PaD: JOINT INTERPRETATION OF PERTURBATION EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Inferring causal direction from single- and double-perturbation assays requires understanding how a response depends on its background. The question is how much predictive value independent scoring must lose, and which measured sources change a decision. The central insight is that exchange-invariant evidence can carry no directional sign on its own yet determine how another channel should be read. PaD implements this conditional interpretation with channel-specific and joint kernel terms. Under balanced exchange symmetry with invariant context, arbitrary nonlinear additive scoring recovers at most half of the classification gain available through joint access; exchange-consistent additive decisions recover none. A common-threshold condition makes the half-gain bound exact. Observation-level bounds and independent-test certification connect this limit to fitted learners: nine simulated classifiers attain risk upper bounds at most 9.534%, below the sharp additive limit of 16.361%. Measured yeast comparisons attribute the empirical benefit to joint access rather than to a specific kernel mixture. On the Jonikas development panel, exchange-invariant measurement metadata can change directional decisions, with contributions that depend on the other retained inputs. Source removal, refitted block-mismatch controls and resampling of the evaluation records assess these contributions and their sensitivity to record composition. Together, the results connect a sharp limit on independent scoring to the practical evaluation of which evidence changes which decision, and under which retained context.

1 INTRODUCTION

Researchers combining expression knockouts with double-mutant assays face two decisions: which candidate relation deserves priority, and which endpoint is upstream. Third-party single knockouts supply response fingerprints, while double-mutant phenotypes and measurement records provide complementary evidence. A higher aggregate score alone does not show which source changed either decision.

Independent scoring assigns each channel a contribution whose meaning cannot change with the other input. Yet the same response asymmetry can arise under different propagation backgrounds. A joint predictor can represent that dependence. Showing that it does requires more than beating a single-channel baseline, which may simply use its channel poorly; the comparison must connect the *complete inputs* to the loss imposed by independent scoring.

The counterintuitive case is background evidence unchanged when the two genes are exchanged. It carries no directional sign on its own, yet can decide whether another fingerprint supports or opposes an ordering. *Context changes what directional evidence means.* Section 6 shows both sides: an informative second channel can be redundant, while one with no marginal background signal can remove shared noise.

To make that conditional role testable, PaD makes three linked contributions: a sharp function-class bound with exact attainment; observation bounds and independent simulated tests establishing noise-tolerant separation from that limit; and measured-panel comparisons of joint access, kernel mixture and source contributions. The first two concern a specified mechanism; the third examines predictive use across retained contexts and task readouts.

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.

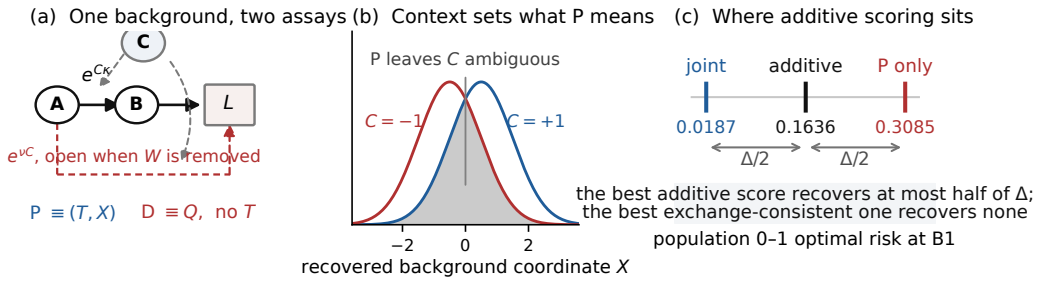

Figure 1: The conditional role of evidence, and the limit it puts on independent scoring. (a): one latent background C scales both the propagation elasticity $e^{C\kappa}$ from A to B and the bypass strength $e^{\nu C}$ in the phenotype readout, whose bypass term opens only when the third partner W is removed (Section 5). The retained fingerprint is information-equivalent to (T, X), the retained phenotype vector to Q, which carries no T. (b): X leaves C partly ambiguous, so one sign of the fingerprint score is consistent with either truth $Y = TC$; joint access reduces this ambiguity. (c): theorem 1 puts the optimal source-additive risk midway between the joint and P-only Bayes risks, and pins the optimal exchange-consistent additive risk at the P-only Bayes risk. Panels (a) and (b) are schematic; panel (c) is corollary 1 at B1, where the common-threshold condition holds, so the midpoint is an equality; it is the floor Section 6.1 certifies against.

