K-PRS: ZERO-TRAINING KERNEL REGRESSION WITH FROZEN FOUNDATION-MODEL EMBEDDINGS FOR COLD-START DRUG RESPONSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models for drug response prediction frequently fail to generalize to novel compounds under cold-start evaluation. We show that this failure stems from shortcut memorization: conventional architectures memorize historical compound identities during training, causing performance on unscreened molecules to collapse below naive cell-line baseline averages. To eliminate this vulnerability, we introduce a closed-form kernel estimator over frozen biological and chemical representations, gated by a positive-semidefinite prior that restricts evidence propagation strictly to compounds sharing identical mechanisms of action. By avoiding parameter optimization on response labels, the model is structurally prevented from memorizing entity identities. Across two independent cancer screening benchmarks, a matched five-fold protocol, and synthetic tasks with controlled identity availability, our non-parametric estimator consistently outperforms parameterized baselines. Furthermore, because predictions are formulated as closed-form weighted averages, the estimation weights inherently provide an exact evidence trail, calibrated uncertainty intervals, and verifiable robustness certificates from a single unified computation without auxiliary surrogates. All algebraic claims are formally machine-checked. Our findings establish both a practical predictor and a broader inductive principle: when entity identity is absent at deployment, explicitly refusing to learn it yields substantially stronger out-of-distribution generalization.

1 INTRODUCTION

Drug-response models are ultimately asked to support decisions about molecules that have not yet accumulated response labels. This *drug cold-start* setting is different from hiding random cell-drug pairs: the complete response column of a query drug is absent. A useful model must therefore transfer biological structure rather than recall a drug-specific effect. The distinction is operationally important in precision oncology, where screens such as GDSC provide broad but still sparse evidence over cell lines and compounds (Yang et al., 2013), and where a new candidate arrives long before a dense response profile exists. Recent evaluation work makes the same point from the benchmarking side: much of the apparent progress in this field survives neither a stricter split nor a naive baseline (Bernett et al., 2026).

Modern predictors encode cells and molecules and train a flexible fusion head on response labels. Such a head *can* exploit a shortcut: each observed drug embedding repeatedly co-occurs with one response profile, allowing the head to bind an embedding to a response identity. That shortcut helps random-pair interpolation but supplies no supervised anchor beyond the support of labelled drugs. In our matched leave-one-drug-out (LODO) study, a ridge head over the same frozen features reaches mean per-drug correlation 0.148, whereas simply reusing each cell's mean response reaches 0.644; a nested-tuned interaction fusion head reaches only 0.638 under the matched grouped protocol (Appendix I). We call this empirical pattern the *memorisation trap*. It is a mechanism hypothesis, not a claim that every optimiser or architecture must fail; the accompanying impossibility statement is weaker and sharper, and it applies to *every* estimator: outside the labelled support, two equally Lipschitz surfaces can agree on all labels and disagree at a query by twice the Lipschitz budget times

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.

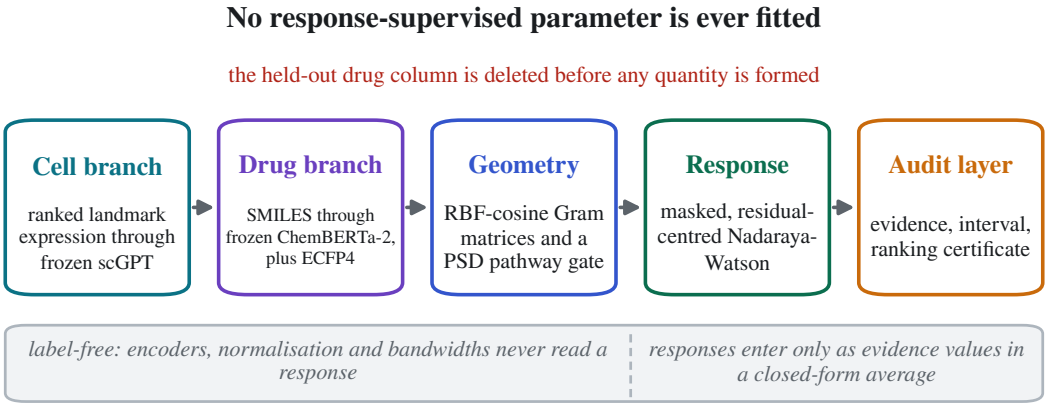


Figure 1: **The K-PRS pipeline and its information boundary.** Frozen encoders define cell and drug geometry; a PSD pathway kernel and a masked closed-form regression transfer responses without fitting a response head. The shaded band separates the label-free stages from the single point at which observed responses enter, namely as evidence values inside a closed-form average.

the distance to support, so some predictor must be wrong by at least that half-gap (Appendix H; machine-checked).

We take the shortcut away. K-PRS combines frozen scGPT cell embeddings (Cui et al., 2024), frozen ChemBERTa-2 and ECFP4 drug features (Ahmad et al., 2022), a pathway-aware positive-semidefinite kernel, and a masked Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964). No response-supervised parameter is optimised. A LODO query is answered by a weighted transfer from other drugs, with bandwidths selected from feature geometry. This deliberately classical estimator reaches 0.7085 mean per-drug Pearson on GDSCv2, and improves over the cell-mean baseline on an independently processed GDSC1 protocol.

The absence of a learned head also makes the output unusually auditable. The centred prediction and restored drug offset each have exact weighted-evidence decompositions; removing one observation has a closed-form effect. We add a LODO conformal interval and a Lipschitz certificate for whether a cell-embedding perturbation can reverse a drug ranking. Their assumptions, derivations, and the scope of the machine audit are stated explicitly rather than hidden behind empirical attack success.

Contributions.

(1) We isolate a cold-start failure mode with matched frozen-feature controls and prove a minimax lower bound: no estimator identifies the response surface outside the labelled support (Appendix H).
(2) We introduce a zero-training dyadic estimator with a PSD pathway prior, per-drug residualisation, label-free bandwidth selection, and vectorised inference.
(3) We evaluate strict LODO on two GDSC releases and a matched five-fold grouped protocol with nested tuning, against nonparametric and trained baselines, with ablations, novelty strata, support-based abstention, and a synthetic task.
(4) We provide exact evidence and leave-one-out decompositions, empirical conformal calibration at three granularities, and a ranking certificate we also prove tight.
(5) We release code, tests, sixteen source-generated figures, a standalone supplement, and a Mathlib-backed Lean development in which 38 theorems about the real-valued estimator are checked with no `sorry` and no custom axioms, including a counterexample to a natural but invalid PSD argument (Appendix S).

2 Related Work and Background

Drug-response prediction. Matrix-factorisation and dyadic methods interpolate response matrices; deep models instead fuse drug graphs or sequences with cell features (Liu et al., 2020; Nguyen et al., 2022; Jin & Nam, 2021; Shin et al., 2022; Xia et al., 2024), and foundation-model hybrids freeze or adapt molecular and transcriptomic encoders before training a response head (Ross et al., 2022; Rössner et al., 2025; Li et al., 2024). Published protocols and feature sets differ, so our central comparison is the controlled one: ridge, kernel-ridge and fusion heads trained on exactly the embeddings K-PRS uses. Appendix C expands this to pairwise kernel learning (Stock et al., 2018; Basilico & Hofmann, 2004; He et al., 2017), multi-task Gaussian processes (Bonilla et al., 2007), clinical transfer (Geeleher et al., 2014), combination screens (Holbeck et al., 2017), and extrapolation (Xu et al., 2021; Gulrajani & Lopez-Paz, 2021).

Kernel prediction for dyads. Nadaraya-Watson regression is a normalised local smoother (Nadaraya, 1964; Watson, 1964; Stone, 1977); product kernels extend it to paired objects (Vert & Kanehisa, 2002; Gärtner et al., 2002). Positive definiteness of the Gaussian kernel is Bochner's theorem (Bochner, 1933), closure under entrywise products is Schur's (Schur, 1911), median-distance bandwidths are a geometry-only heuristic (Garreau et al., 2017), and random features give the scalable variant we do not need at this size (Rahimi & Recht, 2007). We add frozen foundation representations, masking, a mechanism prior, and strict column holdout.

Explanations, uncertainty, and certificates. Post-hoc methods approximate local influence (Ribeiro et al., 2016; Lundberg & Lee, 2017; Baehrens et al., 2010; Sundararajan et al., 2017), attention weights need not be faithful (Jain & Wallace, 2019), and explanations can themselves be unstable (Alvarez-Melis & Jaakkola, 2018); kernel representer decompositions instead expose the coefficients that compute the output (Yeh et al., 2018). Split conformal prediction turns exchangeable residual ranks into marginal coverage (Vovk et al., 2005; Angelopoulos & Bates, 2023), with extensions under shift and dependence (Tibshirani et al., 2019; Chernozhukov et al., 2018), and Lipschitz bounds certify a region rather than searching it with an attack (Hein & Andriushchenko, 2017). Dyadic regression also admits eccentricity-driven bias that RMSE hides (Paz-Ruza et al., 2025), and counterfactual and regulatory framings ask for exactly the artifacts we emit (Wachter et al., 2017; European Commission, 2021). K-PRS places all of these views on one closed-form estimator.

3 Method

3.1 Cold-start task and frozen representations

Let $R \in \mathbb{R}^{n \times m}$ be a response matrix and $M \in \{0,1\}^{n \times m}$ its observation mask. LODO for drug j^* removes *all* labels $R_{\cdot j^*}$ before prediction; Figure 9 in Appendix B contrasts this with random-pair holdout and with the grouped protocol. Cell embeddings $z_i^c \in \mathbb{R}^{d_c}$ come from frozen scGPT applied to rank-normalised landmark-gene expression. Drug embeddings $z_j^d \in \mathbb{R}^{d_d}$ concatenate an ECFP4 fingerprint with a frozen ChemBERTa-2 representation and are normalised. Encoder inference is label-free with respect to R; detailed preprocessing is in Appendix E.

For unit-normalised embeddings, we use the RBF-cosine kernel

$$\kappa_\sigma(z, z') = \exp\left(-\frac{1 - z^\top z'}{2\sigma^2}\right) = \exp\left(-\frac{\|z - z'\|_2^2}{4\sigma^2}\right). \tag{1}$$

The equality makes this a Gaussian RBF restricted to the unit sphere, hence PSD by Bochner's theorem; the identity itself is machine-checked. Bandwidths are median pairwise feature distances computed without response labels.

3.2 Architecture and information flow

Figure 1 separates the system into four modules with a strict information boundary. The first three modules determine *where* evidence is retrieved; the observed response matrix supplies only *what* values are aggregated. This distinction is what we mean by zero training: pretrained encoders remain frozen, no response head is fitted, and neither feature normalisation nor bandwidth selection

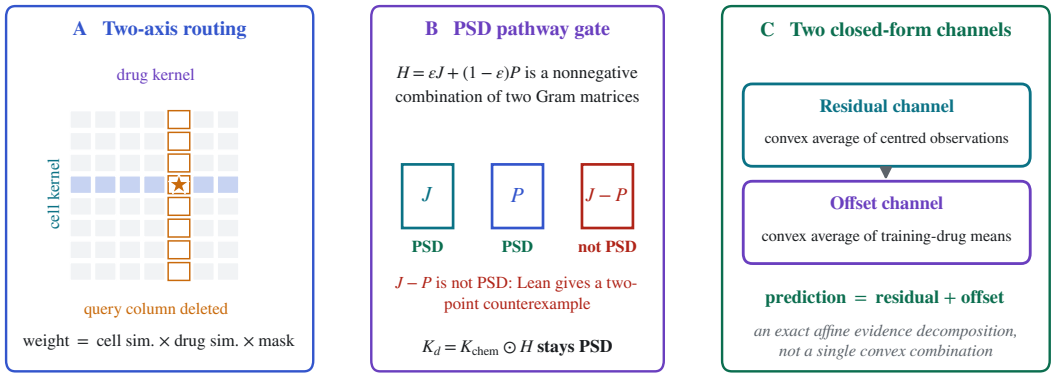


Figure 2: **Anatomy of the estimator. A** A query is routed along two axes at once, and the deleted column contributes to neither numerator nor denominator. **B** The pathway prior is a nonnegative combination of two Gram matrices, so the gated drug kernel stays PSD; the tempting $J - P$ form is not. **C** The residual and offset channels are separately convex, which makes the prediction an exact affine - not convex - decomposition of the evidence.

reads an IC50 label. Figure 2 shows the three mechanisms this buys, and Appendix D gives the module-by-module description, the offline caching plan, and the update cost of adding one drug.

3.3 A PSD PATHWAY-AWARE DRUG KERNEL

Let P_{jk} indicate whether two drugs share a target pathway, and let $J = \mathbf{1}\mathbf{1}^\top$. We require same-pathway weight one and different-pathway weight $\varepsilon \in [0, 1]$. The corresponding matrix is

$$H = \varepsilon J + (1 - \varepsilon)P, \qquad K_d = K_{\text{chem}} \odot H. \tag{2}$$

J and the block-indicator matrix P are PSD; their nonnegative combination is PSD, and Schur's product theorem makes K_d PSD. This form is important: $J - P$ is generally *not* PSD, and the proof assistant exhibits a two-point counterexample; the proof, and the explicit finite feature map that makes the Schur step constructive rather than appealed-to, are in Appendix F.

