.64$]	0.23 [$-0.17, 0.63$]
S/24	15.96 [$-15.30, 47.21$]	16.10 [$-15.22, 47.42$]	-0.14 [$-1.16, 0.87$]
S/25	26.68 [$-14.98, 68.34$]	27.06 [$-14.77, 68.89$]	-0.38 [$-1.53, 0.78$]
S/26	18.36 [$-9.49, 46.22$]	18.07 [$-9.83, 45.96$]	0.30 [$-0.98, 1.57$]
S/27	29.68 [$-8.19, 67.55$]	29.72 [$-8.25, 67.69$]	-0.04 [$-1.13, 1.05$]
S/28	-0.24 [$-3.79, 3.32$]	-0.50 [$-4.12, 3.13$]	0.26 [$-0.07, 0.59$]
S/29	7.75 [$-14.01, 29.52$]	7.80 [$-14.02, 29.62$]	-0.04 [$-1.14, 1.05$]
S/30	0.99 [$-5.19, 7.17$]	0.43 [$-5.77, 6.63$]	0.56 [0.17, 0.96]
S/31	1.94 [0.43, 3.45]	1.21 [$-0.16, 2.57$]	0.73 [0.22, 1.25]
S/32	0.58 [$-0.71, 1.87$]	0.17 [$-1.11, 1.46$]	0.41 [0.13, 0.69]
S/33	4.82 [$-7.09, 16.73$]	2.72 [$-8.70, 14.13$]	2.10 [0.62, 3.58]
S/34	8.56 [$-8.93, 26.04$]	6.86 [$-10.28, 24.00$]	1.70 [0.27, 3.13]
S/35	-1.20 [$-9.25, 6.86$]	-1.96 [$-10.12, 6.20$]	0.76 [0.18, 1.34]
S/36	-0.07 [$-4.62, 4.48$]	-0.77 [$-5.41, 3.87$]	0.70 [0.15, 1.25]
S/37	1.28 [$-1.06, 3.62$]	0.18 [$-2.28, 2.65$]	1.10 [0.53, 1.67]
S/38	3.84 [$-5.21, 12.89$]	3.33 [$-5.69, 12.34$]	0.51 [$-0.23, 1.26$]
S/39	2.82 [$-1.11, 6.75$]	2.12 [$-1.68, 5.92$]	0.70 [0.24, 1.16]
S/40	0.15 [$-0.22, 0.53$]	-0.15 [$-0.49, 0.18$]	0.31 [0.09, 0.53]
S/41	-3.67 [$-7.34, 0.0002$]	-3.36 [$-7.09, 0.37$]	-0.31 [$-0.57, -0.05$]
S/42	2.95 [0.44, 5.46]	1.38 [$-1.04, 3.80$]	1.57 [0.99, 2.15]
S/43	-0.42 [$-1.91, 1.07$]	-0.20 [$-1.61, 1.22$]	-0.23 [$-0.52, 0.07$]
S/44	4.88 [$-8.76, 18.51$]	3.58 [$-10.18, 17.34$]	1.30 [0.19, 2.40]
S/45	16.92 [$-18.91, 52.75$]	16.39 [$-19.69, 52.47$]	0.52 [$-0.61, 1.66$]
S/46	1.33 [0.28, 2.37]	0.52 [$-0.43, 1.48$]	0.81 [0.41, 1.20]
S/47	2.03 [$-14.26, 18.33$]	1.77 [$-14.37, 17.92$]	0.26 [$-0.36, 0.89$]
S/48	0.97 [$-0.44, 2.38$]	-0.04 [$-1.42, 1.34$]	1.01 [0.42, 1.60]
S/49	5.98 [$-1.84, 13.80$]	5.44 [$-2.06, 12.95$]	0.54 [$-0.28, 1.36$]
S/50	2.16 [$-1.15, 5.46$]	1.26 [$-1.90, 4.42$]	0.90 [0.35, 1.44]

Table 9 (continued)