2 Related work

From perturbations to pathway order. Expression screens, double-mutant maps and large interventional single-cell screens reveal different cellular dependencies (Kemmeren et al., 2014; Costanzo et al., 2016; Brown et al., 2025), and several lines of work already combine them. Nested effects models infer structure from intervention-associated patterns (Markowetz et al., 2007). Pathway reconstruction from Jonikas interactions (Jonikas et al., 2009; Battle et al., 2010) includes Réd's factorized epistasis scores (Žitnik & Zupan, 2014), which fit a joint factor model with gene-dependent nonlinear transforms and score relations from single and double perturbation phenotypes together. Quantitative epistasis studies combine complementary readouts, such as fitness and reporter expression, to resolve signal propagation, bypasses, readout effects and motif ambiguity (Phenix et al., 2011; 2013). Where these studies reconstruct pathways from complementary assays, this paper quantifies the classification risk imposed by independent scoring on a specified retained interface and the conditions for exact attainment.

From flexible fusion to a task-level guarantee. Synergistic and redundant information, nonadditive fusion, and sums and products of kernels have established foundations (Williams & Beer, 2010; Bertschinger et al., 2014; Barrett, 2015; Kim et al., 2024; Wörtwein et al., 2022; Sonnenburg et al., 2006), and recent work continues to isolate synergistic components (Wen et al., 2025). PaD uses the established Gaussian product construction, equivalent to a Gaussian kernel on scaled concatenated inputs. Its theoretical contribution concerns the resulting task-level question: after observations pass through retained interfaces, how much predictive value is lost by scoring the channels independently?

3 Evidence and prediction tasks

Let $x = (A, B)$ and let $\tau x = (B, A)$ denote endpoint exchange. The response matrix $M_{g,e}$ records gene g under intervention e. The P channel's fixed Proposer encoder (S1) reads

$$\mathcal{E}_x^P = \{M_{A,e}, M_{B,e} : e \notin \{A, B\}\}. \tag{1}$$

The encoder builds on established moment and distribution statistics (Hyvärinen & Smith, 2013; Fonollosa, 2019; Lopez-Paz et al., 2015). Its five antisymmetric features have fixed coefficients giving $s_P(\tau x) = -s_P(x)$, and the nonlinear input z_P holds $\tanh(s_P/2)$ with seven symmetric summaries. S1–S2 specify every feature, mask and coefficient.

The channel z_D retains measured single and double phenotypes, a no-interaction expectation, quality and missingness descriptors, and for Jonikas the third-partner profiles $L_{A,E}$, $L_{B,E}$ with their residuals (S2). Jonikas has 8 P and 23 D coordinates, Costanzo 8 P and 33 D with no matched third-partner context. Throughout, P and D denote the complete retained encoder outputs rather than the raw assays: "complete" covers every coordinate supplied to the learner, since ambiguity in a scalar score may be resolved by the remaining coordinates.

The panels have distinct evaluation targets. Costanzo's are expression-response orientations carried over from an earlier stage of this programme, not direct-edge labels (S1); Jonikas's are KEGG-supported orderings, with unannotated records used as reference negatives rather than confirmed reverse relations. The theoretical label $Y \in \{-1, +1\}$ below is instead a true orientation in a specified structural model.