3.4 MASKED RESIDUAL NADARAYA-WATSON PREDICTION

For each training drug $k \neq j^*$, compute its observed mean μ_k and residuals $R^{\text{res}}_{ik} = (R_{ik} - \mu_k)M_{ik}$. Query weights are

$$w_{ik} = K_c(i^*, i)K_d(j^*, k)M_{ik}, \qquad \widehat{r}_{i^*j^*} = \frac{\sum_{i,k\neq j^*} w_{ik}R^{\text{res}}_{ik}}{\sum_{i,k\neq j^*} w_{ik}}. \tag{3}$$

The unseen drug offset is interpolated separately, $\widehat{\mu}_{j^*} = \sum_{k\neq j^*} q_k\mu_k / \sum_{k\neq j^*} q_k$, with $q_k = K_d(j^*, k)$, and $\widehat{R}_{i^*j^*} = \widehat{r}_{i^*j^*} + \widehat{\mu}_{j^*}$. This separation prevents the drug's unknown mean from leaking into evaluation.

The residual term is a convex combination of centred observations, and the offset is a convex combination of training-drug means. The final prediction is therefore an exact *affine* evidence decomposition; we do not claim it is a single convex combination of raw responses. For all queries at once,

$$\widehat{R}^{\text{res}} = \frac{K_c R^{\text{res}} K_d^\top}{K_c M K_d^\top} \quad \text{elementwise}, \tag{4}$$

with the held-out column excluded. Dense computation costs $\mathcal{O}(nm(n + m))$, seconds at this size; per-drug residualisation and the bandwidth heuristic are in Appendix G, timings and caching in Appendix P.

3.5 EVIDENCE, INTERVALS, AND RANKING CERTIFICATES

Normalised weights $\alpha_{ik} = w_{ik} / \sum w$ are the exact coefficients of $\widehat{r}$. If one residual is removed, the old-minus-new change is

$$\Delta_{ik} = \frac{\alpha_{ik}(R_{ik}^{\mathrm{res}} - \widehat{r})}{1 - \alpha_{ik}}. \tag{5}$$

We report both residual evidence and offset evidence so the explanation reconstructs the complete prediction.

For uncertainty, calibration drugs receive predictions made without their own columns, producing scores $|R_{ij} - \widehat{R}_{ij}|$. Their corrected empirical quantile q gives $[\widehat{R} - q, \widehat{R} + q]$. Standard finite-sample marginal coverage holds under exchangeability of the calibration and test nonconformity scores; because a pharmacological screen contains dependent cells within a drug, our main result is described conservatively as empirical calibration. We report three granularities - pairwise, within-profile and simultaneous - from the same predictions in Appendix N.

Finally, differentiating Equation 3 yields a factor $1/(2\sigma_c^2)$, not $1/\sigma_c^2$. We bound the gradient on an embedding ball to obtain $L_d(r)$. If drug d_1 is preferred to d_2 by margin $g(z) > 0$, the ranking is certified throughout radius r whenever

$$g(z) > r\{L_{d_1}(r) + L_{d_2}(r)\}. \tag{6}$$

The bound is tight: at equality the ranking can already be tied on the boundary of the ball, so no larger radius follows from Lipschitz information alone. The derivation, the direction convention for log IC50, and the PGD comparison are in Appendix O.

4 EXPERIMENTS

4.1 PROTOCOL

The primary dataset is GDSCv2 with 941 cell lines having matched RMA expression and 295 drugs. Every method receives the same LODO split and frozen features. We report Pearson correlation across cells for each held-out drug, then average across drugs; uncertainty is a bootstrap interval over drug tasks. We also report RMSE, rank correlation and the standard deviation across drugs; a hybrid 956-cell construction is a stability check, and GDSC1 supplies an independent 958×367 evaluation. Data inclusion, preprocessing, baseline hyperparameters, statistical tests, and provenance are in Appendix E-I.

4.2 PRIMARY COMPARISON

K-PRS reaches mean per-drug Pearson 0.7085 with a 95% bootstrap interval $[0.6945, 0.7218]$, median 0.7230, per-drug standard deviation 0.1193, and RMSE 2.294 log-IC50 units. The cell-mean, kernel-ridge, and drug-only k-NN baselines reach 0.6442, 0.6383, and 0.5879. The trained ridge head reaches 0.1480 on the same frozen inputs. Under the matched five-fold grouped protocol with nested tuning, K-PRS reaches 0.6991 $[0.6846, 0.7126]$ against 0.6620 for a two-step kernel-ridge model, 0.6437 for the cell mean - the same baseline on these folds - and 0.6379 for an interaction fusion head, which loses to the naive prior on 226 of 295 drugs after honest model selection (paired Wilcoxon $p = 1.8 \times 10^{-20}$; Appendix I reports every paired test and the cross-protocol FUSED reference).

4.3 WHERE THE GAIN COMES FROM

With no pathway prior, chemistry-only smoothing reaches 0.6561. A pathway-only kernel reaches 0.6926; adding ChemBERTa-2 or ECFP4 reaches 0.6944 or 0.7049; the dual representation with multiplicative pathway weighting reaches 0.7085. Thus most signal comes from the cell response profile and mechanism restriction, while local chemistry supplies the final increment. On GDSC1, K-PRS reaches 0.5379 $[0.5257, 0.5506]$ versus 0.4863 $[0.4735, 0.5002]$ for the cell mean. The 956-cell hybrid result, 0.7075, shows that the headline is not sensitive to the exact expression overlap. Bandwidth grids, missing-annotation stress tests, the conservative unknown-pathway fallback, and dyadic bias diagnostics are in Appendix J.

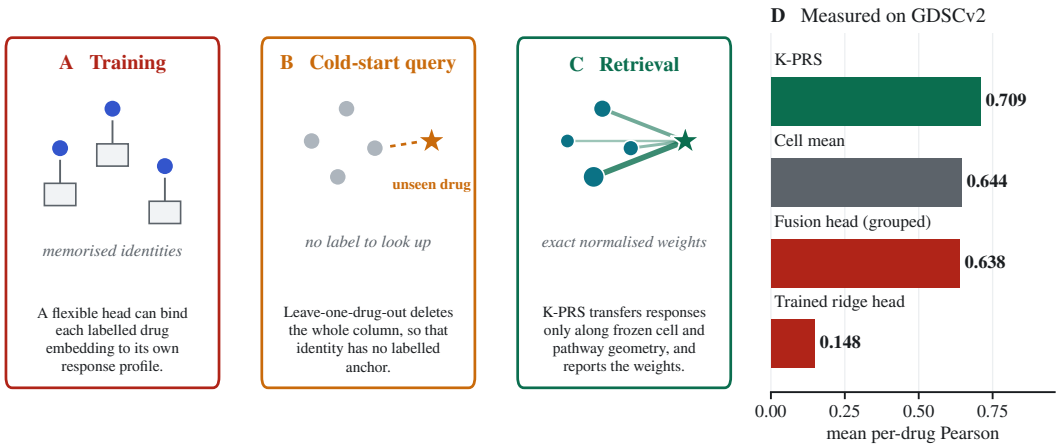


Figure 3: **From shortcut to retrieval. A-C** A supervised head may associate labelled embeddings with response identities; a LODO query removes that identity, and K-PRS instead exposes the weighted evidence transferred to the unseen drug. **D** On matched frozen inputs neither trained head passes the cell-mean prior, while the closed-form estimator exceeds it; the fusion bar is the grouped-protocol control.

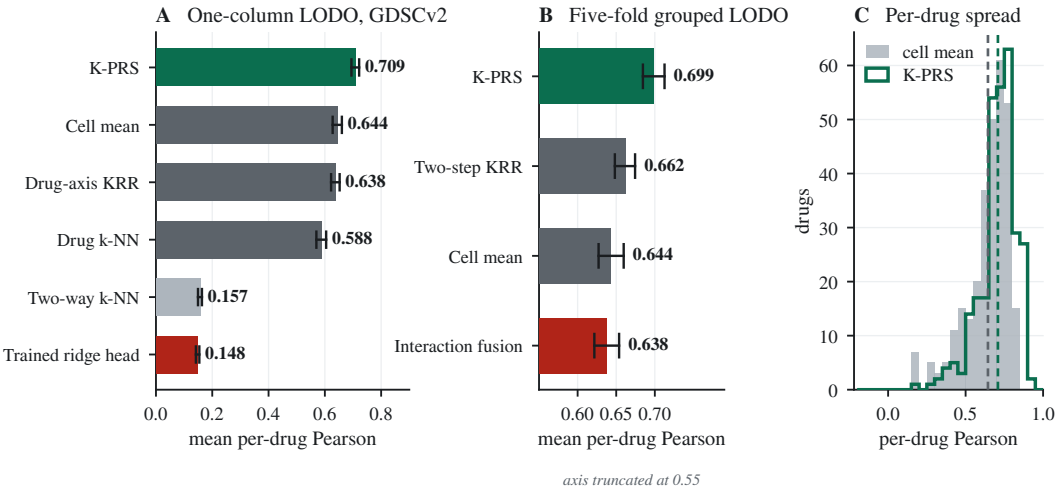


Figure 4: **Strict and grouped GDSCv2 LODO. A** One-column holdout on the full scale, so the collapse of the trained ridge head is visible rather than cropped away. **B** Five-fold grouped LODO with nested tuning; the axis is truncated to resolve the differences and is labelled as such. **C** Per-drug Pearson distributions, dashed lines at the two means. Error bars are bootstrap intervals over drug tasks.

Novelty-stratified curves on a separate 690-cell probe show no catastrophic collapse away from nearby drugs, but nearest-neighbour fingerprint distance proves a weak proxy for difficulty, so we read them descriptively (Appendix K). Support diagnostics turn the idea into an actionable rule, and which one is used decides whether it works: of three label-free scores on the gated drug kernel, only nearest neighbour is usable. Answering the 50% of drugs with the closest chemical neighbour raises mean Pearson from 0.7085 to 0.7354, whereas kernel mass gains little and effective neighbour count moves the wrong way (Appendix J; Appendix L covers the synthetic generator).

4.4 Auditability results

Equation 5 reconstructs explicit removal effects with Spearman 1.000. The pairwise LODO conformal interval reaches 91.84% empirical coverage at the 90% nominal level, with mean width 7.46

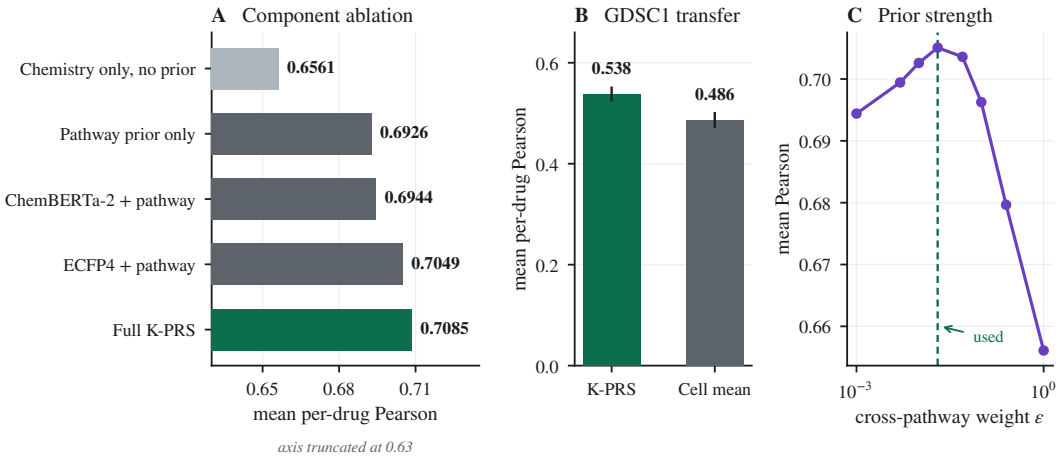


Figure 5: **Ablations and transfer.** **A** The pathway prior supplies the largest increment over chemistry-only smoothing; dual chemical features add a smaller gain. **B** On independently processed GDSC1, K-PRS again exceeds the cell-mean baseline. **C** Accuracy as a function of the cross-pathway weight, which is never selected using response labels.

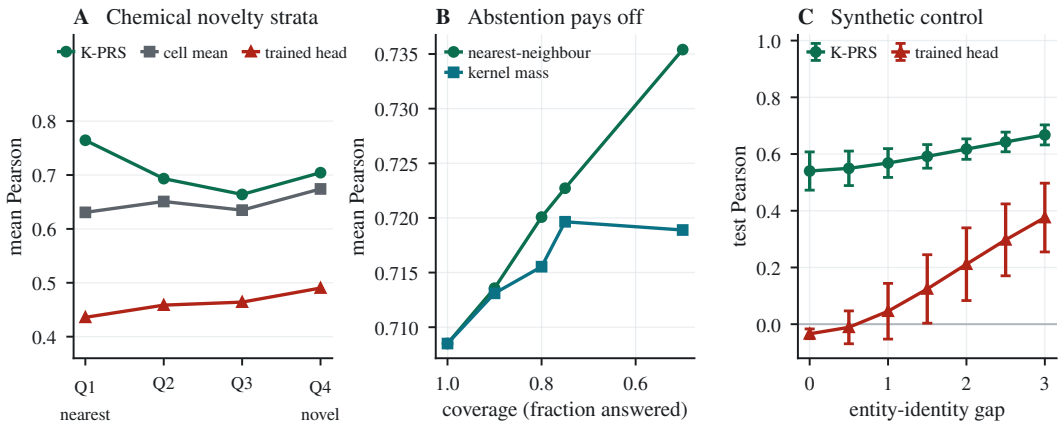


Figure 6: **Where the estimator works.** **A** Novelty strata from the separate 690-cell probe of Appendix K; not comparable with panel B. **B** Abstaining on low-support queries raises accuracy monotonically for the nearest-neighbour diagnostic and only weakly for kernel mass. **C** A five-seed synthetic sweep; the separation between the curves is the claim, not the trend along the gap axis (Appendix L). Bars are standard deviations over seeds.

log-IC50 units; the width makes clear that calibration is not the same as sharpness. Treating whole drug columns as the exchangeable unit costs sharpness for a stronger event: 90.54% simultaneous coverage of every observed cell at width 14.29, or 93.24% within-profile coverage at width 9.48. Across one hundred warm-start rankings the median certified embedding radius is only 6.5×10^{-9} and none reaches the preregistered threshold of 0.01; PGD finds no flip, but attack failure is not a proof. The same estimator on the same rankings, certified instead against perturbation of the *response evidence*, gives a median radius of 0.085 log-IC50 units, with 46% above 0.10 - a more informative statement about what the retrieval depends on. Both certificates concern the mathematical estimator, not biological efficacy; score definitions, radius distributions and settings are in Appendix M-O.