Model/set	P_{gene}	P_{score}	Δ
HE2RNA			
H/01	0.88 [0.22, 1.53]	0.66 [0.002, 1.31]	0.22 [0.10, 0.35]
H/02	11.85 [0.58, 23.12]	11.86 [0.56, 23.16]	−0.01 [−0.15, 0.13]
H/03	0.09 [−1.97, 2.15]	0.09 [−1.97, 2.15]	0.004 [−0.01, 0.02]
H/04	4.23 [−6.90, 15.36]	4.20 [−6.96, 15.36]	0.02 [−0.12, 0.17]
H/05	3.01 [−4.02, 10.04]	2.78 [−4.24, 9.80]	0.23 [0.02, 0.45]
H/06	1.81 [−2.60, 6.21]	1.57 [−2.86, 6.00]	0.24 [0.01, 0.47]
H/07	1.47 [−0.47, 3.40]	1.34 [−0.61, 3.29]	0.13 [0.02, 0.24]
H/08	0.58 [−0.84, 2.00]	0.57 [−0.84, 1.98]	0.01 [−0.02, 0.04]
H/09	0.12 [−0.55, 0.79]	−0.04 [−0.71, 0.64]	0.16 [0.04, 0.27]
H/10	3.57 [−2.05, 9.19]	3.30 [−2.34, 8.94]	0.27 [0.04, 0.50]
H/11	4.36 [−0.12, 8.85]	4.27 [−0.24, 8.78]	0.09 [−0.10, 0.29]
H/12	−0.18 [−1.20, 0.85]	−0.28 [−1.34, 0.77]	0.11 [0.01, 0.20]
H/13	1.22 [−0.68, 3.13]	1.19 [−0.72, 3.11]	0.03 [−0.04, 0.10]
H/14	18.47 [−16.54, 53.48]	17.80 [−17.05, 52.66]	0.67 [−0.13, 1.47]
H/15	0.23 [−1.60, 2.06]	−0.10 [−1.99, 1.79]	0.33 [0.16, 0.49]
H/16	0.28 [−0.61, 1.16]	−0.01 [−0.95, 0.94]	0.28 [0.15, 0.42]
H/17	0.39 [−1.11, 1.89]	0.33 [−1.18, 1.84]	0.06 [−0.01, 0.12]
H/18	1.05 [−0.50, 2.59]	0.96 [−0.57, 2.48]	0.09 [−0.004, 0.19]
H/19	0.56 [0.01, 1.11]	0.41 [−0.15, 0.97]	0.16 [0.04, 0.27]
H/20	2.43 [−7.54, 12.40]	1.98 [−7.96, 11.92]	0.45 [0.06, 0.84]
H/21	0.44 [−0.14, 1.02]	0.41 [−0.17, 0.99]	0.03 [−0.01, 0.08]
H/22	2.67 [−1.01, 6.36]	2.18 [−1.52, 5.89]	0.49 [0.20, 0.78]
H/23	2.75 [−0.61, 6.10]	2.69 [−0.67, 6.05]	0.06 [−0.04, 0.15]
H/24	6.90 [−0.66, 14.46]	6.92 [−0.66, 14.50]	−0.02 [−0.15, 0.11]
H/25	9.08 [−1.44, 19.60]	9.15 [−1.41, 19.71]	−0.08 [−0.31, 0.16]
H/26	5.83 [−0.22, 11.88]	5.80 [−0.26, 11.86]	0.03 [−0.10, 0.17]
H/27	9.74 [0.38, 19.10]	9.74 [0.37, 19.11]	−0.001 [−0.05, 0.05]
H/28	0.31 [−1.03, 1.65]	0.33 [−1.01, 1.68]	−0.02 [−0.08, 0.03]
H/29	4.02 [−3.10, 11.14]	4.03 [−3.11, 11.16]	−0.01 [−0.25, 0.24]
H/30	0.28 [−1.52, 2.07]	0.18 [−1.61, 1.98]	0.09 [0.03, 0.16]
H/31	0.47 [0.12, 0.83]	0.37 [0.04, 0.69]	0.11 [0.03, 0.19]
H/32	0.18 [−0.28, 0.63]	0.16 [−0.30, 0.62]	0.02 [−0.01, 0.05]
H/33	1.43 [−1.80, 4.66]	1.28 [−1.93, 4.50]	0.14 [0.002, 0.29]
H/34	8.06 [−1.98, 18.11]	7.71 [−2.27, 17.69]	0.35 [0.05, 0.66]
H/35	−0.48 [−4.23, 3.28]	−0.65 [−4.42, 3.11]	0.17 [0.02, 0.33]
H/36	−0.49 [−2.73, 1.74]	−0.61 [−2.88, 1.67]	0.11 [0.02, 0.21]
H/37	0.32 [−0.10, 0.74]	0.20 [−0.23, 0.63]	0.12 [0.02, 0.22]
H/38	0.91 [−2.23, 4.06]	0.95 [−2.18, 4.08]	−0.04 [−0.14, 0.07]
H/39	0.90 [−0.34, 2.13]	0.86 [−0.36, 2.08]	0.04 [−0.02, 0.10]
H/40	0.03 [−0.10, 0.17]	−0.001 [−0.13, 0.13]	0.03 [0.005, 0.06]
H/41	−0.61 [−1.64, 0.43]	−0.57 [−1.60, 0.46]	−0.04 [−0.08, −0.0004]
H/42	0.59 [−0.005, 1.19]	0.21 [−0.38, 0.80]	0.38 [0.23, 0.53]
H/43	−0.19 [−0.62, 0.24]	−0.17 [−0.59, 0.26]	−0.02 [−0.06, 0.01]
H/44	1.02 [−6.32, 8.35]	0.83 [−6.54, 8.19]	0.19 [0.05, 0.34]
H/45	4.89 [−3.80, 13.59]	4.83 [−3.90, 13.55]	0.07 [−0.07, 0.21]
H/46	0.14 [−0.12, 0.40]	0.07 [−0.18, 0.31]	0.07 [0.02, 0.13]
H/47	0.53 [−8.10, 9.16]	0.45 [−8.13, 9.03]	0.08 [−0.10, 0.27]
H/48	0.16 [−0.27, 0.60]	0.10 [−0.35, 0.54]	0.07 [−0.004, 0.14]
H/49	1.80 [−1.70, 5.30]	2.02 [−1.44, 5.47]	−0.21 [−0.56, 0.13]
H/50	0.48 [−0.40, 1.36]	0.48 [−0.38, 1.34]	−0.001 [−0.10, 0.10]
DeepPT			
D/01	2.35 [−0.52, 5.22]	0.75 [−2.07, 3.58]	1.60 [0.84, 2.35]
D/02	42.78 [3.19, 82.38]	42.94 [3.30, 82.59]	−0.16 [−2.01, 1.69]
D/03	−0.21 [−5.75, 5.33]	−0.31 [−5.96, 5.34]	0.10 [−0.48, 0.67]
D/04	11.39 [−14.06, 36.83]	10.09 [−15.41, 35.58]	1.30 [−0.23, 2.83]
D/05	4.85 [−8.30, 18.00]	2.33 [−10.64, 15.31]	2.52 [0.98, 4.06]
D/06	1.42 [−13.68, 16.52]	−3.01 [−18.71, 12.68]	4.44 [2.13, 6.74]
D/07	4.42 [−1.53, 10.37]	3.24 [−2.69, 9.18]	1.18 [0.28, 2.08]
D/08	2.22 [−5.72, 10.16]	2.19 [−5.71, 10.10]	0.03 [−0.08, 0.14]
D/09	2.58 [−1.45, 6.62]	0.17 [−3.22, 3.55]	2.42 [0.91, 3.92]
D/10	7.69 [−5.16, 20.55]	5.32 [−7.21, 17.86]	2.37 [0.59, 4.15]

Table 9 (continued)