4 PaD: JOINT INTERPRETATION OF PERTURBATION EVIDENCE

PaD combines Proposer fingerprints (P) with single/double-perturbation phenotypes (D): channel-specific terms represent within-channel patterns, while a product kernel allows one channel to change the interpretation of the other. For fixed channel inputs, define Gaussian kernels using mean squared coordinate distance,

$$k_j(x, x') = \exp\left\{-\frac{\|z_j(x) - z_j(x')\|^2}{2h^2 d_j}\right\}, \qquad j \in \{P, D\}, \quad h = 0.5, \tag{2}$$

where d_j is the input dimension. Training-normalized linear kernels act on the fixed scalar s_P and the exchange-odd D descriptors ϕ_D. The single learned score uses a kernel whose last term is a pointwise product of kernel values,

$$k_\theta = \theta_P k_P^{\text{lin}} + \theta_D k_D^{\text{lin}} + \theta_J k_P^{\text{rbf}} k_D^{\text{rbf}}, \qquad \theta \geq 0, \quad \theta_P + \theta_D + \theta_J = 1. \tag{3}$$

The three terms enter one class-balanced square-loss objective,

$$\min_{F \in \mathcal{H}_{k_\theta}} \frac{1}{n} \sum_{i=1}^{n} w_i (t_i - F(x_i))^2 + \lambda \|F\|^2_{\mathcal{H}_{k_\theta}}, \quad t_i = 2y_i - 1. \tag{4}$$

The relative weights allow either independent or joint terms to contribute, and P remains available through the product kernel even when its linear weight is zero.

The output is decomposed as $d(x) = [F(x) - F(\tau x)]/2$ and $u(x) = [F(x) + F(\tau x)]/2$; orientation uses d, implemented by an exchange-antisymmetrized form of the same kernel, while ordered-reference ranking uses $F = d + u$ and retains symmetric support. S3 gives the solver, normalization and selection rules, and S2 the label-free panel-level scaling of some nonlinear inputs.

Both panels use seven simplex weights and five regularizers, giving 35 candidates. Inner endpoint-excluded predictions define a paired one-standard-error eligibility set, within which the strongest regularization is selected, a stability heuristic (S3). Encoders, kernel family and selection rule are fixed once and used unchanged throughout.

Directional and contextual contributions. The kernel admits direction from either channel, alone or modulated by the other's exchange-invariant component. The data determine which of these contributions is used. With $\text{Odd}[F] = \frac{1}{2}[F(\cdot) - F(\tau\cdot)]$ and superscripts o, e for the odd and even parts of a channel function under the same τ:

Proposition 1 (Additive and product terms admit different paths)**.** $\text{Odd}[a(P) + b(D)] = a^o(P) + b^o(D)$ *and* $\text{Odd}[f(P)g(D)] = f^o(P)g^e(D) + f^e(P)g^o(D)$. *A fitted function in $\mathcal{H}_{k_\theta}$ is a weighted sum of such terms, so its orientation score decomposes as*

$$d(x) = a^o(P) + b^o(D) + \textstyle\sum_\ell \alpha_\ell \left[f_\ell^o(P) g_\ell^e(D) + f_\ell^e(P) g_\ell^o(D)\right]. \tag{5}$$

Under Odd, the $f^o g^o$ and $f^e g^e$ terms cancel, leaving the four classes of contribution in eq. (5). For exchange-invariant metadata M and remaining inputs Z the product path reduces to $\text{Odd}[f(Z)g(M)] = f^o(Z)g(M)$: metadata can enter an orientation score only through joint terms. This describes representable functions, and g^e is a signed modulation, not a reliability weight. Theorem 1 and the certification of Section 6.1 address the achievable loss.