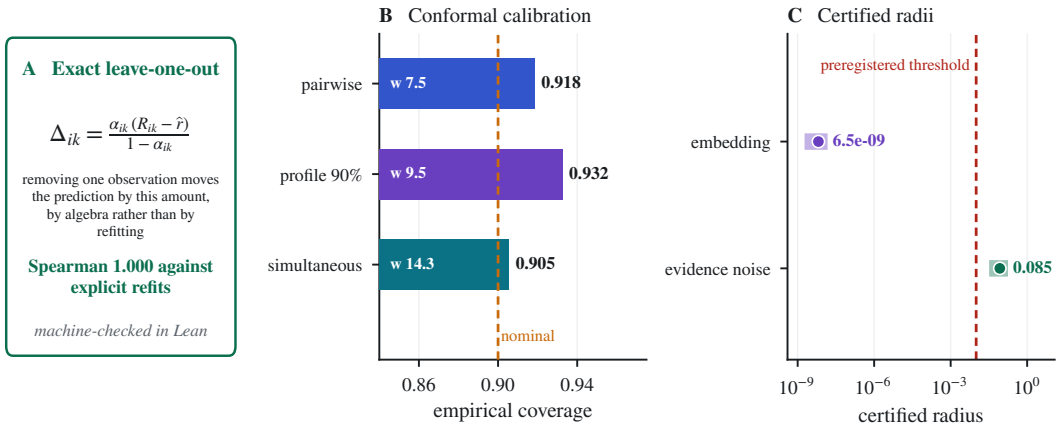


Figure 7: **Auditable outputs, all from the same weights.** One run, one split, one set of rankings. **A** The leave-one-out effect is an identity, not an approximation. **B** Empirical coverage at three granularities against a 90% nominal target, annotated with mean interval width so that calibration is never reported without sharpness. **C** Certified radii on a logarithmic axis, with interquartile bands. The embedding-space radii are minuscule; the radii against noise in the response evidence are not.

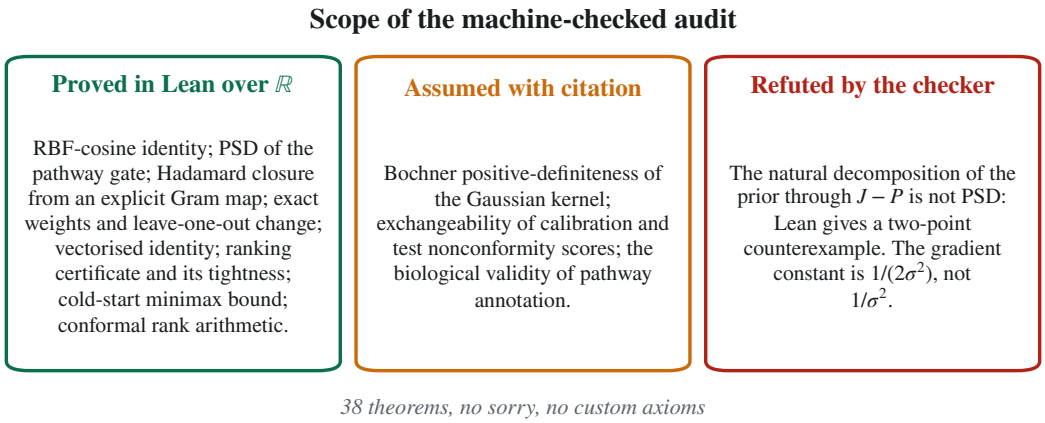


Figure 8: **What is proved and what is assumed.** The finite and inequality content of the method is checked in Lean over the reals with Mathlib. Analytic and probabilistic facts are used with their standard hypotheses and citations and are deliberately not asserted by the development. The checker also refutes the natural $J - P$ decomposition and fixes the gradient constant at $1/(2\sigma^2)$.

4.5 Machine-checked mathematics

We machine-check 38 theorems covering Equations 1-6: the RBF-cosine identity, PSD-ness of the pathway gate through an explicit finite feature map and its closure under the entrywise product, exactness of the normalised weights and of Equation 5, the vectorised identity, the ranking certificate with a tightness witness, the cold-start minimax bound, and the conformal rank arithmetic - with no `sorry` and no custom axioms. Appendix S lists every statement, and Figure 8 the boundary between what is proved and what is assumed with citation.

5 Limitations

Appendix U gives the complete statement; the four themes are these. **Support and representation.** K-PRS is an interpolator, not a mechanism simulator: a query with little kernel mass or an uninformative pathway assignment receives diffuse evidence and drifts toward smoothed training

means, so the method should abstain rather than score everything with equal confidence. Its geometry inherits its inputs' blind spots - scGPT is applied to bulk rather than single-cell expression, fingerprints and language embeddings may miss stereochemistry, exposure or target engagement, and mutation, copy number, dose and microenvironment are absent. **Evaluation and comparison.** Strict LODO measures transfer inside two historical screens, not prospective medicinal chemistry, and GDSC1 and GDSCv2 are not independent clinical cohorts. Per-drug Pearson rewards ranking cells within a drug and can coexist with sizable absolute error; we did not rerun every published architecture, so the cross-protocol reference stays contextual. **Auditability is not causality.** Exact weights explain arithmetic, not biology; a large coefficient exposes dependence, not a causal pathway. Pairwise conformal coverage is empirical under within-drug dependence, the cleaner group analysis pays with very wide intervals, and the certified embedding radii, while valid, are minuscule. The machine audit covers algebra and inequalities, not Bochner's theorem, exchangeability, or the biology of pathway curation. **External validity.** All conclusions are preclinical, the HCC827 example is a warm-start illustration only, and the decisive next test is prospective: freeze the evidence base, register a support-based abstention rule, then assay new compounds.

6 Conclusion

Cold-start evaluation changes what counts as useful learning. When a complete query drug column is removed, success cannot come from remembering that drug's historical response identity. In our matched experiments, response-supervised ridge and fusion heads over strong frozen representations underperform even a cell-mean prior, whereas pathway-aware retrieval improves strict LODO accuracy on both GDSCv2 and GDSC1 and under a grouped protocol with nested tuning. The result does not establish that trained models are intrinsically unsuitable; it shows that random-pair performance and architectural sophistication are insufficient evidence for inductive drug transfer.

The architecture of K-PRS makes its inductive bias explicit. Frozen encoders define cell and drug geometry, a PSD pathway gate restricts mechanistic transfer, and two closed-form channels reconstruct the cell-specific residual and drug-wide offset. The same normalised weights then support exact evidence attribution, conformal calibration, and a differentiable ranking certificate, so the audit artifacts describe the predictor that produced the score rather than post-hoc surrogates. They also expose where that transparency is uncomfortable - intervals are wide and certified embedding radii tiny - which beats concealing weak assurance behind a point estimate.

The design principle is modest but portable. If an entity will be unseen at deployment, evaluation should remove all of its labels, baselines should include simple priors, and every learned component should demonstrate that it transfers rather than identifies. A transparent smoother over well-chosen frozen geometry is a strong control against which a trained head must earn its complexity (Appendix T).

7 Impact Statement

Appendix V-W give the full statement. **Intended benefit.** The appropriate use of K-PRS is to prioritise laboratory measurements for an unscreened compound: identify responsive cell models, expose the observations behind the ranking, and direct assay capacity toward informative follow-up (Garnett et al., 2010). **Risks and affected groups.** The central harm is automation bias: a precise-looking score, explanation, or certificate may be mistaken for treatment evidence. Cell lines incompletely represent tumour heterogeneity, ancestry, immune context, prior therapy and toxicity (World Health Organization, 2026; Lynch et al., 2004; Chapman et al., 2011), so errors may concentrate in rare subtypes, novel scaffolds or underrepresented tissues, and our eccentricity audit finds that concentration for every method tested, ours included (Appendix J). Retrieval also reproduces biases in the screened-drug library and in pathway curation. **Safeguards and prohibited use.** Outputs should display support diagnostics, the evidence decomposition, interval width as well as coverage, and a cold-start or warm-start label; low-support queries should trigger abstention. No prediction should prescribe therapy, exclude a patient, or bypass prospective assays, dose-response validation, toxicity assessment, or qualified human review. Any move to patient-derived or identifiable data requires consent, access control, subgroup audits and regulatory review (European Commission, 2021); redistribution must respect GDSC and encoder checkpoint terms.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022. doi: 10.48550/arXiv.2209.01712.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018. doi: 10.48550/arXiv.1806.08049.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions: An approximation method. *Journal of Machine Learning Research*, 11(61):1803–1831, 2010. URL https://jmlr.org/papers/v11/baehrens10a.html.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. doi: 10.1038/nature11003.

Justin Basilico and Thomas Hofmann. Unifying collaborative and content-based filtering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 65–72, 2004. doi: 10.1145/1015330.1015394.

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. doi: 10.1038/s41467-026-72903-w.

Salomon Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410, 1933. doi: 10.1007/BF01452844.

Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html.

Paul B Chapman, Axel Hauschild, Caroline Robert, John B Haanen, Paolo Ascierto, James Larkin, Reinhard Dummer, Claus Garbe, Alessandro Testori, Michele Maio, et al. Improved survival with vemurafenib in melanoma with braf v600e mutation. *New England Journal of Medicine*, 364(26): 2507–2516, 2011. doi: 10.1056/NEJMoa1103782.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 732–749, 2018. URL https://proceedings.mlr.press/v75/chernozhukov18a.html.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

European Commission. EU artificial intelligence act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Roman Garnett, Michael A. Osborne, and Stephen J. Roberts. Bayesian optimization for sensor set selection. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 209–219, 2010. doi: 10.1145/1791212.1791238.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017. doi: 10.48550/arXiv.1707.07269.

Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alex J Smola. Multi-instance kernels. In *ICML*, pp. 179–186, 2002. URL https://dl.acm.org/doi/10.5555/645531.656025.

Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3): R47, 2014. doi: 10.1186/gb-2014-15-3-r47.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Tong He, Marc Heidemeyer, Frank Ban, Artem Cherkasov, and Martin Ester. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9:24, 2017. doi: 10.1186/s13321-017-0209-z.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/e077e1a544eec4f0307cf5c3c721d944-Abstract.html.

Susan L Holbeck, Richard Camalier, James A Crowell, Jeevan P Govindharajulu, Melinda Hollingshead, Lawrence W Anderson, Eric Polley, Lawrence Rubinstein, Anubha Srivastava, Deborah Wilsker, et al. The national cancer institute almanac: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Research*, 77(13): 3564–3576, 2017. doi: 10.1158/0008-5472.CAN-17-0489.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Iljung Jin and Hojung Nam. HiDRA: Hierarchical network for drug response prediction with attention. *Journal of Chemical Information and Modeling*, 61(8):3858–3867, 2021. doi: 10.1021/acs.jcim.1c00706.

Kun Li, Weiwei Liu, Yong Luo, Xiantao Cai, Jia Wu, and Wenbin Hu. Zero-shot learning for preclinical drug screening. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2117–2125, 2024. URL https://www.ijcai.org/proceedings/2024/234.

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 2020. doi: 10.1093/bioinformatics/btaa822.

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

Thomas J Lynch, Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A Okimoto, Brian W Brannigan, Patricia L Harris, Sara M Haserlat, Jeffrey G Supko, Frank G Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004. doi: 10.1056/NEJMoa040938.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964. doi: 10.1137/1109020.

Tuan Nguyen, Giang T. T. Nguyen, Thin Nguyen, and Duc-Hau Le. Graph convolutional networks for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):146–154, 2022. doi: 10.1109/TCBB.2021.3060430.

Adrián Paz-Ruza, Andrés F. Vilas, Francisco Cedrón, Xavier Carrió, Antonio Frattini, Héctor Fernández-Ledesma, Luis Borrajo, Fernando Saavedra, Nuria Matos-Carro, Teresa J. Pintor, Óscar Fontenla-Romero, Bertha Guijarro-Berdiñas, Amparo Alonso-Betanzos, Glaucia Maria Machado-Santelli, Heydi Mendez-Vazquez, and Enrique Fernández-Blanco. Beyond RMSE and MAE: Introducing EAUC to unmask hidden bias and unfairness in dyadic regression models. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19619–19630, 2025. doi: 10.1109/TNNLS.2025.3593059.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.18653/v1/N16-3020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Till Rössner, Jonas Balke, and Ming Tang. FUSED: Cross-domain integration of foundation models for cancer drug response prediction. *bioRxiv*, 2025. doi: 10.1101/2025.09.30.679434.

Issai Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1911(140):1–28, 1911. doi: 10.1515/crll.1911.140.1.

Yunseok Shin, Yinhua Kim, and Sun Kim. DRPreter: Interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. *International Journal of Molecular Sciences*, 23(22):13919, 2022. doi: 10.3390/ijms232213919.

Michiel Stock, Tapio Pahikkala, Antti Airola, Bernard De Baets, and Willem Waegeman. A comparative study of pairwise learning methods based on kernel ridge regression. *Neural Computation*, 30(8):2245–2283, 2018. doi: 10.1162/neco_a_01096.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–645, 1977. doi: 10.1214/aos/1176343886.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Jean-Philippe Vert and Minoru Kanehisa. Graph-driven feature extraction from microarray data using diffusion kernels and kernel CCA. In *Advances in Neural Information Processing Systems*, volume 15, 2002. URL https://proceedings.neurips.cc/paper/2002/hash/074177d3eb6371e32c16c55a3b8f706b-Abstract.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2017. doi: 10.2139/ssrn.3063289.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964. URL https://www.jstor.org/stable/25049340.