Model/set	P_{gene}	P_{score}	Δ
D/11	14.02 [−1.46, 29.49]	13.32 [−1.96, 28.61]	0.69 [−0.89, 2.27]
D/12	0.94 [−1.50, 3.38]	−0.66 [−2.87, 1.54]	1.60 [0.88, 2.32]
D/13	11.66 [−5.28, 28.60]	11.06 [−5.42, 27.55]	0.60 [−0.58, 1.78]
D/14	42.37 [−9.50, 94.25]	38.48 [−12.16, 89.13]	3.89 [−0.85, 8.63]
D/15	4.87 [0.55, 9.20]	0.47 [−3.40, 4.35]	4.40 [2.48, 6.32]
D/16	4.14 [0.75, 7.52]	1.00 [−1.80, 3.80]	3.14 [1.80, 4.47]
D/17	2.46 [−3.11, 8.03]	1.09 [−4.33, 6.51]	1.37 [0.61, 2.13]
D/18	8.82 [−4.45, 22.09]	7.73 [−5.06, 20.51]	1.09 [−0.04, 2.23]
D/19	4.16 [0.86, 7.45]	2.17 [−1.03, 5.37]	1.99 [1.02, 2.95]
D/20	9.47 [−15.15, 34.09]	4.69 [−19.48, 28.86]	4.79 [1.79, 7.78]
D/21	2.74 [0.01, 5.48]	1.93 [−0.71, 4.56]	0.82 [0.39, 1.24]
D/22	7.69 [−2.12, 17.51]	4.79 [−5.40, 14.98]	2.90 [1.00, 4.81]
D/23	10.11 [−1.57, 21.79]	9.71 [−1.91, 21.34]	0.40 [−0.32, 1.11]
D/24	18.38 [−10.25, 47.01]	18.70 [−9.95, 47.34]	−0.31 [−2.30, 1.68]
D/25	29.01 [−8.36, 66.38]	29.65 [−7.85, 67.15]	−0.65 [−2.52, 1.22]
D/26	26.79 [−0.93, 54.52]	26.34 [−1.52, 54.19]	0.46 [−1.62, 2.53]
D/27	40.11 [3.50, 76.73]	40.20 [3.49, 76.90]	−0.08 [−1.92, 1.76]
D/28	1.39 [−6.15, 8.92]	0.98 [−6.39, 8.35]	0.41 [−0.17, 0.98]
D/29	10.62 [−7.77, 29.02]	10.74 [−7.60, 29.08]	−0.11 [−2.04, 1.82]
D/30	0.78 [−6.88, 8.44]	−0.17 [−7.74, 7.39]	0.95 [0.35, 1.55]
D/31	4.11 [0.12, 8.10]	2.87 [−0.82, 6.57]	1.24 [0.40, 2.08]
D/32	1.64 [−0.47, 3.75]	0.94 [−0.97, 2.86]	0.69 [0.25, 1.14]
D/33	6.66 [−8.41, 21.73]	2.96 [−11.20, 17.11]	3.70 [1.53, 5.87]
D/34	16.70 [−3.50, 36.90]	14.57 [−5.27, 34.42]	2.13 [0.35, 3.90]
D/35	0.39 [−9.31, 10.09]	−1.31 [−11.05, 8.42]	1.71 [0.94, 2.47]
D/36	−0.32 [−4.31, 3.67]	−1.34 [−5.38, 2.70]	1.02 [0.28, 1.76]
D/37	2.42 [−0.29, 5.12]	0.39 [−2.64, 3.42]	2.03 [1.05, 3.00]
D/38	2.17 [−10.69, 15.04]	2.55 [−10.23, 15.32]	−0.37 [−1.11, 0.36]
D/39	2.89 [−1.03, 6.81]	2.24 [−1.71, 6.19]	0.65 [0.17, 1.13]
D/40	0.62 [−0.11, 1.36]	0.20 [−0.47, 0.88]	0.42 [0.14, 0.71]
D/41	−5.24 [−10.69, 0.21]	−4.69 [−10.16, 0.78]	−0.55 [−1.00, −0.10]
D/42	3.01 [0.15, 5.86]	0.72 [−2.05, 3.49]	2.29 [1.51, 3.07]
D/43	−1.11 [−5.86, 3.64]	−0.80 [−5.47, 3.87]	−0.31 [−0.71, 0.09]
D/44	7.90 [−4.22, 20.01]	5.80 [−6.39, 18.00]	2.09 [0.63, 3.55]
D/45	13.10 [−25.85, 52.05]	12.21 [−27.10, 51.52]	0.89 [−1.09, 2.87]
D/46	2.12 [0.02, 4.22]	0.77 [−1.21, 2.76]	1.35 [0.74, 1.95]
D/47	4.10 [−8.60, 16.81]	3.81 [−8.77, 16.38]	0.29 [−0.40, 0.99]
D/48	2.37 [−0.10, 4.85]	0.73 [−1.69, 3.14]	1.64 [0.76, 2.53]
D/49	5.91 [−13.84, 25.66]	4.62 [−15.02, 24.25]	1.29 [−0.49, 3.07]
D/50	1.76 [−3.00, 6.53]	0.86 [−3.69, 5.41]	0.91 [0.30, 1.51]
Path2Omics			
P/01	3.51 [−0.24, 7.26]	1.00 [−2.56, 4.56]	2.51 [1.23, 3.79]
P/02	42.72 [−5.06, 90.51]	42.96 [−4.83, 90.74]	−0.24 [−2.81, 2.33]
P/03	−4.47 [−12.56, 3.62]	−4.59 [−12.76, 3.58]	0.12 [−0.75, 0.99]
P/04	7.81 [−25.16, 40.78]	6.36 [−26.71, 39.43]	1.46 [−0.30, 3.21]
P/05	1.19 [−18.66, 21.04]	−3.10 [−22.29, 16.08]	4.29 [1.64, 6.95]
P/06	−6.61 [−28.23, 15.02]	−13.11 [−35.39, 9.17]	6.50 [2.62, 10.39]
P/07	4.23 [−3.71, 12.16]	2.49 [−5.34, 10.32]	1.74 [0.34, 3.13]
P/08	3.49 [−4.19, 11.16]	3.47 [−4.19, 11.13]	0.02 [−0.09, 0.13]
P/09	3.41 [−2.64, 9.46]	−0.28 [−5.36, 4.79]	3.69 [1.31, 6.07]
P/10	7.78 [−12.53, 28.09]	4.19 [−15.67, 24.05]	3.59 [0.79, 6.39]
P/11	13.19 [−6.37, 32.75]	12.30 [−6.96, 31.55]	0.90 [−1.24, 3.04]
P/12	−0.40 [−4.47, 3.67]	−3.25 [−7.43, 0.93]	2.85 [1.48, 4.22]
P/13	8.83 [−12.36, 30.01]	7.77 [−12.98, 28.52]	1.06 [−1.11, 3.23]
P/14	35.84 [−35.84, 107.51]	30.05 [−40.00, 100.10]	5.79 [−1.47, 13.05]
P/15	4.43 [−2.73, 11.60]	−1.42 [−7.92, 5.08]	5.85 [3.05, 8.66]
P/16	4.55 [−0.47, 9.56]	0.22 [−3.96, 4.40]	4.32 [2.27, 6.38]
P/17	3.05 [−3.41, 9.51]	0.83 [−5.50, 7.16]	2.22 [1.04, 3.41]
P/18	8.67 [−6.63, 23.97]	6.88 [−7.89, 21.64]	1.79 [−0.11, 3.69]
P/19	4.16 [−0.32, 8.65]	1.08 [−3.03, 5.20]	3.08 [1.52, 4.63]
P/20	2.03 [−34.55, 38.62]	−4.35 [−39.96, 31.26]	6.38 [1.88, 10.89]
P/21	2.33 [−0.86, 5.52]	1.23 [−1.92, 4.38]	1.10 [0.53, 1.68]

Table 9 (continued)

Model/set	P_{gene}	P_{score}	Δ
P/22	$8.21\ [-4.80, 21.23]$	$4.17\ [-9.21, 17.56]$	$4.04\ [1.29, 6.79]$
P/23	$8.76\ [-6.75, 24.27]$	$8.21\ [-7.17, 23.59]$	$0.55\ [-0.49, 1.58]$
P/24	$15.16\ [-20.96, 51.27]$	$15.62\ [-20.36, 51.60]$	$-0.46\ [-3.22, 2.29]$
P/25	$24.13\ [-21.10, 69.36]$	$25.02\ [-20.39, 70.42]$	$-0.88\ [-3.37, 1.60]$
P/26	$27.45\ [-2.49, 57.39]$	$26.82\ [-3.33, 56.97]$	$0.63\ [-2.39, 3.64]$
P/27	$38.54\ [-3.04, 80.12]$	$38.69\ [-3.01, 80.39]$	$-0.15\ [-2.88, 2.58]$
P/28	$-3.37\ [-14.28, 7.54]$	$-3.70\ [-14.51, 7.12]$	$0.33\ [-0.42, 1.08]$
P/29	$8.12\ [-16.76, 33.00]$	$8.28\ [-16.58, 33.15]$	$-0.16\ [-2.55, 2.22]$
P/30	$0.54\ [-7.93, 9.00]$	$-1.02\ [-9.30, 7.25]$	$1.56\ [0.52, 2.59]$
P/31	$4.24\ [-0.93, 9.40]$	$2.23\ [-2.54, 7.00]$	$2.01\ [0.59, 3.43]$
P/32	$1.16\ [-1.50, 3.81]$	$-0.09\ [-2.69, 2.50]$	$1.25\ [0.41, 2.08]$
P/33	$2.97\ [-17.51, 23.45]$	$-4.17\ [-23.53, 15.18]$	$7.14\ [2.72, 11.56]$
P/34	$12.18\ [-18.67, 43.02]$	$8.95\ [-21.37, 39.28]$	$3.22\ [0.40, 6.04]$
P/35	$-4.53\ [-20.14, 11.08]$	$-6.55\ [-22.14, 9.05]$	$2.01\ [0.87, 3.15]$
P/36	$-0.72\ [-5.49, 4.04]$	$-2.40\ [-7.40, 2.59]$	$1.68\ [0.45, 2.91]$
P/37	$2.50\ [-1.14, 6.14]$	$-0.58\ [-4.44, 3.29]$	$3.07\ [1.59, 4.56]$
P/38	$5.62\ [-9.63, 20.86]$	$6.15\ [-9.15, 21.45]$	$-0.53\ [-1.48, 0.41]$
P/39	$3.61\ [-0.48, 7.69]$	$2.49\ [-1.51, 6.49]$	$1.12\ [0.34, 1.90]$
P/40	$0.67\ [-0.37, 1.72]$	$0.01\ [-0.96, 0.97]$	$0.67\ [0.19, 1.15]$
P/41	$-11.12\ [-19.29, -2.95]$	$-10.26\ [-18.35, -2.17]$	$-0.86\ [-1.58, -0.15]$
P/42	$3.02\ [-1.19, 7.24]$	$-0.26\ [-4.36, 3.85]$	$3.28\ [2.09, 4.47]$
P/43	$-1.56\ [-7.66, 4.55]$	$-1.12\ [-7.04, 4.81]$	$-0.44\ [-1.03, 0.16]$
P/44	$4.57\ [-12.10, 21.24]$	$1.72\ [-14.88, 18.31]$	$2.85\ [0.77, 4.94]$
P/45	$9.48\ [-31.41, 50.37]$	$8.19\ [-33.22, 49.60]$	$1.29\ [-1.66, 4.24]$
P/46	$2.76\ [0.17, 5.35]$	$0.56\ [-1.75, 2.87]$	$2.20\ [1.10, 3.29]$
P/47	$-2.73\ [-23.26, 17.81]$	$-3.06\ [-23.38, 17.25]$	$0.34\ [-0.48, 1.15]$
P/48	$2.03\ [-1.91, 5.96]$	$-0.60\ [-4.21, 3.02]$	$2.62\ [1.16, 4.09]$
P/49	$-0.43\ [-26.35, 25.49]$	$-1.82\ [-27.68, 24.03]$	$1.39\ [-1.00, 3.79]$
P/50	$2.72\ [-3.52, 8.97]$	$1.20\ [-4.62, 7.01]$	$1.53\ [0.55, 2.50]$