Nonlinear controls. The primary additive control A_{match} replaces only $k_P^{\mathrm{rbf}} k_D^{\mathrm{rbf}}$ by $(k_P^{\mathrm{rbf}} + k_D^{\mathrm{rbf}})/2$ in eq. (3). The secondary additive control A_{rbf} uses $w k_P^{\mathrm{rbf}} + (1-w) k_D^{\mathrm{rbf}}$ over seven fixed w and the same five regularizers, again 35 candidates. Both model within-channel nonlinearities and share the solver, the selection procedure and one selection budget (Cawley & Talbot, 2010). The nonadditive control J uses only $k_P^{\mathrm{rbf}} k_D^{\mathrm{rbf}}$ and the same five regularizers. It assesses whether channel-specific terms add predictive value, using five candidates rather than PaD's 35. These fitted controls complement the unrestricted function-class analysis below.

5 A MECHANISM FOR CONDITIONAL INTERPRETATION

A propagation model makes the conditional role of context explicit and traces it through the full inputs used by PaD.

5.1 LOG-RESPONSE PROPAGATION AND A BACKGROUND-SENSITIVE READOUT

Let $Y = +1$ mean that $(U, V) = (A, B)$ is upstream/downstream order, and $Y = -1$ the reverse. A balanced background $C \in \{-1, +1\}$, independent of balanced Y, sets the relative-response elasticity $e^{C\kappa}$, with $\kappa > 0$. For arbitrary positive baselines u_0, v_0, define

$$\begin{aligned} S &= \delta C + \sigma_P Z_P, \qquad & U_e &= u_0\, 2^{e^S z_e}, \\ V_e &= v_0 (U_e/u_0)^{e^{C\kappa}}, \qquad & \sigma_P &> 0. \end{aligned} \tag{6}$$

With z the fixed third-party effect list of S4.1, the measured responses are exactly $M_U(e) = \log_2(U_e/u_0) = e^S z_e$ and $M_V(e) = \log_2(V_e/v_0) = e^{S+C\kappa} z_e$. Here δ is the background-dependent amplitude offset, σ_P its variability, and S carries between-pair amplitude variation including shared assay variability. The elasticity acts on relative changes, so it survives baseline normalization and matches the expression measurement scale (Kemmeren et al., 2014; Blum et al., 2018; Savageau, 1969).

A third partner W is normally present, $W = 1$, and its removal exposes the bypass. The phenotype assay observes $L(U, V, W) = UV/(u_0 v_0) + e^{\nu C}(1 - W)U/u_0$, a knockout replacing its structural equation by zero. With W removed, upstream knockout gives zero while downstream knockout retains the bypass (S4.2), whose observed positive readout is $B_{\mathrm{obs}} = e^{\nu C + \sigma_D Z_D}$ with $\sigma_D > 0$ and ν the background-dependent bypass strength. The standard normal variables (Z_P, Z_D) have correlation ρ, $|\rho| < 1$, and are independent of (Y, C). The common C explicitly couples propagation and bypass sensitivity, making the second assay a readout of the background affecting the first.

5.2 COMPLETE SUMMARIES AND RESIDUAL AMBIGUITY

Set $T = YC$ and $X = S + C\kappa/2$. The entire named response pair is

$$M_A = e^{X - T\kappa/2} z, \qquad M_B = e^{X + T\kappa/2} z. \tag{7}$$

All eight P coordinates are therefore functions of (T, X, κ), including those that can reveal background. Symmetric z, the fixed mask and the fixed coefficients give $\mathrm{sign}(s_P) = T$, and the variance-product coordinate is an invertible transform of $e^{2X} \mathrm{Var}_0(z)$; at fixed κ complete P is therefore information-equivalent to (T, X) in exact arithmetic. Complete D is information-equivalent to $Q = \log(mq) = \nu C + \sigma_D Z_D$ with m matched third partners and q the mean absolute residual-profile difference, every D coordinate here being constant or a function of Q (Appendix Section S4). Two conditions fix the subfamily analyzed here: all matched third partners have zero null residuals, and P uses one significance mask across states. In it, direction is absent from the retained D vector but not from the raw measurements, since ordered uncompressed double profiles can reveal the upstream endpoint (S4.2).