World Health Organization. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer, 2026. URL https://www.who.int/news-room/fact-sheets/cancer. Fact sheet; accessed 29 August 2026.

Xiaoqiong Xia, Chaoyu Zhu, Fan Zhong, and Lei Liu. TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. *BMC Biology*, 22(1):227, 2024. doi: 10.1186/s12915-024-02023-8.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=UH-cmocLJC.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013. doi: 10.1093/nar/gks1111.

Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/8a7129b8f3edd95b7d969dfc2c8e9d9d-Abstract.html.

A Appendix Overview

The appendices give the full methodological, theoretical, experimental and reproducibility detail behind the main text, and are also compiled as a standalone supplement. Table 1 maps each topic to its location. Every headline result uses the 941-cell GDSCv2 protocol; the few analyses run on a smaller 690-cell intersection are labelled as such where they appear (Appendix K and Appendix Q).

Table 1: Where each topic is treated.

Topic	Location
Full related-work comparison	Section C
Module-by-module architecture and information flow	Section D
Detailed formulation, encoders, kernels, and NW derivation	Main Method section and Sections F-G
Exact representers and counterfactual derivation	Sections M and O
Main, baseline, ablation, GDSC1, and variance tables	Sections I and J
Cross-platform, runtime, and warm-start case study	Sections Q, P, and R
Non-identifiability, minimax bound, and all proofs	Sections H, F, and S
Bandwidth, novelty, synthetic, and per-class analyses	Sections J, K, L, and Q
Full Limitations text	Section U
Full Impact Statement text	Sections V and W
Reproducibility details and code inventory	Section X

B Evaluation Protocol and Leakage Controls

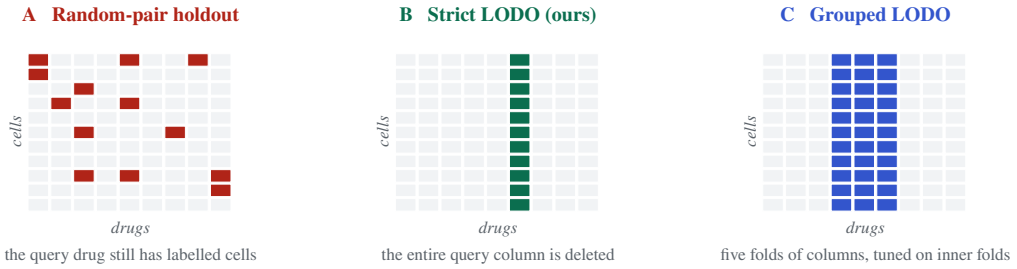


Figure 9: **Three evaluation protocols on the same response matrix.** **A** Random-pair holdout leaves the query drug with labelled cells. **B** Strict LODO, used throughout, deletes the entire query column before any statistic is computed. **C** Grouped LODO deletes five folds of columns and tunes on inner folds. The HCC827 illustration of Appendix R is warm-start and follows none of these.

For drug j, the evaluation routine constructs the Boolean selector $\mathcal{D}_{-j} = \{k : k \neq j\}$ before computing drug means, residuals, denominators, or offsets. The query embedding may be used for distances-the task is inductive in labels, not transductive in features-but $R_{\cdot j}$ and every statistic derived from that column remain inaccessible. The drug bandwidth is computed from training drugs only when not supplied from a label-free global feature computation. The cell kernel may include the query cell itself because the task holds out a *drug*, not a cell; its responses to other drugs are legitimate training evidence.

This protocol differs from three easier settings: random observed-pair holdout, warm-start matrix completion, and leave-one-cell-out. We never mix their metrics. Per-drug Pearson correlation is computed over the observed cells for that held-out drug. The reported mean weights every drug task equally rather than allowing densely measured drugs to dominate. The bootstrap resamples drug tasks with replacement.

C Extended Related Work

Dyadic and pairwise prediction. Cell-drug response is a dyadic regression problem, and the classical treatment is pairwise kernel ridge regression, which forms a Kronecker kernel over the two

entity spaces and solves a single linear system (Stock et al., 2018). That literature already distinguishes four evaluation regimes - both entities seen, one entity new, the other entity new, and both new - and observes that difficulty increases sharply in the last three. Our strict LODO protocol is the "new drug" regime. Related constructions include unified collaborative and content-based filtering (Basilico & Hofmann, 2004), gradient-boosted read-across for drug-target affinity (He et al., 2017), and multi-task Gaussian processes, which share strength across tasks through a learned task covariance (Bonilla et al., 2007). Our pathway gate can be read as a fixed, mechanistically motivated, non-learned task covariance - deliberately non-learned, because a task covariance fitted on observed drugs is exactly the object that cannot be estimated for an unobserved one.

Deep drug-response models. DeepCDR couples a graph convolutional drug encoder with multi-omic cell features (Liu et al., 2020); GraphDRP uses graph representations of compounds (Nguyen et al., 2022); HiDRA introduces hierarchical pathway attention (Jin & Nam, 2021); DRPreter adds knowledge-guided graph structure with a transformer readout (Shin et al., 2022); TransCDR transfers multimodal representations (Xia et al., 2024). Foundation-model hybrids freeze or adapt large molecular and transcriptomic encoders and then train a response head (Ross et al., 2022; Rössner et al., 2025). Zero-shot preclinical screening addresses the unscreened-compound regime explicitly (Li et al., 2024). These systems differ in data version, cell inclusion, feature construction and split definition, which is why we treat published numbers as context and rely on controls trained on our own frozen features. The DrEval analysis makes the same methodological argument at scale and finds that naive baselines are frequently competitive once splits are tightened (Bernett et al., 2026); our cell-mean result is an instance of that finding, and our contribution is a method that beats the naive baseline under the strict split rather than only under the permissive one.

Extrapolation and domain shift. The failure we document is a specific instance of a general phenomenon: neural networks extrapolate in ways determined by their architecture and training distribution rather than by the target function (Xu et al., 2021), and careful benchmarking of domain generalisation repeatedly finds that simple baselines are hard to beat once model selection is made honest (Gulrajani & Lopez-Paz, 2021). Our grouped protocol with nested tuning is designed to satisfy that standard.

Clinical and combinatorial extensions. Transfer from cell-line screens to clinical response has an established, and appropriately cautious, literature (Geeleher et al., 2014), and combination screens such as the NCI ALMANAC extend the dyadic problem to triples (Holbeck et al., 2017). We do not attempt either; both are named in Appendix U as future work rather than as claims.

D Architecture and Information Flow in Detail

This section expands Section 3.2.

Representation and geometry. The cell branch maps ranked landmark-gene expression through scGPT to $Z_c \in \mathbb{R}^{n \times d_c}$. In parallel, the drug branch concatenates a ChemBERTa-2 SMILES embedding with an ECFP4 fingerprint to form $Z_d \in \mathbb{R}^{m \times d_d}$. Row normalisation places both branches on unit spheres. The geometry module then constructs a reusable cell Gram matrix K_c and, for each query drug, a chemical similarity vector to the remaining drugs. A pathway gate modulates the latter without changing the encoders or introducing fitted parameters.

Cold-start retrieval and response synthesis. At inference for j^*, the entire column $R_{\cdot j^*}$ is removed before drug means, residuals, denominators, or offsets are formed. The product $K_c(i^*, i)K_d(j^*, k)$ routes a query to observed cell-drug pairs along two axes: biologically similar cells and chemically or mechanistically related drugs. The response module has two parallel channels. One estimates the cell-specific centred effect from pairwise evidence; the other estimates the unseen drug's global offset from drug neighbours. Adding the channels yields the prediction in Equation 3. Missing observations are eliminated by M, so they contribute to neither numerator nor normalisation, a property that is itself machine-checked.

Shared audit layer. The normalised retrieval weights are retained rather than discarded. The evidence view reports which observations and drug offsets formed the answer; the uncertainty view

calibrates errors from complete held-out columns; and the robustness view differentiates the same closed-form map. Thus prediction, explanation, interval, and certificate do not come from separate surrogate models (Figure 10).

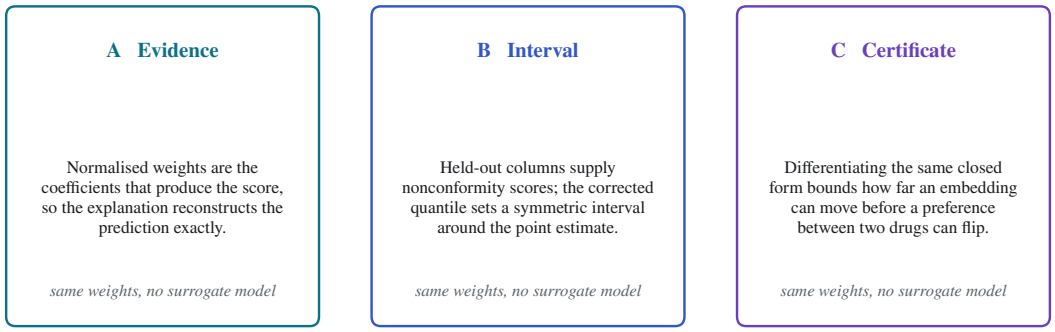


Figure 10: Three audit outputs from one set of weights. **A** The normalised weights are the coefficients that produce the score, so the explanation reconstructs the prediction exactly. **B** Held-out columns supply the nonconformity scores that set the interval. **C** Differentiating the same closed form bounds how far an embedding can move before a preference between two drugs can flip.

The offline stage caches embeddings and K_c; a new drug requires only its two chemical representations, pathway metadata when available, and a query-to-training kernel vector. Vectorised execution and update costs are in Appendix G and P.

E Data, Features, and Provenance

GDSCv2. The response matrix is built from the Sanger fitted dose-response release dated 27 October 2023, with log IC50 as the response. Duplicate cell-drug observations are aggregated by the data loader, and missing pairs remain masked rather than imputed. The headline intersection contains 941 cell lines with RMA expression and 295 drugs. A separate hybrid construction contains 956 lines: 941 RMA lines plus 15 DepMap-only lines, aligned on 892 shared landmark genes and rank-normalized within each cell.

GDSC1. The independent evaluation uses the fitted dose-response release dated 24 July 2022, RMA expression, and the drug mapping assembled by the released build script. The assembled response matrix is 987×367; after intersecting with available expression it contains 958 cells and 367 drugs, which is the matrix evaluated. Its ECFP4 representation, pathway prior, and LODO split are built independently from GDSCv2. Shared drug names do not transfer response labels between screens. Two settings differ from the GDSCv2 headline and are not tuned to match it: the drug branch is the 2048-bit ECFP4 fingerprint alone, because the GDSC1 SMILES set is resolved independently and we did not re-encode it with ChemBERTa-2, and the cross-pathway weight is $\varepsilon = 0.05$ rather than 0.02. GDSC1 is therefore a transfer check on a deliberately simpler configuration, which makes its lower absolute accuracy (0.5379 against a 0.4863 cell-mean prior) the relevant comparison rather than its distance from the GDSCv2 number.

Cell representation. Expression is restricted to available L1000 landmark genes (Subramanian et al., 2017), ranked within each cell, and passed through the frozen whole-human scGPT checkpoint (Cui et al., 2024). The embedding scripts expose the pooling style and checkpoint arguments. The primary run uses cached embeddings; no response label is used to alter encoder weights.

Drug representation. Canonical SMILES are encoded by a frozen ChemBERTa checkpoint and concatenated with a 2048-bit radius-two Morgan fingerprint (ECFP4). Missing or unresolved structures are tracked by the build scripts. Rows are normalized before cosine geometry. Target pathways come from the GDSC screened-compounds metadata. The multiplicative prior uses $\varepsilon = 0.02$ in the headline dual-feature variant; this was a modeling choice, not a label-tuned optimum. A metadata-only alternative sets ε to the median ratio of same-pathway to other neighbors, giving 0.042553. The

conservative fallback treats missing, "Other," and "Unclassified" categories as diagonal-only rather than a shared mechanism block.

Public assets and redistribution. GDSC and CCLE are public research resources (Yang et al., 2013; Barretina et al., 2012). Large raw files and third-party checkpoints are not duplicated in the anonymous code archive. Their expected filenames, upstream locations, processing commands, and hashes when locally available are documented. Users remain responsible for upstream license and access terms.