E.2 Reference-partition sensitivity

Table 10: Descriptive statistics across 256 reference partitions, each with eight reference patients and prior weight eight. Counts use the signs of 250 point estimates; correlations compare with the primary partition.

Statistic	Minimum	Median	Maximum
Count of $P_{\text{gene}} > 0$	214	225	233
Count of $P_{\text{score}} > 0$	190	201	215
Positive-count gap	15	24	32
Spearman P_{score}	0.934177	0.985229	0.994400

Table 11: Nine reference-count and prior-weight settings, each with 64 partitions. The gap counts positive P_{gene} minus positive P_{score} point estimates. Stable cases have positive Δ in at least 90% of partitions, among the 24 primary cases with positive paired intervals.

Reference count	Prior weight	Median gap	Positive gaps	Stable cases
4	4	23.0	100%	24
4	8	25.0	100%	24
4	16	24.0	100%	24
8	4	23.0	100%	24
8	8	24.5	100%	24
8	16	25.0	100%	24
16	4	23.0	100%	24
16	8	24.0	100%	24
16	16	25.0	100%	24

F Complete Drug Results

Table 12 reports all 28 replacement gains and variation contributions in pp, with the 88-contrast intervals. Table 13 gives the database-level H and descriptive reference gains R. Point estimates of S follow from $R + J$; their intervals are computed from the S draws. The alignment table uses the expanded 116-contrast family.

Table 12: All 28 drug comparisons in pp, with 95% simultaneous intervals over 88 contrasts.

Model	J	P
CCLE		
DIPK	-2.43 $[-30.52, 25.65]$	-15.71 $[-46.90, 15.48]$
ENet	13.87 $[-6.93, 34.68]$	0.59 $[-3.02, 4.21]$
GBDT	19.32 $[2.89, 35.75]$	6.04 $[-2.22, 14.30]$
RF	22.45 $[6.70, 38.20]$	9.18 $[-0.90, 19.25]$
SRMF	-76.49 $[-146.95, -6.02]$	-89.76 $[-159.64, -19.89]$
SNN	16.43 $[1.99, 30.87]$	3.15 $[-10.95, 17.25]$
SF	13.45 $[-7.93, 34.84]$	0.17 $[-0.01, 0.36]$
CTRPv1		
DIPK	26.25 $[2.17, 50.32]$	0.66 $[-11.82, 13.14]$
ENet	26.90 $[2.40, 51.41]$	1.31 $[-1.28, 3.91]$
GBDT	29.81 $[6.86, 52.76]$	4.22 $[-0.13, 8.58]$
RF	32.01 $[9.86, 54.17]$	6.42 $[0.70, 12.15]$
SRMF	-5.83 $[-41.60, 29.94]$	-31.42 $[-47.16, -15.68]$
SNN	28.24 $[4.34, 52.15]$	2.65 $[-4.83, 10.13]$
SF	25.66 $[0.72, 50.61]$	0.07 $[0.00, 0.14]$
GDSC1		
DIPK	5.82 $[2.05, 9.59]$	0.08 $[0.03, 0.13]$
ENet	5.83 $[2.05, 9.61]$	0.09 $[-0.51, 0.70]$
GBDT	5.93 $[2.16, 9.71]$	0.20 $[0.06, 0.33]$
RF	5.40 $[1.58, 9.22]$	-0.34 $[-0.59, -0.09]$
SRMF	-641.80 $[-702.20, -581.40]$	-647.54 $[-706.97, -588.10]$
SNN	4.25 $[0.37, 8.12]$	-1.49 $[-2.02, -0.97]$
SF	5.79 $[2.01, 9.57]$	0.05 $[0.03, 0.07]$
GDSC2		
DIPK	6.33 $[1.78, 10.89]$	0.13 $[0.05, 0.21]$
ENet	6.76 $[2.21, 11.32]$	0.56 $[-0.26, 1.39]$
GBDT	6.62 $[2.08, 11.17]$	0.43 $[0.24, 0.61]$
RF	5.65 $[1.03, 10.27]$	-0.55 $[-0.90, -0.20]$
SRMF	-1001.16 $[-1186.89, -815.42]$	-1007.35 $[-1192.74, -821.97]$
SNN	5.10 $[0.43, 9.77]$	-1.10 $[-1.83, -0.37]$
SF	6.20 $[1.63, 10.76]$	1.16×10^{-8} $[0.90\times10^{-8}, 1.41\times10^{-8}]$

Table 13: Reference gains and baseline-removal gains shared within each database, in pp. R is descriptive; H intervals belong to the 88-contrast family.

Database	R	H (95% simultaneous CI)
CCLE	17.18	13.28 $[-8.22, 34.78]$
CTRPv1	4.22	25.59 $[0.61, 50.57]$
GDSC1	4.80	5.74 $[1.96, 9.52]$
GDSC2	4.14	6.20 $[1.63, 10.76]$

Table 14: Alignment A, amplitude cost Q, and contribution $P = A - Q$, in pp. Intervals for A and P use the 116-contrast family; Q is a point estimate.