The remaining ambiguity is quantified by the conditional distribution $X \mid C \sim N\big(C(\delta + \kappa/2), \sigma_P^2\big)$. The densities overlap at every finite signal-to-variability ratio, so complete P reveals background only partially and leaves a quantified orientation ambiguity (fig. 1b).

5.3 SHARP LIMITS OF ADDITIVE INTERPRETATION

Under (A1) below, whose information-equivalence clause S4 establishes for the retained interface at fixed κ on zero-null summaries, write R^*_P, R^*_{PD} and R^*_{add} for the population 0–1 risks of the best measurable rule on (T, X), on (T, X, D), and of the best threshold of a source-additive score $a(T, X) + b(D)$ with arbitrary measurable a, b. That class contains every independently scored late fusion and both matched additive controls of Section 4, and excludes only cross-channel interaction (Wörtwein et al., 2022).

Exchange sends (T, X, D) to $(-T, X, D)$, and $R^*_{\text{odd-add}}$ is the optimum over additive rules whose decision reverses under it. Put $\Delta = R^*_P - R^*_{PD}$, the classification gain joint access obtains over the complete first channel.

Theorem 1 (Sharp limits of additive interpretation)**.** *Assume (A1): T balanced, $T \perp (C, X, D)$, $Y = TC$, and P information-equivalent to (T, X). With $\Delta = R^*_P - R^*_{PD}$,*

$$R^*_{\text{odd-add}} = R^*_P, \qquad R^*_{\text{add}} \;\geq\; R^*_P - \tfrac{1}{2}\Delta. \tag{8}$$

If in addition $\operatorname{sign}_+[u(X) + v(D)]$ is a Bayes classifier for C given (X, D) for some measurable u, v (the common-threshold condition CT), the inequality is an equality, attained by an explicit score given in S4.4. Without CT it can be strict.

The whole obstruction is one step (fig. 1c): at a fixed value of X a threshold is monotone, so the sign of the difference between the two orientation decisions cannot depend on the context value. The context channel can move either decision, or both; what it cannot do is reverse their order. Attainment spends that freedom, running the Bayes rule for C on one orientation branch and the P-only Bayes rule on the other, which is why it must be exchange inconsistent whenever $\Delta > 0$. S4.4 gives both proofs, the tie convention, a strict finite-support example and the scope limits. *Context changes what directional evidence means, and the reading rule sets how much of that conditional information becomes predictive gain.*

The bound holds at every threshold of a fixed additive score, not only at its optimum, so the balanced-label ROC envelope caps additive AUROC as well; ranking never follows from a classification bound by itself and needs that separate step (Menon & Williamson, 2014). S4.3 evaluates the envelope at the channel-oracle corner $R^*_{PD} = 0$, where an overlap constant e between the two conditional densities of X gives $R^*_{\text{add}} \geq e/2$ and caps balanced-orientation AUROC at $1 - e^2/2$. That oracle form is weaker; it calibrates the noise scales of Section 5.4 and gives a secondary certificate in Section 6.1; theorem 1 needs no oracle and is never weaker on the actual channels (S4.3, S4.4, S4.6).

What is new here. At $R^*_{PD} = 0$ with uninformative X the bound reduces to the exclusive-or cell, where a threshold of a sum misclassifies one of four corners: that instance is classical (Minsky & Papert, 1969). Hessel & Lee (2020) study the same source-additive class, projecting a fixed model's scores onto it in L_2. Here the population 0–1 risk is optimized over that class against the target labels instead, and what is new is the resulting inequality for arbitrary X and D with Δ non-degenerate, exact attainment under CT, and the decision-level exchange statement.