F Proofs of Kernel and Evidence Properties

F.1 RBF-cosine kernel

Proposition 1 (PSD RBF-cosine kernel)**.** *For unit vectors z, z', the RBF-cosine equation in the main paper defines a positive-semidefinite kernel for every $\sigma > 0$.*

Proof. Unit normalization gives $\|z-z'\|_2^2 = 2-2z^\top z' = 2(1-z^\top z')$. Thus $\kappa_\sigma(z,z') = \exp[-\|z-z'\|_2^2/(4\sigma^2)]$, the Gaussian RBF with bandwidth $\sqrt{2}\sigma$. It is PSD by Bochner's theorem (Bochner, 1933); restriction from $\mathbb{R}^d$ to the unit sphere preserves PSD. □

F.2 Corrected pathway prior

Proposition 2 (PSD pathway weighting)**.** *Let $P_{jk} = 1$ when drugs j, k share a pathway and zero otherwise. For $\varepsilon \in [0,1]$, $H = \varepsilon J + (1-\varepsilon)P$ is PSD and has entries one within a pathway and ε across pathways. Consequently $K_{\text{chem}} \odot H$ is PSD whenever K_{chem} is PSD.*

Proof. $J = \mathbf{1}\mathbf{1}^\top$ is rank-one PSD. After ordering drugs by pathway, P is block diagonal with all-ones blocks, hence PSD. A nonnegative linear combination of PSD matrices is PSD. The entry calculation follows directly: $\varepsilon + (1-\varepsilon) = 1$ within a block and ε otherwise. The last claim is the Schur product theorem (Schur, 1911). □

A constructive Schur step. The proof above invokes the Schur product theorem. The machine-checked development avoids that appeal by proving the special case the method needs and exhibiting the required factorisation. Concretely, if $A_{jk} = \sum_f u_f(j)u_f(k)$ for a finite family $\{u_f\}$ and B is PSD, then for any x,

$$\sum_{j,k} x_j(A_{jk}B_{jk})x_k = \sum_f \sum_{j,k} \big(u_f(j)x_j\big)B_{jk}\big(u_f(k)x_k\big) \;\geq\; 0,$$

each summand being a quadratic form of B. The prior H has the explicit map

$$u_\bullet(j) = \big(\sqrt{\varepsilon},\ \sqrt{1-\varepsilon}\,\mathbf{1}[p(j) = g]\big)_{g\in\mathcal{P}},$$

where $p(j)$ is the pathway of drug j and $\mathcal{P}$ the pathway set, since $\sqrt{\varepsilon}^2 + (1-\varepsilon)\mathbf{1}[p(j)=p(k)] = H_{jk}$. Hence $K_{\text{chem}} \odot H$ is PSD whenever K_{chem} is, and the only remaining analytic input is Bochner's theorem for K_{chem} itself. See `PSD.hadamard_of_gram` and `gram_pathwayPrior` in Appendix S.

A tempting decomposition that fails. The same entries can be written $P + \varepsilon(1-P)$, which suggests arguing that both summands are PSD. The entry formula is correct but the argument is not, because $1-P$ is generally indefinite: for two singleton pathways, $1-P = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ and $(1,-1)(1-P)(1,-1)^\top = -2$. The form $\varepsilon J + (1-\varepsilon)P$ is the one that yields a valid proof.

F.3 Product kernel

Let A and B be cell and drug Gram matrices over a common list of observed pairs. Their joint Gram matrix is $A \odot B$. Writing $A = U^\top U$, for any x,

$$x^\top(A \odot B)x = \sum_r y_r^\top B y_r \geq 0, \qquad (y_r)_i = x_i U_{ri}.$$

Thus the dyadic product kernel is PSD.

F.4 Residual and offset evidence

Proposition 3 (Two exact convex decompositions)**.** *If the denominator in the main-text NW estimator is positive, then $\widehat{r} = \sum_{ik} \alpha_{ik} R^{\mathrm{res}}_{ik}$ with $\alpha_{ik} \geq 0$ and $\sum \alpha_{ik} = 1$. Likewise $\widehat{\mu} = \sum_k \beta_k \mu_k$ with $\beta_k \geq 0$ and $\sum \beta_k = 1$. Therefore the full prediction has an exact affine decomposition into raw training responses.*

Proof. Normalize the nonnegative kernel-mask products by their positive sums. Expanding $R^{\mathrm{res}}_{ik} = R_{ik} - \mu_k$ and each μ_k as an observed-cell mean gives the final affine coefficients. They need not all be nonnegative, which is why the main text avoids calling the raw-response decomposition convex. □

Proposition 4 (Leave-one-out identity)**.** *For residual prediction $\widehat{r} = S/Z$ and normalized weight $\alpha = w/Z < 1$, the old-minus-new change after removing observation (i, k) is the main-text leave-one-out equation.*

Proof. The removed prediction is $(S - wR^{\mathrm{res}}_{ik})/(Z - w)$. Subtracting it from S/Z and substituting $\alpha = w/Z$ yields the expression. The Lean file checks the cross-multiplied polynomial identity without division automation. □

G Additional Estimator Derivations

G.1 Per-drug residualization

Write $R_{ik} = \mu_k + e_{ik}$ for observed pairs. A cold-start predictor must estimate both the drug-wide offset and the cell-specific interaction. Direct NW smoothing can let large cross-drug mean differences obscure the interaction. We therefore smooth e_{ik} over the cell-drug product geometry and interpolate μ_k only over drug geometry. Every quantity for query j^* uses $k \neq j^*$.

G.2 Vectorized inference

The (i, j) entry of $K_c R^{\mathrm{res}} K_d^\top$ is $\sum_{a,k} K_c(i, a) R^{\mathrm{res}}_{ak} K_d(j, k)$, exactly the numerator of the pairwise estimator. Replacing the residual matrix by M gives its denominator. Two dense matrix multiplications cost $\mathcal{O}(n^2 m + n m^2) = \mathcal{O}(nm(n + m))$. The Lean audit expands and checks a two-by-two instance, which is the finite algebraic core of the index identity.

G.3 Bandwidth heuristic

For a representation matrix Z, the implementation samples feature pairs with a fixed random generator and takes the median positive cosine distance before applying the RBF scaling. This is label-free but not parameter-free: representation choice, pooling, fingerprint radius, and the pathway weight remain modeling decisions.

H Non-identifiability Outside Training Support

The result below is an impossibility construction showing that response labels alone cannot identify a unique extrapolation. It is deliberately not a claim that trained approximators fail; it bounds every estimator, ours included.

Proposition 5 (Cold-start extension ambiguity)**.** *Let H be a nonempty closed training support in drug-embedding space and let the observed response be the constant μ on H. For any $L > 0$, define*

$$f_+(z) = \mu + L \operatorname{dist}(z, H), \qquad f_-(z) = \mu - L \operatorname{dist}(z, H).$$

Both functions are L-Lipschitz and agree with every training label. At a query z^* at distance r^* from H, they differ by $2Lr^*$.

Proof. Distance to a nonempty closed set is one-Lipschitz. Scaling and adding a constant therefore makes each function L-Lipschitz. Distance is zero on H, so both match the labels there. Their difference at z^* is $2L\mathrm{dist}(z^*, H)$. □

The construction has a direct minimax consequence, which is the form the main text cites: no estimator - trained, retrieval-based, or otherwise - can be accurate on both extensions at once.

Corollary 6 (Cold-start minimax lower bound)**.** *For any single prediction $\hat{y} \in \mathbb{R}$ at the query z^*,*

$$\max\{|\hat{y} - f_+(z^*)|,\ |\hat{y} - f_-(z^*)|\} \ \geq\ L\,\mathrm{dist}(z^*, H).$$

Proof. Write $f_+(z^*) - f_-(z^*) = (\hat{y} - f_-(z^*)) - (\hat{y} - f_+(z^*))$ and apply the triangle inequality: $2L\,\mathrm{dist}(z^*, H) \leq |\hat{y} - f_+(z^*)| + |\hat{y} - f_-(z^*)| \leq 2\max\{\cdot\}$. The bound is zero exactly on the support, so it degrades gracefully rather than being vacuous or catastrophic. □

This corollary does not prove that every trained head collapses or that K-PRS always wins; it establishes a floor that applies equally to us. What an empirical-risk objective lacks is any label-derived basis for choosing between opposite extensions in even a simple realisable case. Regularisers, architectural priors, and pretrained geometry choose an extension; the matched LODO results measure whether that choice works here, and the support diagnostics of Appendix K are the practical response to the bound, since $\mathrm{dist}(z^*, H)$ is observable at test time without any label.

The Lean theorems `cold_start_ambiguity`, `cold_start_minimax` and `cold_start_ambiguity_zero_on_support` check the proposition, the corollary, and the degeneracy on the support, over an arbitrary real metric space.

I Full Primary Results and Statistical Reporting

Table 2: GDSCv2 strict LODO on the matched 941-cell protocol. Confidence intervals resample held-out drug tasks. The MolFormer plus scGPT row is a contextual result from the FUSED paper's distinct protocol, not a matched rerun and not the FUSED fusion architecture itself.

Method	Mean PCC	95% CI	RMSE
K-PRS dual feature + pathway	**0.7085**	**[0.6945, 0.7218]**	**2.294**
Cell mean	0.6442	[0.6277, 0.6604]	2.625
Drug-axis kernel ridge diagnostic	0.6383	[0.6221, 0.6525]	2.469
Drug-only k-NN	0.5879	[0.5699, 0.6046]	-
Reported MolFormer + scGPT (561 cells, 223 drugs)	0.4470	-	-
Two-way k-NN	0.1566	[0.1493, 0.1634]	2.633
Trained ridge head	0.1480	[0.1419, 0.1541]	3.280

The drug-axis kernel-ridge row above smooths each cell's response across the drug kernel and is retained as a diagnostic; it is not a fully separable dyadic KRR. The matched stronger controls below use common outer folds and response-supervised inner selection.

Trained-head controls. Under strict one-column LODO the trained-head control is the ridge head of Table 2, fitted on every observed pair of the remaining drugs. The fusion-head control is evaluated under the grouped protocol below, where nested drug-group validation selects its configuration; a single untuned configuration under one-column holdout would be the weaker comparison.

Two-step KRR tuning. The separable KRR uses the same K_c and conservative K_d as K-PRS. Missing centered training responses are filled with zero, which equals the corresponding training-drug mean before centering. Five outer drug folds use seed 2027. Within each outer training set, three drug-group folds select $\lambda_c \in \{0.01, 0.1, 1\}$ and $\lambda_d \in \{0.01, 0.1, 1, 10\}$ by mean per-drug PCC. Selected pairs are $(0.01, 1)$ in four folds and $(0.01, 10)$ in one. The response of every outer test drug remains inaccessible during selection and fitting.

Table 3: Matched five-fold grouped LODO. Each outer test fold removes all of its drug columns simultaneously. Confidence intervals bootstrap the resulting 295 held-out drug tasks. The cell-mean row is the same baseline as in Table 2, re-evaluated under this protocol; the two values differ by 0.0005 purely because the folds differ, and neither is a rerun of the other.

Method	Mean PCC	95% CI	RMSE
K-PRS (metadata-only ε)	**0.6991**	**[0.6846, 0.7126]**	**2.373**
Two-step/Kronecker KRR	0.6620	[0.6482, 0.6744]	2.601
Cell mean	0.6437	[0.6272, 0.6597]	2.632
Interaction fusion head	0.6379	[0.6217, 0.6537]	2.618

Interaction fusion tuning. Label-free PCA reduces each frozen branch to 64 components. Separate linear-ReLU projections form h_c and h_d; the regressor receives $[h_c, h_d, h_c \odot h_d]$. The multiplicative term is the whole point of the control: it lets the head represent a cell-by-drug interaction rather than two additive effects. There is no attention module anywhere in this head, and we describe it throughout as an interaction fusion head rather than a cross-attention one. It minimizes MSE on standardized log-IC50 with AdamW. For each outer fold, 20% of the training drugs are an inner validation set; early stopping has patience four and a 20-epoch cap, and selection uses mean per-drug PCC. The grid is (hidden units, dropout, learning rate, weight decay) $\{(64, 0, 10^{-3}, 10^{-4}), (64, 0.1, 3\times10^{-4}, 10^{-3}), (128, 0.1, 10^{-3}, 10^{-4})\}$ with batch size 8192. The first configuration is selected in all five folds, after which it is refit on the outer training drugs for the selected epoch count. Fold PCCs are 0.6131, 0.6496, 0.6233, 0.6516, and 0.6521.

Paired tests. The bootstrap intervals above resample drug tasks independently for each method, so overlapping intervals do not settle a within-drug comparison; the paired two-sided Wilcoxon signed-rank test over the same 295 per-drug correlations does. On the grouped folds K-PRS beats the cell mean on 209 of 295 drugs ($p = 3.3 \times 10^{-18}$) and the interaction fusion head on 239 of 295 ($p = 6.8 \times 10^{-27}$). The fusion head's mean sits only 0.0057 below the cell mean and its interval overlaps, yet it loses to that prior on 226 of 295 drugs and the paired test is decisive ($p = 1.8\times10^{-20}$): the small mean gap is a consistent per-drug deficit, not a coin flip. Under strict one-column LODO the same test puts K-PRS above the cell mean on 199 of 295 drugs ($p = 2.0 \times 10^{-14}$), while the ridge head loses on *all* 295 ($p = 4\times10^{-50}$). We report these as evidence about this matrix and these controls, not as a general claim about trained architectures.

Table 4: Distribution of the 295 per-drug correlations for the headline run.

Statistic	Value
Mean / median	0.7085 / 0.7230
Standard deviation across drugs	0.1193
Fifth / twenty-fifth percentile	0.4801 / 0.6577
Seventy-fifth / ninety-fifth percentile	0.7862 / 0.8663
Drugs above 0.70 / 0.50	177 / 279
Drugs below 0.30	2
Mean per-drug Spearman	0.6944
Overall RMSE	2.2936

Table 4 summarises the spread behind the headline mean, which matters because a per-drug average can hide a long low tail; here only two of 295 drugs fall below 0.30. The bootstrap uses 2000 resamples and seed zero. Confidence intervals do not remove dataset or protocol uncertainty and should not be interpreted as confidence about clinical deployment.

Figure 11 summarises accuracy across the three evaluated matrices. The two GDSCv2 constructions agree to within 0.001, so the headline does not depend on the exact expression overlap; GDSC1 is lower in absolute terms under its simpler configuration (Appendix E) but again exceeds its own cell-mean prior.