Model	A	Q	P
CCLE			
DIPK	49.13 $[4.17, 94.10]$	64.84	-15.71 $[-48.26, 16.84]$

Table 14 (continued)

Model	A	Q	P
ENet	$2.88\ [-2.15, 7.91]$	2.28	$0.59\ [-3.18, 4.37]$
GBDT	$9.99\ [-0.66, 20.64]$	3.95	$6.04\ [-2.58, 14.66]$
RF	$17.91\ [2.64, 33.17]$	8.73	$9.18\ [-1.34, 19.69]$
SRMF	$19.63\ [-5.22, 44.48]$	109.39	$-89.76\ [-162.70, -16.83]$
SNN	$21.82\ [2.42, 41.21]$	18.67	$3.15\ [-11.57, 17.87]$
SF	$0.17\ [-0.02, 0.37]$	0.00	$0.17\ [-0.02, 0.37]$
CTRPv1			
DIPK	$27.62\ [11.25, 43.99]$	26.96	$0.66\ [-12.37, 13.68]$
ENet	$2.53\ [-0.77, 5.83]$	1.21	$1.31\ [-1.39, 4.02]$
GBDT	$7.37\ [2.01, 12.73]$	3.14	$4.22\ [-0.32, 8.77]$
RF	$10.46\ [3.06, 17.86]$	4.04	$6.42\ [0.45, 12.40]$
SRMF	$14.36\ [3.44, 25.28]$	45.78	$-31.42\ [-47.85, -14.99]$
SNN	$12.38\ [2.76, 21.99]$	9.72	$2.65\ [-5.16, 10.46]$
SF	$0.07\ [0.00, 0.14]$	0.00	$0.07\ [0.00, 0.14]$
GDSC1			
DIPK	$0.10\ [0.04, 0.15]$	0.02	$0.08\ [0.03, 0.14]$
ENet	$0.80\ [0.18, 1.41]$	0.70	$0.09\ [-0.53, 0.72]$
GBDT	$0.30\ [0.16, 0.44]$	0.10	$0.20\ [0.06, 0.34]$
RF	$-0.08\ [-0.33, 0.17]$	0.26	$-0.34\ [-0.60, -0.08]$
SRMF	$4.11\ [-2.84, 11.07]$	651.65	$-647.54\ [-709.57, -585.50]$
SNN	$0.52\ [0.03, 1.00]$	2.01	$-1.49\ [-2.04, -0.95]$
SF	$0.06\ [0.04, 0.07]$	0.00	$0.05\ [0.03, 0.07]$
GDSC2			
DIPK	$0.15\ [0.07, 0.24]$	0.02	$0.13\ [0.05, 0.22]$
ENet	$1.37\ [0.48, 2.26]$	0.81	$0.56\ [-0.29, 1.42]$
GBDT	$0.55\ [0.35, 0.75]$	0.12	$0.43\ [0.23, 0.62]$
RF	$-0.21\ [-0.56, 0.15]$	0.34	$-0.55\ [-0.92, -0.18]$
SRMF	$-0.18\ [-14.33, 13.96]$	1007.17	$-1007.35\ [-1200.85, -813.86]$
SNN	$1.28\ [0.53, 2.03]$	2.38	$-1.10\ [-1.87, -0.34]$
SF	$1.16\ [0.89, 1.43] \times 10^{-8}$	2.57×10^{-17}	$1.16\ [0.89, 1.43] \times 10^{-8}$

Table 15: All 28 leave-one-cell-out ranges, in pp. Models, references, and source coefficients are fixed. n counts deletions; ranges summarize recomputed point estimates and are not confidence intervals.

Model	n	H range	P range	J range
CCLE				
DIPK	286	$[7.70, 14.26]$	$[-18.21, -9.31]$	$[-6.14, 2.94]$
ENet	286	$[7.70, 14.26]$	$[0.29, 1.01]$	$[8.27, 14.77]$
GBDT	286	$[7.70, 14.26]$	$[5.44, 7.29]$	$[14.59, 20.65]$
RF	286	$[7.70, 14.26]$	$[8.18, 10.54]$	$[17.85, 23.47]$
SRMF	286	$[7.70, 14.26]$	$[-111.55, -84.81]$	$[-97.38, -71.54]$
SNN	286	$[7.70, 14.26]$	$[1.89, 5.98]$	$[12.25, 17.77]$
SF	286	$[7.70, 14.26]$	$[0.16, 0.20]$	$[7.89, 14.44]$
CTRPv1				
DIPK	152	$[16.76, 26.48]$	$[-0.60, 2.68]$	$[17.56, 28.40]$
ENet	152	$[16.76, 26.48]$	$[1.17, 1.76]$	$[18.24, 28.18]$
GBDT	152	$[16.76, 26.48]$	$[3.90, 4.75]$	$[21.45, 30.76]$
RF	152	$[16.76, 26.48]$	$[5.65, 7.21]$	$[23.97, 33.05]$
SRMF	152	$[16.76, 26.48]$	$[-36.10, -30.12]$	$[-19.34, -4.06]$
SNN	152	$[16.76, 26.48]$	$[2.13, 4.15]$	$[19.54, 30.16]$
SF	152	$[16.76, 26.48]$	$[0.06, 0.08]$	$[16.84, 26.55]$
GDSC1				
DIPK	860	$[5.28, 5.83]$	$[0.08, 0.09]$	$[5.36, 5.91]$
ENet	860	$[5.28, 5.83]$	$[0.03, 0.13]$	$[5.37, 5.91]$
GBDT	860	$[5.28, 5.83]$	$[0.19, 0.20]$	$[5.47, 6.02]$
RF	860	$[5.28, 5.83]$	$[-0.36, -0.32]$	$[4.94, 5.49]$
SRMF	860	$[5.28, 5.83]$	$[-654.27, -646.11]$	$[-648.51, -640.30]$
SNN	860	$[5.28, 5.83]$	$[-1.53, -1.45]$	$[3.78, 4.33]$

Table 15 (continued)

Model	n	H range	P range	J range
SF	860	$[5.28, 5.83]$	$[0.05, 0.05]$	$[5.33, 5.88]$
GDSC2				
DIPK	857	$[5.57, 6.36]$	$[0.12, 0.14]$	$[5.71, 6.49]$
ENet	857	$[5.57, 6.36]$	$[0.52, 0.63]$	$[6.15, 6.90]$
GBDT	857	$[5.57, 6.36]$	$[0.41, 0.44]$	$[6.00, 6.79]$
RF	857	$[5.57, 6.36]$	$[-0.58, -0.52]$	$[5.01, 5.81]$
SRMF	857	$[5.57, 6.36]$	$[-1037.36, -1003.63]$	$[-1031.13, -997.40]$
SNN	857	$[5.57, 6.36]$	$[-1.14, -1.01]$	$[4.46, 5.22]$
SF	857	$[5.57, 6.36]$	$[1.13\times10^{-8}, 1.17\times10^{-8}]$	$[5.57, 6.36]$

G Histology Risks and Source-Fitted Amplitudes

Table 16 summarizes original and risk-matched amplitude classifications. Three amplitudes and three risks give 135 original conditions: 70 positive, 19 negative, and 46 zero-crossing contribution intervals. Fitting an amplitude for each risk gives 45 risk-matched fits: 31 positive and 14 crossing zero. Table 18 reports coefficients and intervals. Table 19 gives all 45 paired differences from the original-scale fit, including increases, decreases, and unresolved differences.

Table 16: Positive/negative/zero-crossing interval counts by cancer and model. Original settings have nine conditions per row, totaling 70/19/46; risk-matched amplitude fits have three per row, totaling 31/0/14.

Cancer	Model	Original settings	Matched fits
BRCA	MORSA	9/0/0	3/0/0
BRCA	SEQUOIA	4/0/5	3/0/0
BRCA	HE2RNA	9/0/0	3/0/0
BRCA	DeepPT	2/3/4	0/0/3
BRCA	Path2Omics	1/2/6	0/0/3
COAD	MORSA	9/0/0	3/0/0
COAD	SEQUOIA	6/0/3	3/0/0
COAD	HE2RNA	2/0/7	0/0/3
COAD	DeepPT	2/4/3	1/0/2
COAD	Path2Omics	0/4/5	0/0/3
GBM	MORSA	6/0/3	3/0/0
GBM	SEQUOIA	6/0/3	3/0/0
GBM	HE2RNA	6/0/3	3/0/0
GBM	DeepPT	4/3/2	3/0/0
GBM	Path2Omics	4/3/2	3/0/0

Table 17: Contributions for all 135 original histology conditions, in pp. I/N/Z denote unit, original-scale fitted, and standardized fitted amplitudes. M/N/Z risk columns denote gene-macro, native-pooled, and source-standardized pooling. Intervals use 270 contrasts.