For the correlated family, define standardized signals and conditional separation

$$\begin{aligned} &\bar{X} = aC + Z_P, \quad \bar{Q} = bC + Z_D, \quad a = \frac{\delta + \kappa/2}{\sigma_P}, \quad b = \frac{\nu}{\sigma_D}, \\ &s_\perp = \frac{b - \rho a}{\sqrt{1 - \rho^2}}, \qquad r = \sqrt{a^2 + s_\perp^2}. \end{aligned} \tag{9}$$

The residual $\bar{Q} - \rho\bar{X}$ has signal $b - \rho a$ and noise variance $1 - \rho^2$: it is what D contributes once the information and the noise already present in P are accounted for.

Corollary 1 (Conditional value and the exact additive optimum)**.** *For fixed κ under the stated mechanism and summary assumptions,*

$$\begin{aligned} R^*_P &= \Phi(-|a|), & R^*_{P,D} &= \Phi(-r), \\ \text{AUROC}^*_P &= \Phi(\sqrt{2}|a|), & \text{AUROC}^*_{P,D} &= \Phi(\sqrt{2}r). \end{aligned} \tag{10}$$

The background posterior mean is $\tanh(\alpha\bar{X} + \beta\bar{Q})$ *with* $\alpha = (a - \rho b)/(1 - \rho^2)$ *and* $\beta = (b - \rho a)/(1 - \rho^2)$. *CT therefore holds for every finite* a, b *and* $|\rho| < 1$*, whatever the signs of* β *and* ρ*, and*

$$R^*_{\mathrm{add}} = \frac{\Phi(-|a|) + \Phi(-r)}{2} \;\geq\; \frac{R^*_P}{2}, \qquad \mathrm{AUROC}^*_{\mathrm{add}} \leq 1 - e^2_\rho/2, \quad e_\rho = \Phi\left(-\frac{|a|}{\sqrt{1-\rho^2}}\right), \tag{11}$$

e_ρ *being that overlap constant recomputed once the shared noise is revealed to both scorers. For finite parameters,* $R^*_{PD} < R^*_{\mathrm{add}} < R^*_P$ *if and only if* $b \neq \rho a$.

Gaussian likelihood ratios give eq. (10) and the posterior; eq. (11) is then theorem 1 under CT, and its ranking statement is the S4.3 envelope with the exogenous D noise revealed (S5.1–S5.3). The criterion $b \neq \rho a$ is exact rather than a sufficient region, and Section 6.2 tests it against a design fixed before the corollary existed. S5 evaluates the parameter-free form $R^*_{\mathrm{add}} \geq R^*_P/2$ on $|a| \leq 0.5$, $b \geq 2$, $|\rho| \leq 0.5$ and gives the continuous-elasticity extension (S5.6).

Fixing every parameter but ρ leaves the laws of (P, Y) and (D, Y) unchanged while $R^*_{P,D}$ follows eq. (10) and, under CT, the additive optimum tracks their midpoint: identical channel-label marginals do not determine joint risk (S8.4). PaD represents these conditional interpretations through its joint term while retaining independent components.

5.4 Observation perturbations and usable task margins

Observation condition O. Require a joint realization $\widetilde{P} = G_P(P, \xi_P)$, $\widetilde{D} = G_D(D, \xi_D)$, where (ξ_P, ξ_D) is independent of the original mechanism and its two components may be correlated, each map reading only its own ideal channel and external noise. Separate marginal factorizations are insufficient, since they could conceal label-dependent cross-channel coupling (S8).

Theorem 2 (Observation loss spends the joint margin)**.** *Under condition O and the fixed-elasticity mechanism, observed additive scores still satisfy* $R^*_{\mathrm{add,obs}} \geq R^*_{\mathrm{add}}$ *of eq.* (11) *and* $\mathrm{AUROC}^*_{\mathrm{add,obs}} \leq 1 - e^2_\rho/2$. *Decoding the two ideal coordinates from the observed channels with symbol-error budgets* ϵ_T, ϵ_Q *and a magnitude budget* $\zeta_y(t)$ *leaves an explicit margin over the joint witness, and a corresponding AUROC bound (S8.1).*