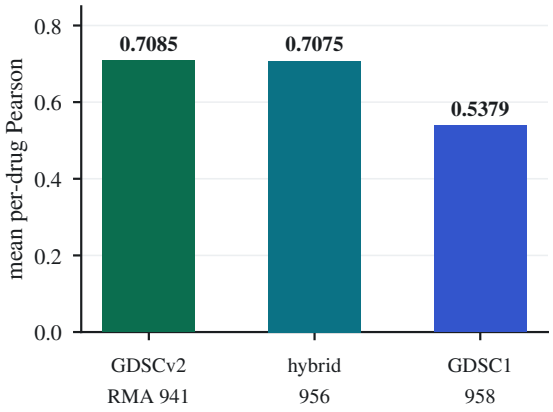


Figure 11: Mean per-drug Pearson of K-PRS on the three evaluated matrices: GDSCv2 with RMA expression (941 cells), the hybrid RMA-plus-DepMap expression construction (956 cells), and GDSC1 (958 cells).

Table 5: Drug representation and pathway ablations. Minor differences between the exploratory 690-cell and final 941-cell runs are labeled rather than pooled.

Variant	Mean LODO PCC	Protocol
Full ECFP4 + ChemBERTa-2 + pathway, $\varepsilon = 0.02$	0.7085	941 cells
ECFP4 + pathway, $\varepsilon = 0.05$	0.7049	941 cells
ChemBERTa-2 + pathway, $\varepsilon = 0.05$	0.6944	941 cells
Pathway only, no chemistry	0.6926	941 cells
Chemistry only, no pathway	0.6561	941 cells
Additive pathway, best $\lambda = 5$	0.6913	exploratory
ECFP radius 3, 2048 bits	0.7051	exploratory
ECFP radius 2, 4096 bits	0.7064	exploratory
Multiplicative pathway, $\varepsilon = 0.10$	0.6974	exploratory

Table 6: Full-protocol bandwidth sensitivity. Values multiply the geometry-only median-distance bandwidth. Rows for cell multipliers 0.25-2 are identical to four decimals; at 4, the largest deviation is below 3×10^{-5}.

Drug multiplier	0.25	0.5	1	2	4
Mean LODO PCC	0.6341	0.6859	**0.7085**	0.7005	0.6973

J Ablations and Bandwidth Sensitivity

Table 5 separates the component ablations by protocol and Table 6 reports the drug-bandwidth row of the full 5×5 multiplier grid. The cell-axis invariance is not generic bandwidth robustness: the selected cell bandwidth is 1.07×10^{-4} and the resulting kernel is almost self-local. The drug axis is materially sensitive, especially when narrowed. The median drug rule is best on this declared grid, but the result should not be extrapolated to other representations or datasets.

Table 7: Pathway-weight sensitivity under the conservative unknown-category fallback. No response labels choose ε; the count-balance rule gives 0.042553 and lies between the displayed 0.02 and 0.05 settings. Every row applies the conservative fallback, so the $\varepsilon = 0.02$ entry is 0.7051 rather than the headline 0.7085: the headline uses the literal shared-pathway indicator at the same ε, and the 0.0034 difference is the price of refusing to treat "Unknown", "Other" and "Unclassified" as a shared mechanism.

ε	.001	.005	.01	.02	.05	.10	.25	1
PCC	.6944	.6994	.7026	.7051	.7036	.6963	.6797	.6561

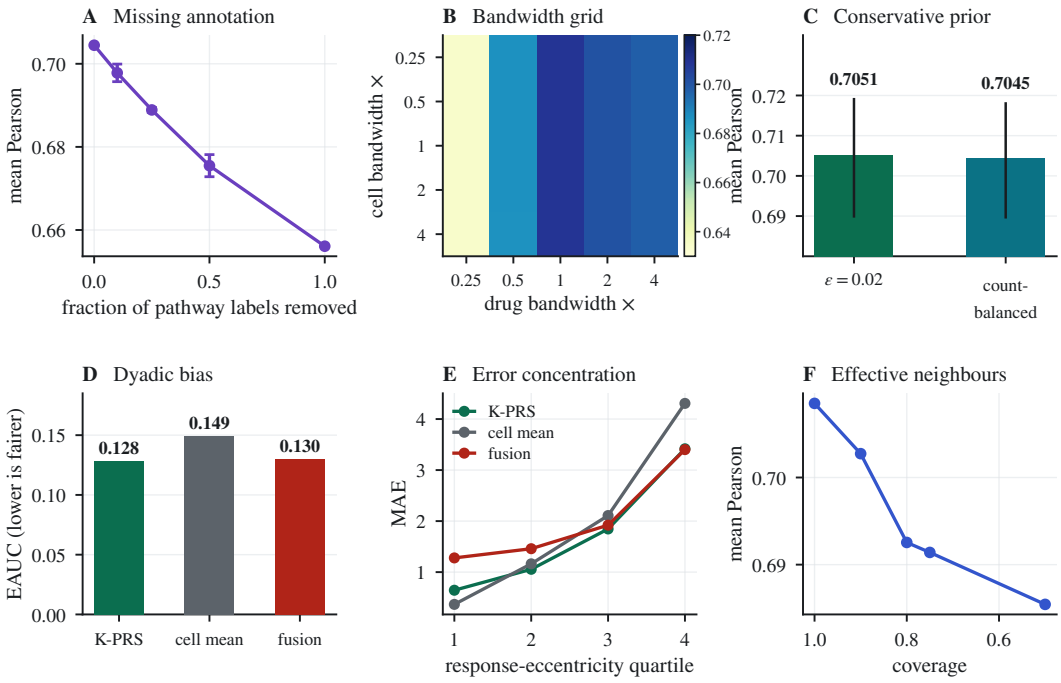


Figure 12: **Stress tests and bias audit. A** Missing pathway annotation; error bars are standard deviations over three masks. **B** The cell and drug bandwidth grid. **C** The conservative unknown-category prior across ε. **D** Cold-start EAUC, which adapts EAUC to unseen drugs through the deployable cell-mean prior; lower is better. **E** Error by eccentricity quartile. **F** Abstention by effective neighbour count. Full declared ranges are shown rather than only their optima.

Table 7 gives the cross-pathway weight sweep, and Figure 12 the remaining stress tests and the bias audit. At missing-pathway fractions 0, 0.10, 0.25, 0.50, and 1, the metadata-only count-balance variant reaches PCC 0.7045, 0.6978 ± 0.0021, 0.6889 ± 0.0006, 0.6755 ± 0.0027, and 0.6561. The final value is the chemistry-only fallback.

Three support diagnostics, one gated kernel. All three scores are computed from the off-diagonal rows of the headline gated drug kernel and use no response labels, so their thresholds are quantiles of kernel support only and none of these curves tunes on the retained drugs' responses. Nearest neighbour is the largest off-diagonal similarity, kernel mass is the mean similarity, and effective neighbour count is $(\sum_k K_{jk})^2 / \sum_k K_{jk}^2$. At retained coverages 90%, 80%, 75%, and 50%:

- nearest neighbour, thresholds 0.04212, 0.05054, 0.05538, 0.08061, PCC 0.7136, 0.7201, 0.7227, 0.7354 - monotone improvement;
- kernel mass, PCC 0.7131, 0.7155, 0.7196, 0.7189 - a smaller and finally flat gain;
- effective neighbour count, PCC 0.7027, 0.6925, 0.6914, 0.6854 - monotone degradation.

Only the first is a usable abstention rule. Effective neighbour count rewards drugs whose similarity is spread evenly over many neighbours, which is exactly the diffuse evidence regime the estimator handles worst, so it selects the wrong drugs to answer. Figure 6B in the main text plots the first two curves and Figure 12F the third.

Standard EAUC defines eccentricity using training means of both dyad entities (Paz-Ruza et al., 2025). Because an unseen drug has no training mean, we define cold-start EAUC (CS-EAUC) using the cell-mean predictor as the deployable prior and normalize by the observed response range. K-PRS, grouped interaction fusion, and cell mean score 0.1281, 0.1302, and 0.1490 (lower is better). Across increasing eccentricity quartiles, K-PRS MAE is 0.645, 1.056, 1.846, and 3.417, compared with cell mean 0.368, 1.162, 2.109, and 4.307. Thus K-PRS is worse in the easiest quartile but reduces the high-eccentricity failure tail; reporting only average PCC would hide that tradeoff.

K Novelty-Stratified Analysis

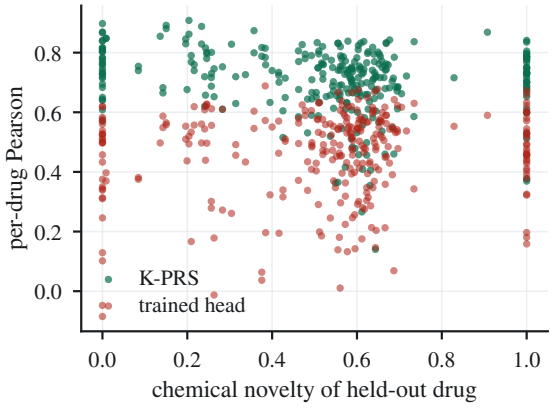


Figure 13: Per-drug correlation against drug novelty, one point per held-out drug, for K-PRS and the ridge probe; the quartile means of the same data are in Figure 6A. Both series are the exploratory 690-cell protocol rather than the 941-cell headline one, so no value here should be equated with a full-protocol number, the trained-head probe least of all. The scatter is shown because the quartile means hide how wide the within-bin spread is relative to the differences between bins.

Protocol. This analysis uses the 690-cell intersection with ECFP4-only drug features and $\varepsilon = 0.05$. Its overall means are K-PRS 0.7066, cell mean 0.6476 and ridge head 0.4622, none equal to the corresponding 941-cell values (0.7085, 0.6442, 0.1480); in particular the ridge probe is much stronger on the smaller intersection. The quantity of interest is the shape of accuracy across novelty strata, which Figure 6A also plots.

Novelty is one minus the maximum cosine similarity to any training drug in the feature representation used by the probe. The quartile counts are 74, 73, 74, and 74. K-PRS achieves 0.764, 0.693, 0.664, and 0.704 across the bins; cell mean achieves 0.631, 0.651, 0.635, and 0.674; the ridge probe achieves 0.436, 0.459, 0.464, and 0.490. Two features of this plot argue against over-reading it. K-PRS's last bin is non-monotone, so a scalar nearest-neighbour novelty measure does not capture pathway support or response-profile difficulty. And every curve, including the trained head's, is flat or mildly *increasing* in novelty, which is the opposite of what a difficulty axis should produce: nearest- neighbour distance in a 2048-bit fingerprint space separates scaffold families more than it separates predictability, so the most "novel" quartile is not the hardest one. We therefore present this as a descriptive stratification, not as a causal test of the memorization mechanism, and the minimax bound of Appendix H rather than this figure is what formalises the cost of leaving the support.

L SYNTHETIC COLD-START BENCHMARK

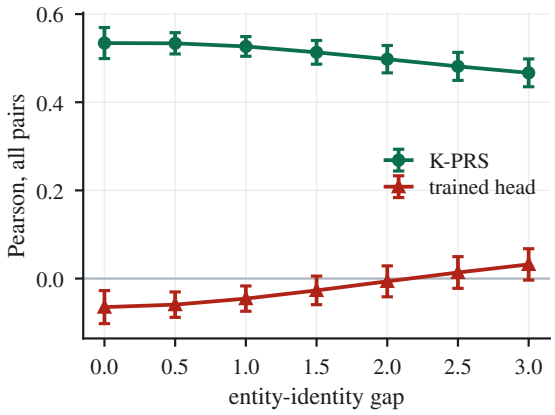


Figure 14: Five-seed synthetic sweep, averaged over all drugs; error bars are one standard deviation over seeds. Figure 6C shows the held-out drugs. The finding is the separation between the curves, not the drift along the horizontal axis (see text).

The generator. For 100 cells and 40 drugs, of which 8 are test drugs, we draw unit cell embeddings $z_i^c \in \mathbb{R}^8$ and a linear cell effect c_i, draw a smooth drug function g from a Gaussian process with a short-lengthscale (0.3) RBF kernel over unit drug embeddings $z_j^d \in \mathbb{R}^{24}$, standardise g, and form $R_{ij} = c_i\{1 + g(z_j^d)\} + \epsilon_{ij}$ with ϵ Gaussian at scale 0.3 and density 0.9. Each test drug starts as a copy of a randomly chosen training drug and is displaced: $z^d_{\text{test}} = \text{normalise}(z^d_{\text{train}} + \text{gap} \cdot \varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, I_{24})$.

What the gap axis measures. Two properties of the design limit what the horizontal axis can show.

First, the gap axis saturates almost immediately. The displacement has expected norm $\text{gap} \cdot \sqrt{24} \approx 4.9\,\text{gap}$ against a unit base vector, so by $\text{gap} = 0.5$ the perturbation already dominates and, after renormalisation, a test embedding is close to a uniformly random direction. Measured on this generator at its default size (50 drugs, 10 of them test, noise 0.2, five seeds), the zero-training estimator's held-out correlation is 0.698 at $\text{gap} = 0$ and then 0.6846, 0.6841, 0.6840, 0.6840, 0.6840, 0.6840 across $\text{gap} = 0.5$ to 3 - flat to four decimals. Over the same first step the mean largest normalised drug weight falls from 0.147 to 0.058 and the effective neighbour count rises from 23 to 34, and neither moves afterwards. Increasing the gap beyond 0.5 therefore does not make the task harder; it cannot, because the test drug is already unrelated to its parent.

Second, per-drug Pearson on this generator is close to a sign test. The truth for drug j is $c_i\{1 + g_j\}$, so any prediction proportional to a smoothed c_i correlates at $+\rho$ when $1 + g_j > 0$ and $-\rho$ when $1 + g_j < 0$; measured per drug, the two modes sit near $+0.90$ and -0.57. The mean over eight test drugs is then governed by how many of them the Gaussian-process draw happened to give a positive response slope, which is a property of the generator's randomness rather than of novelty.