Amplitude	α	M	N	Z
BRCA / MORSA				
I	1.0000	7.59 [4.22, 10.95]	10.78 [6.41, 15.15]	5.40 [2.79, 8.00]
N	0.7561	7.15 [4.48, 9.82]	9.92 [6.36, 13.48]	4.97 [2.87, 7.07]
Z	0.7090	6.96 [4.43, 9.49]	9.63 [6.24, 13.01]	4.82 [2.82, 6.82]
BRCA / SEQUOIA				
I	1.0000	−0.14 [−6.24, 5.95]	3.67 [−2.77, 10.12]	0.51 [−3.76, 4.78]
N	0.6108	3.09 [−0.47, 6.65]	5.61 [1.58, 9.65]	2.29 [−0.31, 4.89]
Z	0.4684	3.26 [0.55, 5.97]	5.25 [2.10, 8.41]	2.31 [0.30, 4.32]

Table 17 (continued)

Amplitude	α	M	N	Z
BRCA / HE2RNA				
I	1.0000	1.72 [0.76, 2.67]	2.38 [1.06, 3.71]	1.22 [0.48, 1.96]
N	0.9938	1.71 [0.75, 2.66]	2.37 [1.05, 3.69]	1.21 [0.48, 1.95]
Z	0.8074	1.46 [0.67, 2.24]	2.01 [0.92, 3.10]	1.03 [0.42, 1.64]
BRCA / DeepPT				
I	1.0000	−43.08 [−64.58, −21.57]	−4.76 [−11.57, 2.05]	−11.19 [−16.14, −6.25]
N	0.4665	−6.35 [−11.56, −1.15]	3.13 [0.02, 6.25]	−0.35 [−2.42, 1.72]
Z	0.2456	−0.35 [−2.25, 1.55]	2.82 [1.14, 4.49]	0.88 [−0.23, 1.99]
BRCA / Path2Omics				
I	1.0000	−26.03 [−36.59, −15.48]	−7.67 [−16.01, 0.67]	−12.42 [−18.35, −6.50]
N	0.4913	−2.66 [−6.34, 1.01]	2.99 [−0.79, 6.77]	−0.47 [−2.93, 1.99]
Z	0.2934	0.76 [−1.28, 2.81]	3.36 [1.08, 5.64]	1.03 [−0.43, 2.48]
COAD / MORSA				
I	1.0000	3.68 [1.30, 6.07]	4.98 [2.62, 7.33]	2.95 [1.28, 4.62]
N	0.7010	3.55 [1.85, 5.24]	4.42 [2.68, 6.16]	2.66 [1.44, 3.88]
Z	0.6623	3.47 [1.86, 5.08]	4.29 [2.63, 5.95]	2.58 [1.43, 3.74]
COAD / SEQUOIA				
I	1.0000	−0.32 [−3.80, 3.15]	1.17 [−1.86, 4.19]	0.27 [−2.01, 2.54]
N	0.5306	1.91 [0.19, 3.63]	2.50 [0.86, 4.15]	1.44 [0.22, 2.66]
Z	0.4154	1.89 [0.52, 3.27]	2.32 [0.99, 3.66]	1.38 [0.39, 2.36]
COAD / HE2RNA				
I	1.0000	−1.09 [−2.41, 0.23]	0.81 [−0.04, 1.67]	0.21 [−0.40, 0.82]
N	0.8549	−0.69 [−1.75, 0.37]	0.74 [0.01, 1.47]	0.24 [−0.28, 0.76]
Z	0.6688	−0.30 [−1.07, 0.47]	0.63 [0.06, 1.20]	0.25 [−0.15, 0.65]
COAD / DeepPT				
I	1.0000	−40.50 [−51.68, −29.32]	−4.32 [−8.35, −0.30]	−11.89 [−16.44, −7.34]
N	0.4606	−5.74 [−8.51, −2.98]	1.90 [0.24, 3.55]	−0.70 [−2.26, 0.87]
Z	0.2582	−0.50 [−1.86, 0.85]	1.88 [0.86, 2.90]	0.61 [−0.29, 1.52]
COAD / Path2Omics				
I	1.0000	−70.89 [−95.91, −45.87]	−18.59 [−27.96, −9.22]	−26.71 [−37.07, −16.34]
N	0.4061	−8.23 [−12.82, −3.63]	0.31 [−2.09, 2.72]	−2.25 [−4.64, 0.14]
Z	0.2744	−2.47 [−5.06, 0.11]	1.39 [−0.25, 3.03]	−0.23 [−1.77, 1.31]
GBM / MORSA				
I	1.0000	0.60 [−2.40, 3.59]	1.48 [−1.60, 4.57]	0.53 [−1.58, 2.65]
N	0.3428	2.72 [1.68, 3.77]	2.88 [1.72, 4.04]	1.73 [0.97, 2.48]
Z	0.3249	2.65 [1.65, 3.64]	2.79 [1.68, 3.90]	1.68 [0.95, 2.40]
GBM / SEQUOIA				
I	1.0000	−1.40 [−4.44, 1.65]	0.45 [−2.54, 3.44]	−1.10 [−3.33, 1.13]
N	0.3075	1.68 [0.73, 2.62]	1.89 [0.89, 2.89]	1.06 [0.36, 1.76]
Z	0.2472	1.49 [0.73, 2.26]	1.64 [0.82, 2.46]	0.95 [0.39, 1.51]
GBM / HE2RNA				
I	1.0000	0.64 [−0.32, 1.60]	0.81 [−0.20, 1.82]	0.43 [−0.29, 1.15]
N	0.4492	0.60 [0.16, 1.03]	0.66 [0.19, 1.13]	0.40 [0.07, 0.73]
Z	0.4462	0.59 [0.16, 1.03]	0.65 [0.19, 1.12]	0.40 [0.07, 0.72]
GBM / DeepPT				
I	1.0000	−18.42 [−23.54, −13.30]	−5.88 [−9.17, −2.59]	−12.27 [−15.47, −9.06]
N	0.2734	1.02 [−0.02, 2.06]	1.84 [0.86, 2.83]	0.56 [−0.18, 1.30]
Z	0.1545	1.14 [0.55, 1.73]	1.36 [0.78, 1.94]	0.68 [0.26, 1.10]
GBM / Path2Omics				
I	1.0000	−22.74 [−29.14, −16.35]	−11.38 [−16.27, −6.49]	−14.61 [−18.86, −10.36]
N	0.2863	0.96 [−0.30, 2.22]	1.73 [0.51, 2.95]	0.53 [−0.38, 1.44]
Z	0.1993	1.30 [0.43, 2.17]	1.63 [0.77, 2.49]	0.77 [0.14, 1.40]

Table 18: All 45 risk-matched fits. n_{source} counts source patients; P is in pp. Intervals cover 360 contrasts.