Conditioning on external noise preserves the ideal additive class, so composing a, b with the two observation maps gives an ideal additive score at unchanged joint law; averaging its classification bound over the noise and applying the balanced-label ROC envelope preserves that envelope. S8 proves both for arbitrarily correlated decoding errors. It states the margin against the oracle floor $e_\rho/2$; the certificate of Section 6.1 uses eq. (11) instead. On $|a| \leq 0.5$, $b \geq 2$, $|\rho| \leq 0.5$, $\epsilon_T + \epsilon_Q \leq 0.05$ guarantees an error advantage of 0.068176 and, at $t = 0.1$ with $\zeta_+(t) + \zeta_-(t) \leq 0.01$, an AUROC advantage of 0.026115, both calibrated at the raw-noise scales of S9.

6 Experiments: from task limits to learned behavior

6.1 PaD achieves risk below the additive lower bound

All nine B1 PaD classifiers achieve certified risk below the limit for arbitrary nonlinear additive scoring on independent test draws from the simulated mechanism of Section 5, six of them trained with noise added to individual simulated measurements (fig. 2a). The measured panels are considered separately in Section 6.3.

The simulations use six prespecified mechanisms with the full encoders and the fixed rule of Section 4; each of three seeds supplies 512 training and 4,096 independent test instances (procedures, grids and random streams: S9).

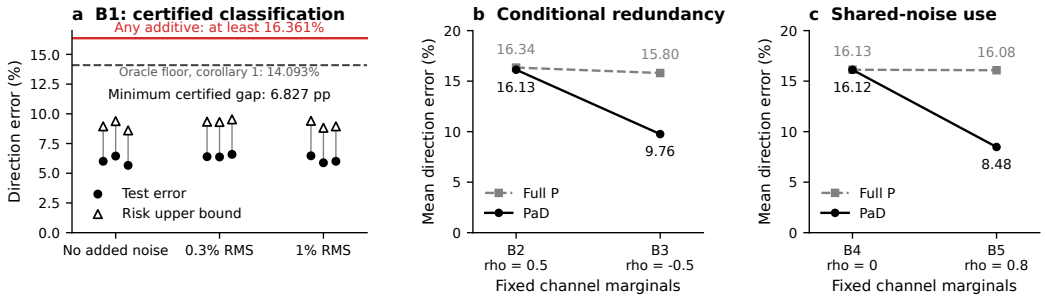

Figure 2: From an additive task limit to a learned advantage. All panels are simulation. (a): dots show B1 test errors and open triangles their simultaneous risk upper bounds. Segments show the confidence correction, not seed variability. Each of the nine fits uses 4,096 independent test draws, and every upper bound lies below both additive risk lower bounds. (b) and (c): three-seed mean errors for full P and joint learning, channel marginals fixed within each pair while the noise correlation changes.

Certification reserves familywise $\alpha = 0.05$ for $M = 30$ joint fits. For B1, eq. (11) gives the additive floor $L_{\mathrm{sharp}} = [\Phi(-0.5) + \Phi(-2.081666)]/2 = 0.1636120164$. The correction is $c_R = \sqrt{\log(1200)/8192} = 0.0294191756$, and all nine fits pass (S9). Their worst risk upper bound satisfies

$$\max U = 270/4096 + c_R = 0.0953371444 < L_{\mathrm{sharp}},$$

leaving a minimum certified separation of 6.827487 percentage points (fig. 2a). On the same nine fits and simultaneous confidence event, the weaker oracle floor $e_\rho/2 = 0.1409257154$ yields 4.558857 points (S4.3, S9). The familywise allocation remains fixed at 30, including the B3/B5 fits, whose unrestricted-additive certificates do not pass.

The fitted-control comparison in ideal B1 is larger, at 24.9268 percentage points. It answers a different question from the 6.827487-point certificate against the unrestricted additive class (S9). In this encoding, $b^o \equiv 0$ and $g^o \equiv 0$ in eq. (5), so the matched control uses only P. The B1 and B5 fitted contrasts therefore compare multiplicative modulation with a pure-P baseline, while theorem 1 bounds the broader additive class.