Finding. In the plotted sweep the zero-training estimator's held-out correlation exceeds the trained MLP's in all 35 gap-and-seed runs, by at least 0.149; the means rise from 0.540 to 0.667 against -0.034 to 0.376 across gaps 0 to 3. That sweep was produced with an earlier configuration of the generator, and its upward drift along the gap axis does not appear in the generator described above, which is flat beyond $\text{gap} = 0.5$; given the two properties above we do not interpret the drift. The finding we draw is the separation: a zero-training smoother over the same geometry transfers to unseen drugs where a head fitted to the response labels does not, matching the pattern of the GDSC results.

M EXACT EVIDENCE AND FAITHFULNESS

For a query, the estimator yields the unnormalised pair kernel w, the normalised residual coefficients α, their sum, and the separately normalised drug-offset coefficients β, all under the pathway-gated drug kernel. An explanation is complete only when both pieces are shown. For a chosen training pair, faithfulness compares the directly recomputed old-minus-new prediction with the closed-form leave-one-out identity. Over 200 sampled query pairs on the 941-cell headline run the rank correlation is 1.000, as algebra predicts. This is not comparable to a human-interpretability score: it says the reported coefficient effect exactly matches the estimator, not that the evidence is causal or clinically meaningful. For the same reason we do not compare against post-hoc attribution methods such as SHAP or LIME, which answer a different question.

N CONFORMAL CALIBRATION

All three intervals below are calibrated on the same LODO predictions from the headline estimator - dual chemical features, pathway gate at $\varepsilon = 0.02$, label-free global bandwidths, mean per-drug PCC 0.7085 - using the same seed-zero division into 147 calibration and 148 test drugs. They differ only in the nonconformity score, so their coverages and widths are directly comparable.

N.1 PAIRWISE EMPIRICAL INTERVAL

Every calibration residual is itself obtained from a LODO prediction. Pooling observed cell-level absolute residuals gives $117{,}131$ calibration scores, corrected quantile $q = 3.7303$, pairwise coverage 0.9184 at nominal 0.90, and width 7.4605. Because scores from cells within one drug share a held-out entity and training fit, arbitrary pairwise exchangeability is a strong assumption. The main paper therefore calls this result empirical calibration, not an unconditional distribution-free guarantee. Dependent data require an exchangeability unit or dependence model matched to the sampling process (Chernozhukov et al., 2018).

Proposition 7 (Standard split-conformal coverage)**.** *If n_{cal} calibration scores and the test score are jointly exchangeable, and q is the $\lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil$ order statistic (clipped at the largest calibration score), then* $\Pr\{S_{\mathrm{test}} \le q\} \ge 1-\alpha$.

Proof. Under exchangeability, the randomized rank of the test score among the pooled scores is uniform. At least a $1-\alpha$ fraction of ranks lie below the selected corrected order statistic. Ties can only increase coverage under the non-strict acceptance rule. □

N.2 DRUG-LEVEL GROUP CONFORMAL

For the exchangeability unit actually induced by LODO, define one calibration score per drug, for example $S_j = \max_{i:M_{ij}=1} |R_{ij} - \widehat{R}_{ij}|$. Applying the same corrected rank quantile over drugs yields simultaneous coverage of the complete observed response profile of a new exchangeable drug. A less conservative prespecified within-drug quantile yields that level of within-profile coverage. The max-score interval has 0.9054 empirical group coverage, 0.9896 pair coverage, $q = 7.1427$, and width 14.2854; it is wider than the pairwise interval because it protects the stronger all-observed-cells event.

For a sharper declared target, define S_j as the 90th percentile of absolute errors within drug j. Calibrating these 147 drug scores gives $q = 4.7410$ and width 9.4819. On the 148 test drugs, 0.9324 have at least 90% of observed cells covered; pair coverage is 0.9515. This interval is 33.6% narrower because it does not promise simultaneous coverage of every cell. It is a block-exchangeable analysis over drugs, not a correction that makes individual cells independent.

Weighted conformal is a distinct possible remedy when test covariate shift is specified and density ratios are known or can be estimated (Tibshirani et al., 2019). We do not report it here because estimating a reliable density ratio for a genuinely new chemical mechanism from 295 drugs would introduce another unvalidated model.

O COUNTERFACTUAL GRADIENT AND CERTIFICATE

Let z be the query-cell embedding, $k_i(z) = exp[-\|z - z_i\|^2/(4\sigma_c^2)]$, and let drug-side weights be fixed. Define $a_i = \sum_k K_d(d,k)R^{\text{res}}_{ik}$, $b_i = \sum_k K_d(d,k)M_{ik}$, $S(z) = \sum_i k_i(z)a_i$, and $Z(z) = \sum_i k_i(z)b_i$. Then

$$\nabla k_i(z) = \frac{k_i(z)}{2\sigma_c^2}(z_i - z),$$

and the quotient rule gives

$$\nabla \widehat{r}(z,d) = \frac{1}{Z(z)} \sum_i \frac{k_i(z)}{2\sigma_c^2}(z_i - z)\{a_i - \widehat{r}(z,d)b_i\}. \tag{7}$$

The factor is $1/(2\sigma_c^2)$, not $1/\sigma_c^2$, because of the $4\sigma^2$ denominator in the RBF-cosine form, and the factor of two is easily lost when moving between the cosine and Euclidean forms. The expression is checked against finite differences; Lean verifies the constant, the denominator arithmetic and the quotient rearrangement.

Let $[m_{\min}, m_{\max}]$ bound centered observed responses and

$$C_i = \max\{|a_i - b_i m_{\min}|, |a_i - b_i m_{\max}|\}.$$

For all z' in a radius-r ball around z,

$$L_d(r) = \frac{1}{2\sigma_c^2} \max_{i:b_i>0} \frac{(\|z_i - z\| + r)C_i}{b_i} \tag{8}$$

bounds the gradient norm. This follows from Equation 7, the convex range of $\widehat{r}$, the triangle inequality, and the weighted-ratio bound.

For log IC50, *lower is preferred*. If d_1 is the recommended drug and $g(z) = \widehat{R}(z,d_2) - \widehat{R}(z,d_1) > 0$, then g is $L_{d_1}(r) + L_{d_2}(r)$ Lipschitz. Hence the main-text certificate inequality ensures $g(z') > 0$ throughout the ball. PGD is retained only as a falsification search; its failure to flip a ranking is not itself a proof.

The certificate is tight. Equation 6 cannot be improved using Lipschitz information alone. Take $f_1(y) = L(r - y)$ and $f_2(y) = L(y - r)$ on the real line with query $z = 0$. Both are L-Lipschitz on the radius-r ball, so each satisfies the hypothesis with constant L, and the margin at the query is exactly $f_1(0) - f_2(0) = 2Lr = r(L + L)$, that is, exactly the certificate's threshold. Yet at $z' = r$, which lies on the ball, both functions equal zero and the strict preference has already been lost. Any claim of a strictly larger certified radius therefore requires information beyond the two Lipschitz constants. This witness is machine-checked as `ranking_certificate_tight`; a companion theorem, `ranking_certified_margin`, records the stronger fact that the margin stays above $g(z) - r\{L_{d_1}(r) + L_{d_2}(r)\}$ everywhere on the ball, not merely positive at the worst point.

We compute warm-start predictions under the headline pathway-gated drug kernel, order drugs by increasing predicted log IC50, and pass the same lower-is-better convention to both the certificate and the PGD search. Across 100 fixed-seed queries with mean margin 0.4190 log-IC50, the certified radii range from 0 to 5.12×10^{-8}, with quartiles 1.86×10^{-9}, 6.52×10^{-9} and 1.49×10^{-8} and mean 1.04×10^{-8}; all fall below the diagnostic threshold 0.01, and twelve of the hundred queries admit no positive certified radius at all. The median radius is 6.1×10^{-5} of the selected cell bandwidth. PGD reports no flip within its search budget, but this does not enlarge the proved ball. The small radii are an important negative result, not evidence of practical stability: they follow from $\sigma_c = 1.07 \times 10^{-4}$, which makes the cell kernel almost self-local and hence the local Lipschitz constant enormous.

O.1 EXACT CERTIFICATE AGAINST RESPONSE-EVIDENCE NOISE

The embedding bound controls movement of the query representation and can be loose because it upper-bounds a quotient over an entire ball. A complementary question is how much the recorded response evidence could change before the selected ranking changes, holding embeddings and kernel weights fixed. Expand residual centering and the offset channel to write each prediction as $\widehat{R}_d =$

$\sum_{ik} \gamma_{d,ik} R_{ik}$. For preferred d_1, runner-up d_2, and margin $g = \widehat{R}_{d_2} - \widehat{R}_{d_1} > 0$, perturbation δ changes the margin by

$$\langle \gamma_{d_2} - \gamma_{d_1}, \delta \rangle \geq -\|\gamma_{d_2} - \gamma_{d_1}\|_1 \|\delta\|_\infty .$$

Therefore every perturbation with $\|\delta\|_\infty < g/\|\gamma_{d_2} - \gamma_{d_1}\|_1$ preserves the ranking. Hölder's bound is attained by a sign-aligned perturbation, so the radius is sharp for this affine map and threat model.

We evaluate it on the same 100 rankings of the same estimator as the embedding certificate, so the two threat models are directly comparable. The median radius is 0.0848 log-IC50, the quartiles are 0.0323 and 0.1774, and the 10th and 90th percentiles are 0.0090 and 0.2735; no query has radius zero. Fractions exceeding 0.05, 0.10, and 0.25 are 66%, 46%, and 16%. This is more informative than the global embedding bound for assay-value noise, but it assumes fixed weights, simultaneous bounded perturbations, and no representation or metadata error. It cannot be converted into a chemical-edit or clinical guarantee.

Algorithm 1 Projected search used only to challenge a certified ranking

Require: query z, preferred d_1, runner-up d_2, radius r, step η

$\delta \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  $g \leftarrow \widehat{R}(z + \delta, d_2) - \widehat{R}(z + \delta, d_1)$
  if $g < 0$ **then**
    return flip found
  end if
  $\delta \leftarrow \delta - \eta \nabla g / \max(\|\nabla g\|_2, 10^{-12})$
  project δ onto the radius-r Euclidean ball
end for
return no flip found (not a certificate)

P Runtime and Scaling

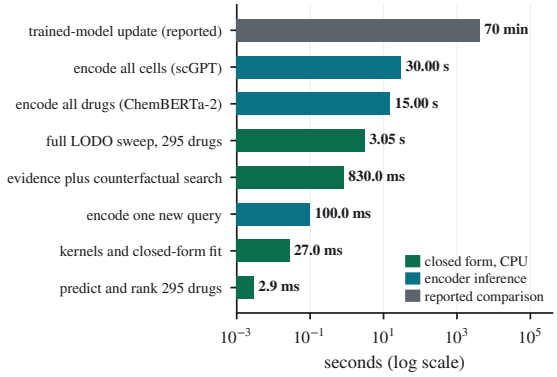


Figure 15: Recorded operation times on a logarithmic scale. Encoder inference is a one-time cached cost; the trained-model bar is a reported update time on different hardware, shown for scale only.

Table 8: Runtime breakdown.

Operation	Time	Scope
Encode cells with scGPT	about 30 s	one-time GPU, cached
Encode drugs with ChemBERTa	about 15 s	one-time GPU, cached
Compute kernels / closed-form fit	0.027 s	CPU NumPy
Encode a new query	about 0.1 s	encoder inference
Predict and rank 295 drugs	2.9 ms	CPU NumPy
Full evidence plus counterfactual search	0.83 s	CPU NumPy
Trained-model update (reported)	about 70 min	different hardware

The trained-model update in the last row of Table 8 is a reported figure on different hardware, quoted as 70-150 minutes depending on architecture; the table and Figure 15 use the conservative lower end. We do not report a speed-up ratio, because hardware, implementation and retraining policy differ. The operational claim is narrower: K-PRS appends observations and recomputes closed-form kernels, whereas a response-supervised head must be optimised again before new labels can change its parameters.

Q Exploratory Analyses on an Earlier Protocol

The analyses in this section use an earlier protocol with a smaller cell intersection and different representations, on which K-PRS reaches mean per-drug Pearson 0.558 (median 0.584, standard deviation 0.118, RMSE 2.193, 18 drugs below 0.30 against two on the headline protocol). They are not comparable with the 941-cell results, and because their intermediate predictions are not among the released artifacts they are reported descriptively and excluded from every headline conclusion.

Cross-platform transfer. With the evidence base built on GDSCv2 and evaluation on shared CCLE measurements, K-PRS reaches 0.491 against 0.381 for FUSED, compared with 0.558 and 0.428 within GDSCv2 under the same protocol.

Pharmacological strata. Table 9 summarises accuracy by post-hoc drug class. These post-hoc classes were not prespecified and should be treated as hypotheses for a future stratified evaluation, not evidence that any class is clinically predictable.

Table 9: Per-class summary on the earlier protocol (exploratory).

Drug class	Mean LODO PCC	Count
Broad-spectrum cytotoxic	0.72	28
Epigenetic inhibitors	0.68	34
Cell-cycle / broad kinase inhibitors	0.57	89
Targeted single-mutation therapies	0.31	18
Other	0.54	126

R Warm-Start HCC827 Illustration

The HCC827 example is *not LODO*. HCC827 and its observed drug responses are present in the evidence matrix; the purpose is to show how evidence can be inspected for a familiar cell, not to estimate cold-start performance. The illustration uses the earlier protocol of Appendix Q. The largest representer term for Romidepsin is the observed HCC827-Romidepsin pair with coefficient 0.139 and log IC50 -5.8, which is legitimate in the declared warm-start setting but would be leakage in a held-out-pair claim.

Table 10: HCC827 warm-start ranking. Mean absolute top-ten error is 0.21 log units, Recall@20 is 55%, and drug-rank Spearman is 0.596.