Risk	α	n_{source}	P (95% simultaneous CI)
BRCA / MORSA			
Gene-macro	0.7024	477	$6.93\ [4.42, 9.44]$
Native-pooled	0.7667	477	$9.99\ [6.39, 13.58]$
Standardized	0.7204	477	$4.86\ [2.84, 6.88]$
BRCA / SEQUOIA			
Gene-macro	0.4574	477	$3.25\ [0.60, 5.90]$
Native-pooled	0.6262	477	$5.62\ [1.49, 9.75]$
Standardized	0.4934	477	$2.33\ [0.22, 4.44]$
BRCA / HE2RNA			
Gene-macro	1.0000	477	$1.72\ [0.76, 2.67]$
Native-pooled	1.0000	477	$2.38\ [1.06, 3.71]$
Standardized	1.0000	477	$1.22\ [0.48, 1.96]$
BRCA / DeepPT			
Gene-macro	0.2066	477	$0.15\ [-1.36, 1.66]$
Native-pooled	0.4707	477	$3.12\ [-0.02, 6.26]$
Standardized	0.2473	477	$0.88\ [-0.24, 1.99]$
BRCA / Path2Omics			
Gene-macro	0.2487	477	$1.10\ [-0.63, 2.82]$
Native-pooled	0.4916	477	$2.99\ [-0.79, 6.77]$
Standardized	0.2932	477	$1.03\ [-0.43, 2.48]$
COAD / MORSA			
Gene-macro	0.6462	386	$3.43\ [1.86, 5.01]$
Native-pooled	0.7162	386	$4.47\ [2.70, 6.24]$
Standardized	0.6799	386	$2.62\ [1.43, 3.80]$
COAD / SEQUOIA			
Gene-macro	0.3741	386	$1.83\ [0.58, 3.08]$
Native-pooled	0.5432	386	$2.51\ [0.84, 4.19]$
Standardized	0.4323	386	$1.40\ [0.38, 2.42]$
COAD / HE2RNA			
Gene-macro	0.9572	386	$-0.97\ [-2.21, 0.28]$
Native-pooled	1.0000	386	$0.81\ [-0.04, 1.67]$
Standardized	1.0000	386	$0.21\ [-0.40, 0.82]$
COAD / DeepPT			
Gene-macro	0.1719	386	$0.44\ [-0.47, 1.34]$
Native-pooled	0.4618	386	$1.89\ [0.23, 3.55]$
Standardized	0.2566	386	$0.62\ [-0.28, 1.52]$
COAD / Path2Omics			
Gene-macro	0.1871	386	$-0.29\ [-1.94, 1.35]$
Native-pooled	0.4046	386	$0.33\ [-2.06, 2.73]$
Standardized	0.2714	386	$-0.20\ [-1.72, 1.32]$
GBM / MORSA			
Gene-macro	0.2644	77	$2.33\ [1.51, 3.15]$
Native-pooled	0.2943	77	$2.62\ [1.61, 3.63]$
Standardized	0.2751	77	$1.51\ [0.90, 2.13]$
GBM / SEQUOIA			
Gene-macro	0.1751	77	$1.18\ [0.64, 1.73]$
Native-pooled	0.2641	77	$1.72\ [0.85, 2.59]$
Standardized	0.2082	77	$0.85\ [0.38, 1.33]$
GBM / HE2RNA			
Gene-macro	0.3065	77	$0.46\ [0.16, 0.76]$
Native-pooled	0.3729	77	$0.58\ [0.19, 0.97]$
Standardized	0.3325	77	$0.33\ [0.08, 0.57]$
GBM / DeepPT			
Gene-macro	0.0951	77	$0.87\ [0.50, 1.24]$
Native-pooled	0.2371	77	$1.75\ [0.88, 2.61]$
Standardized	0.1338	77	$0.64\ [0.28, 1.01]$

Table 18 (continued)

Risk	α	n_{source}	P (95% simultaneous CI)
GBM / Path2Omics			
Gene-macro	0.1164	77	1.11 [0.60, 1.63]
Native-pooled	0.2131	77	1.67 [0.75, 2.59]
Standardized	0.1486	77	0.75 [0.28, 1.22]

G.1 Direct paired amplitude comparisons

Table 19: All 45 paired differences: matched source amplitude minus original-scale fit. Contributions and differences are in pp; intervals use 405 contrasts. M/N/Z denote gene-macro, native-pooled, and source-standardized risks.

Model	Risk	Original P	Matched P	Difference	95% simultaneous CI
BRCA					
MORSA	M	7.147	6.927	-0.219	$[-0.391, -0.048]$
MORSA	N	9.924	9.985	0.061	$[0.020, 0.102]$
MORSA	Z	4.968	4.858	-0.111	$[-0.194, -0.027]$
SEQUOIA	M	3.090	3.252	0.161	$[-0.787, 1.110]$
SEQUOIA	N	5.615	5.620	0.005	$[-0.096, 0.105]$
SEQUOIA	Z	2.287	2.329	0.042	$[-0.464, 0.548]$
HE2RNA	M	1.708	1.715	0.008	$[0.002, 0.013]$
HE2RNA	N	2.372	2.383	0.011	$[0.004, 0.019]$
HE2RNA	Z	1.215	1.221	0.006	$[0.002, 0.010]$
DeepPT	M	-6.354	0.152	6.506	$[2.598, 10.413]$
DeepPT	N	3.132	3.118	-0.014	$[-0.042, 0.014]$
DeepPT	Z	-0.349	0.877	1.225	$[0.223, 2.228]$
Path2Omics	M	-2.663	1.096	3.760	$[1.631, 5.888]$
Path2Omics	N	2.991	2.989	-0.002	$[-0.004, 0.0003]$
Path2Omics	Z	-0.471	1.028	1.499	$[0.424, 2.574]$
COAD					
MORSA	M	3.546	3.432	-0.114	$[-0.248, 0.019]$
MORSA	N	4.421	4.469	0.047	$[0.012, 0.082]$
MORSA	Z	2.658	2.618	-0.040	$[-0.074, -0.005]$
SEQUOIA	M	1.908	1.834	-0.074	$[-0.602, 0.454]$
SEQUOIA	N	2.504	2.512	0.008	$[-0.031, 0.046]$
SEQUOIA	Z	1.443	1.398	-0.045	$[-0.268, 0.178]$
HE2RNA	M	-0.691	-0.966	-0.275	$[-0.467, -0.082]$
HE2RNA	N	0.743	0.814	0.071	$[-0.058, 0.200]$
HE2RNA	Z	0.241	0.211	-0.031	$[-0.128, 0.066]$
DeepPT	M	-5.744	0.435	6.179	$[4.054, 8.304]$
DeepPT	N	1.896	1.892	-0.004	$[-0.008, 0.001]$
DeepPT	Z	-0.699	0.617	1.316	$[0.553, 2.079]$
Path2Omics	M	-8.225	-0.294	7.931	$[4.613, 11.249]$
Path2Omics	N	0.315	0.334	0.019	$[0.007, 0.031]$
Path2Omics	Z	-2.252	-0.201	2.050	$[1.039, 3.062]$
GBM					
MORSA	M	2.725	2.333	-0.391	$[-0.623, -0.160]$
MORSA	N	2.879	2.622	-0.257	$[-0.412, -0.103]$
MORSA	Z	1.726	1.512	-0.213	$[-0.358, -0.069]$
SEQUOIA	M	1.676	1.184	-0.492	$[-0.892, -0.093]$
SEQUOIA	N	1.890	1.717	-0.173	$[-0.308, -0.037]$
SEQUOIA	Z	1.060	0.853	-0.206	$[-0.427, 0.014]$
HE2RNA	M	0.597	0.462	-0.135	$[-0.272, 0.003]$
HE2RNA	N	0.657	0.579	-0.078	$[-0.156, 0.0001]$
HE2RNA	Z	0.400	0.328	-0.072	$[-0.156, 0.013]$
DeepPT	M	1.021	0.872	-0.149	$[-0.837, 0.539]$
DeepPT	N	1.844	1.749	-0.095	$[-0.219, 0.029]$
DeepPT	Z	0.562	0.643	0.081	$[-0.298, 0.461]$