Rank	Drug	Mechanism	Predicted	Observed
1	Romidepsin	HDAC inhibitor	-4.71	-5.80
2	Bortezomib	Proteasome inhibitor	-4.13	-4.10
3	Sepantronium	IAP inhibitor	-3.64	-3.90
4	Docetaxel	Microtubule stabilizer	-2.85	-3.00
5	Daporinad	NAMPT inhibitor	-2.67	-2.80
6	SN-38	Topoisomerase I	-2.63	-2.70
7	Vinblastine	Microtubule disruptor	-2.25	-2.40
8	Dinaciclib	CDK inhibitor	-2.11	-2.00
9	Dactinomycin	DNA intercalator	-1.98	-2.10
10	Vinorelbine	Vinca alkaloid	-1.78	-1.90

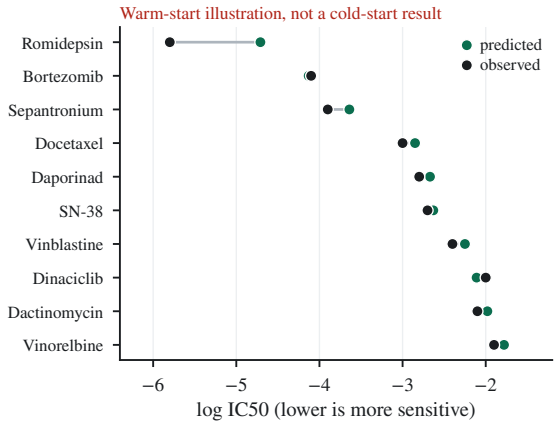


Figure 16: Predicted and observed log IC50 for the ten top-ranked HCC827 recommendations. Lower values indicate greater sensitivity.

S Machine-Checked Mathematics

The development in `formal/` is a Lean 4 project built against Mathlib. It compiles with `lake build`; `formal/check.sh` then runs an axiom audit that prints the axiom dependencies of every cited theorem and fails if a `sorryAx` appears. At the time of writing, 38 theorems are checked and every one depends on exactly the three standard Lean axioms `propext`, `Classical.choice` and `Quot.sound`; no custom axiom is declared.

Every statement below is about real numbers, real inner-product spaces or real metric spaces, so the checked statements are the statements the paper makes rather than integer or rational analogues of them. Table 11 maps each theorem to the claim it discharges.

What is deliberately not formalised. Bochner's theorem (Bochner, 1933), the general Schur product theorem (Schur, 1911), and exchangeability of conformal scores (Vovk et al., 2005) are used with their standard hypotheses and citations. The Schur step is the interesting case: rather than assume the general theorem, the development proves the special form the method actually needs, namely that if A is the Gram matrix of an explicit finite feature map and B is PSD then $A \odot B$ is PSD, and separately exhibits that feature map for H. The remaining gap is therefore only the positive definiteness of K_{chem} itself, which is Bochner. Exchangeability is an assumption about the screen, not a theorem about the estimator, which is precisely why the main text reports empirical rather than guaranteed coverage.

T Extended Discussion

What the result does and does not establish. Our matched experiments show that response-supervised heads over strong frozen representations underperform a cell-mean prior under strict LODO, while pathway-aware retrieval exceeds it on two releases and under a grouped protocol with nested tuning. This does not establish that trained models are intrinsically unsuitable for cold-start drug response. A head with a different inductive bias, a different objective, or access to mechanism-level supervision could do better, and we would expect it to. What the result does establish is narrower and, we think, more useful: random-pair performance and architectural sophistication are not evidence of inductive drug transfer, and the community's default protocol does not distinguish the two. The minimax bound of Appendix H explains why no protocol can fully close that gap, and why an observable support diagnostic is the right practical response.

Why a smoother is a good control rather than merely a weak baseline. A transparent smoother over well-chosen frozen geometry has three properties that a trained head must match before its extra flexibility is justified: it has no response-supervised parameters to overfit, its inductive bias is legible in the kernel rather than implicit in an optimiser trajectory, and every downstream artifact - evidence,

Table 11: Scope of machine checking. Analytic and probabilistic theorems are used with citations and are not misrepresented as formalised.

Lean theorem	Paper statement checked
`sq_dist_of_unit`	Squared distance is affine in the cosine on the sphere
`rbf_cosine_eq`	Equation 1, including the $4\sigma^2$ denominator
`rbf_cosine_pos`, `rbf_cosine_self`	Strict positivity and the value at coincident embeddings
`deriv_gaussian_slice`	The gradient constant is $1/(2\sigma^2)$
`gradient_constants_differ`	$1/(2\sigma^2) \neq 1/\sigma^2$, so the audit is not cosmetic
`PSD.of_gram`	Any Gram matrix of a finite feature map is PSD
`PSD.add`, `PSD.smul`	Closure of PSD under nonnegative combination
`pathwayPrior_same/diff/symm`	Entries and symmetry of the prior in Equation 2
`gram_pathwayPrior`	Explicit finite feature map realising H
`psd_pathwayPrior`	$H = \varepsilon J + (1-\varepsilon)P$ is PSD for $\varepsilon \in [0,1]$
`PSD.hadamard_of_gram`	The Schur step actually used, in constructive form
`psd_drugKernel`	$K_d = K_{\mathrm{chem}} \odot H$ is PSD
`not_psd_onesMinusPathway`	Two-point counterexample refuting the old $J - P$ proof
`alpha_sum_one`, `alpha_nonneg`	The normalised weights form a convex vector
`nw_eq_sum_alpha`	Equation 3 is exactly the weighted average
`convex_comb_mem_Icc`, `nw_mem_Icc`	The residual channel never leaves the evidence range
`prediction_bounds`	The two-channel prediction is affine, with the stated bracket
`masked_sum_eq_observed`	Masked entries contribute to neither numerator nor denominator
`vectorized_eq_double_sum`	Equation 4 equals the elementwise definition
`loo_exact`	Equation 5, exactly and without refitting
`ranking_certified`	Equation 6
`ranking_certified_margin`	The margin stays positive on the whole ball
`ranking_certified_lower_better`	The log-IC50 direction convention
`lipschitz_gives_ball_bound`	The hypothesis follows from a local Lipschitz bound
`ranking_certificate_tight`	A witness showing the radius cannot be enlarged
`cold_start_ambiguity`	Two Lipschitz extensions agreeing on the support
`cold_start_ambiguity_zero_on_support`	The ambiguity gap vanishes on the support, so the bound is not vacuous
`cold_start_minimax`	The $L\,d(z,S)$ lower bound on any single prediction
`card_le_order_statistic`	Counting fact behind the conformal threshold
`conformal_level`	The chosen rank attains the nominal level
`interval_mem_iff`, `interval_width`	Coverage is a score statement; width is $2q$

interval, certificate - is derived from the same weights that produce the score. When a trained model beats it, that is informative. When a trained model does not, as here, the comparison localises the failure to the transfer mechanism rather than to tuning. We therefore recommend reporting such a control in every cold-start dyadic study, alongside the entity-mean prior, and we release the implementation to make that cheap.

U Limitations in Full

Support and representation. K-PRS is an interpolator, not a mechanism simulator. A query with little kernel mass, no close chemical neighbours, or an uninformative pathway assignment receives diffuse evidence and moves toward smoothed training means. The method should therefore abstain or flag low-support queries rather than present every score with equal confidence; Appendix J shows that a nearest-neighbour support rule achieves this, that a kernel-mass rule helps only weakly, and that an effective-neighbour-count rule makes accuracy worse, so the diagnostic must be chosen deliberately. Its geometry also inherits the blind spots of its inputs. scGPT was developed for single-cell measurements but is applied here to bulk expression; cross-platform stability does not remove that domain shift. ECFP4 and ChemBERTa-2 may miss stereochemistry, exposure, metabolism, or target engagement, while pathway labels are curated, coarse, and occasionally missing - Appendix J quantifies the degradation as annotation is removed. Mutation, methylation, copy number, dose, temporal state, and microenvironment are absent even though they can dominate response to targeted therapies.

Evaluation and comparison. Strict LODO measures transfer to unscreened drugs inside two historical screening collections. It does not reproduce prospective medicinal chemistry, where novel scaffolds and mechanisms may lie farther outside support, and GDSC1 and GDSCv2 are not independent clinical cohorts. Per-drug Pearson emphasises ranking cells for each drug; it can coexist with sizable absolute error, as the RMSE and interval width show. The matched ridge, k-NN, kernel-ridge, and fusion controls isolate useful mechanisms, but we did not rerun every published architecture on the identical 941-cell matrix. The reported MolFormer-scGPT number therefore remains a contextual, cross-protocol reference rather than a controlled superiority claim.

Auditability is not causality. Exact weights explain the estimator's arithmetic, not why a drug works biologically. A large coefficient can expose dependence on one observation without validating that observation or identifying a causal pathway. Pairwise conformal coverage is empirical under within-drug dependence, and the cleaner group analysis pays for its assumption with very wide intervals. Likewise, the certified radii are mathematically valid only for perturbations of the frozen embedding and are extremely small in the present experiment. They neither certify invariance to realistic molecular edits nor imply clinical safety. The machine-checked development covers algebra and inequalities; it does not cover Bochner's theorem, exchangeability, or the biological validity of pathway annotation, and Appendix S states that boundary explicitly rather than letting a compile result stand in for a broader claim.

External validity and next tests. All conclusions remain preclinical. The warm-start HCC827 example is an explanation illustration, not evidence of cold-start or patient-level utility, and the cross-platform analysis is exploratory (Appendix Q). The most important next evaluation is prospective: freeze the evidence base and all design choices, register an abstention rule based on support, then assay genuinely new compounds across cell and patient-derived models. Future extensions should add multimodal cell state and dose dependence while preserving the column-level leakage boundary and reporting when extra flexibility reintroduces an identity shortcut. Combination screens (Holbeck et al., 2017) and clinical transfer (Geeleher et al., 2014) are natural but separate problems that we do not claim to address.

V Impact Statement in Full

Intended benefit. The appropriate use of K-PRS is to prioritise laboratory measurements for a previously unscreened compound: identify responsive cell models, expose the historical observations behind that ranking, and direct scarce assay capacity toward informative follow-up. Its closed-form

update may also make it easier for small laboratories to audit and refresh a model without repeatedly training a large response head. Faster prioritisation could reduce redundant screens and make negative or uncertain evidence more visible.

Risks and affected groups. The central harm is automation bias: a precise-looking score, explanation, or certificate may be mistaken for treatment evidence. Cell lines incompletely represent tumour heterogeneity, ancestry, immune context, prior therapy, pharmacokinetics, and toxicity. Errors may therefore concentrate in rare molecular subtypes, novel chemical scaffolds, underrepresented tissues, or mechanisms poorly encoded by expression; the eccentricity analysis in Appendix J shows error concentrating in exactly the atypical response quartile for every method we tested, ours included. The retrieval architecture can also reproduce biases in the screened-drug library and pathway curation. Exact evidence traces make those dependencies inspectable but do not make them causal, fair, or correct. Because the method is readily reusable, it could also be applied to proprietary or patient-derived expression data whose consent, privacy, and governance requirements differ from the public preclinical data used here.

Safeguards and prohibited use. Outputs should display support diagnostics, the complete evidence decomposition, interval width as well as coverage, and an explicit cold-start or warm-start label. Low-mass and out-of-support queries should trigger abstention and additional screening. No prediction should be used to prescribe therapy, exclude a patient, or bypass prospective assays, dose-response validation, pharmacology, toxicity assessment, or qualified human review. Any move to patient-derived or identifiable data requires appropriate consent, access control, privacy protection, subgroup audits, and regulatory review. Redistribution must respect GDSC, expression-resource, and encoder checkpoint terms. These controls preserve the intended role of the system: generating testable preclinical hypotheses, not making clinical decisions.

W Ethics and Intended Use Details

The system is a research prioritization tool for cell-line experiments. It must not be used to prescribe treatment, infer a patient's response from a cell-line embedding, or bypass toxicity and pharmacokinetic studies. Nearest evidence can reproduce biases in the screened compounds, cell-line ancestry, tissue representation, assay design, and curation of target pathways. A high representer weight identifies reliance, not a causal mechanism. A conformal interval calibrates a declared sampling assumption, not biological transport from cell lines to patients. A certificate bounds the software estimator in embedding space, not laboratory or clinical outcomes.

Recommended safeguards are: show the complete evidence decomposition; flag low kernel mass and out-of-support queries; disclose whether a result is LODO or warm-start; report uncertainty width as well as coverage; preserve human review; and require independent experimental validation. If future work uses patient-derived or identifiable data, governance, consent, privacy, subgroup audits, and applicable regulatory review will be required.

X Reproducibility and Code Inventory

The release is organized as follows:

- `kprs/`: kernels, masked NW prediction, baselines, metrics, explanations, conformal utilities, and synthetic generation;
- `scripts/`: data alignment, embedding, primary/baseline/ablation experiments, figure generation, and submission checks;
- `tests/`: unit tests for masking, cold-start exclusion, algebraic identities, conformal quantiles, and certificate conventions;
- `results/`: compact JSON summaries and source-generated PDF figures;
- `formal/`: Lean source and an invocation script;

The minimum CPU-only path regenerates tables and figures from compact results. The full path requires public GDSC data and pretrained encoder checkpoints. Seeds are fixed in scripts; the

bootstrap uses seed zero; the synthetic sweep records each seed in JSON. The package excludes the multi-gigabyte raw and embedding caches and includes a manifest explaining how to rebuild them.

Analyses outside the regeneration path. Two reported analyses do not regenerate exactly from the released code. The synthetic sweep of Appendix L was produced with an earlier configuration of the generator; the released sweep script regenerates it with the current one. The exploratory analyses of Appendix Q lack their intermediate predictions. Every other result in the paper regenerates from the released artifacts, and the release audit checks that the calibration and certificate outputs were computed on the headline estimator.