Table 19 (continued)

Model	Risk	Original P	Matched P	Difference	95% simultaneous CI
Path2Omics	M	0.958	1.112	0.155	$[-0.614, 0.924]$
Path2Omics	N	1.731	1.669	-0.061	$[-0.373, 0.250]$
Path2Omics	Z	0.531	0.748	0.217	$[-0.233, 0.667]$

G.2 All amplitude curves

For each cancer, model, and risk, squared-loss contribution has the form $P(\alpha) = a\alpha - q\alpha^2$. Table 20 supplies coefficients for all 45 curves on $[0, 1]$, computed from prediction and observation second moments. Figure 2a uses the gene-macro curve for COAD/DeepPT.

Table 20: All 45 amplitude curves. Each cell gives a/q for $P(\alpha) = a\alpha - q\alpha^2$, $0 \leq \alpha \leq 1$, in pp. Columns M/N/Z denote gene-macro, native-pooled, and source-standardized risk. Coefficients use target prediction and observation moments.

Model	M	N	Z
BRCA			
MORSA	15.2312/7.6438	20.4013/9.6232	10.2134/4.8173
SEQUOIA	13.2201/13.3608	17.8527/14.1783	8.8205/8.3119
HE2RNA	2.1711/0.4558	2.9382/0.5554	1.5113/0.2906
DeepPT	12.1431/55.2188	16.7484/21.5077	8.3888/19.5818
Path2Omics	14.4871/40.5211	19.3766/27.0481	10.1141/22.5384
COAD			
MORSA	8.2848/4.6020	9.4298/4.4546	5.7580/2.8049
SEQUOIA	8.0272/8.3511	8.7361/7.5705	5.4903/5.2217
HE2RNA	0.8667/1.9598	1.1927/0.3790	0.7053/0.4947
DeepPT	11.4651/51.9650	11.3224/15.6451	7.3415/19.2316
Path2Omics	14.3770/85.2666	14.0189/32.6109	8.9277/35.6332
GBM			
MORSA	11.7841/11.1890	12.0072/10.5237	7.3812/6.8466
SEQUOIA	8.4910/9.8885	8.6756/8.2267	5.4642/6.5633
HE2RNA	1.8913/1.2527	1.9927/1.1825	1.2634/0.8308
DeepPT	12.0700/30.4901	11.4978/17.3795	7.4451/19.7131
Path2Omics	13.8101/36.5538	13.0352/24.4178	8.4582/23.0679

H Precipitation Decomposition and Block Sensitivity

Table 21: Complete precipitation decomposition in mm. Four-week blocks are primary; eight-week blocks assess sensitivity. Each interval family covers five contrasts.

Contrast	Estimate	4-week interval	8-week interval
S	1.799	$[1.003, 2.595]$	$[1.019, 2.579]$
R	1.116	$[0.518, 1.715]$	$[0.607, 1.626]$
H	0.622	$[-0.141, 1.385]$	$[-0.012, 1.257]$
P	0.060	$[-0.427, 0.548]$	$[-0.436, 0.556]$
J	0.683	$[0.007, 1.359]$	$[0.037, 1.328]$

I Models and Training Settings

I.1 Inputs, folds, and source quantities

UNI provides 1,024-dimensional tile features. MORSA, SEQUOIA, and HE2RNA use 100 cluster representatives per slide. DeepPT and Path2Omics use the unclustered tile vectors without a tile-count cap; their slide predictions average tile-level outputs. Five patient-level source folds hold out

20% of patients for testing, with 10% of the remaining patients used for validation. Target-cohort predictions average the five fitted models. Expression means, principal components, and any learned transformations are fitted within each source training fold.

Archived source training pools for MORSA and SEQUOIA contain 855, 458, and 285 patients for BRCA, COAD, and GBM. The HE2RNA, DeepPT, and Path2Omics pools contain 629, 450, and 280 patients. Fixed source gene means and standard deviations use the latter populations. Common archived out-of-fold support used in the amplitude analysis contains 477, 386, and 77 patients; Table 18 gives the support of each fit. Each contribution compares a model's fixed predictions with its corresponding removal control.

I.2 Architectures

MORSA. The input projection maps each of the 100 slide representatives from 1,024 to 96 dimensions. A centered covariance matrix is normalized by its mean diagonal, shrunk toward the identity with coefficient 0.05, and regularized with $10^{-3}I$ before taking its matrix logarithm. The upper-triangular entries pass through a 1,024-unit hidden layer with dropout 0.1 to predict 16 expression coordinates. The training-fold expression mean and principal-component basis reconstruct gene expression from these coordinates.

SEQUOIA. Six linearized-attention layers with 16 attention heads operate on the 100 cluster representatives. A linear output head maps the aggregated representation to gene expression. The input representation is the same UNI feature type used by the other models.

HE2RNA. A tile-level $1024 \to 256 \to 256 \to G$ multilayer perceptron uses ReLU and dropout 0.5 after each hidden layer, where G is the training gene count. For each gene, top-k pooling averages the largest tile predictions. Training draws k uniformly from $\{1, 2, 5, 10, 20, 50, 100\}$; evaluation averages the seven pooled predictions and applies ReLU to the output. The final bias is initialized with the training gene mean.

DeepPT. An autoencoder maps $1024 \to 512 \to 1024$, with ReLU after the encoder and a linear decoder. The decoder accommodates signed UNI features. After reconstruction training, the encoder is fixed. A $512 \to 512 \to G$ regression head uses dropout 0.2 between its linear layers, without a hidden ReLU. Slide-level predictions average all tile predictions. The autoencoder reconstruction loss gives equal weight to slides.

Path2Omics. A $1024 \to 512 \to G$ head uses dropout 0.2 between linear layers, without a hidden ReLU. All tile-level predictions are averaged for each slide. The input layer accepts the 1,024-dimensional UNI features in place of the original feature representation.

I.3 Optimization and checkpoint selection

All gene-regression heads minimize MSE with equal patient and gene weights. Table 22 gives the optimization settings. Adam and AdamW use $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$; weight decay is zero, and no learning-rate schedule is used. MORSA and SEQUOIA select checkpoints by validation MSE/PCC: a lower MSE or an improvement in the tracked PCC updates the checkpoint. HE2RNA retains the best validation gene-macro PCC state. DeepPT and Path2Omics monitor the same PCC for stopping and retain the state at stopping. Patience counts epochs without the corresponding validation improvement. DeepPT's autoencoder first trains for 500 epochs on reconstruction MSE; the head settings in the table apply afterward.

Table 22: Source training settings. Batch size counts patients. Maximum epochs and patience refer to gene regression. All selection uses source validation patients.

Model	Optimizer	Learning rate	Batch	Max. epochs	Patience
MORSA	AdamW	10^{-3}	16	200	20
SEQUOIA	Adam	10^{-3}	16	200	20
HE2RNA	Adam	10^{-3}	16	200	20
DeepPT	Adam	10^{-4}	32	500	50
Path2Omics	Adam	10^{-4}	32	500	50