A pre-registered interface ablation. The additive obstruction is a property of the specified generator and its retained interface (S12.1): enlarging D with three exchange-odd third-partner descriptors removes it inside the generator, where A_{match} mean error falls from 30.615% on D to zero on D^+ over 12,288 independent test instances (S12.2, Table S26), whereas on Jonikas those descriptors are near-random or reversed and every fitted learner degrades (S12.2–S12.3).

Independent calibration selected raw coordinate noise at 0.3% and 1% of fixed RMS scales. At these levels, theorem 2 guarantees gains of at least 0.110275 and 0.093094 on separate confidence families (S9), and classification superiority is certified for the fitted learners. The nine B1 AUROCs range from 0.984770 to 0.988762, but their test-corrected lower bounds remain below the 0.960280 additive ceiling; ranking is therefore not certified (S8–S9).

The same channel marginals, different joint value. B2/B3 and B4/B5 preserve each channel's population marginals while changing the conditional relation, so within each pair full-P performance stays similar while the joint advantage moves (fig. 2b,c). The pairs are the design's own control on attribution: a gain that followed either marginal would have to move with it, and it does not.

6.2 The criterion recovers the prespecified design

The criterion $b \neq \rho a$ matches the gain/no-gain labels assigned to B1, B2, B4 and B5 before corollary 1 was derived. The residual signal $b - \rho a$ vanishes at the two declared null settings, B2 $(1, 0.5, 0.5)$ and B4 $(1, 0, 0)$, and at no other setting.

At the two settings declared to permit a joint gain, B1 $(0.5, 2, 0.5)$ and B5 $(1, 0, 0.8)$, Δ is 0.289851 and 0.110865. The contrast settings B3 $(1, 0.5, -0.5)$ and B6 $(4, 2, 0)$ give 0.095340 and 0.000028. Fitted behavior follows this partition: B2 and B4 show no appreciable gain, while B1 and B5 recover 85.6% and 66.6% of the respective 28.99- and 11.09-point gains available over the ideal P-only rule. These fractions use the P-only Bayes risk, rather than the additive limit, as their reference (fig. 2b,c; S5.3, S12.4).

6.3 Measured comparisons on two yeast panels

Costanzo supplies growth-fitness evidence for 657 expression-response orientations; Jonikas supplies 168 KEGG records over 84 unordered pairs, with 21 positives and 147 reference-negatives (Battle et al., 2010). Both are development panels sharing an expression source (S1–S3, S6). External baselines, RBF SVM, XGBoost and CatBoost (Cortes & Vapnik, 1995; Chen & Guestrin, 2016; Prokhorenkova et al., 2018), receive the same information under endpoint-disjoint folds and inner-only selection; PaD exceeds every prespecified external run on both panels (S10).

Table 1: Nonlinear controls on the same retained inputs. Bold marks the best point estimate. J has five candidates versus PaD's 35; it tests mixture attribution, not matched search capacity. Costanzo measures orientation agreement; Jonikas reports AUROC of F and d. Single-channel results and paired comparisons are in S13.

Method	Costanzo correct / 657	Jonikas F	Jonikas d
PaD	611	**0.881114**	0.844833
Pure joint RBF J	607	0.875607	**0.851312**
Matched nonlinear additive A_{match}	**613**	0.827664	0.714286
Pure additive RBF A_{rbf}	606	0.810982	0.678167

AUROC of d ranks panel labels and is distinct from direction-sign accuracy on annotated relations; $F = d + u$ does not decompose AUROC additively. On Costanzo, PaD and A_{match} agree on all but four pairs; their paired correct-count interval $[-6, +2]$ lies within the pre-declared ± 13-pair tolerance. On Jonikas, neither additive control exceeds the stronger single channel, D (S13). PaD's F increment over D is 0.046647, with descriptive interval $[-0.0157, +0.1116]$. Rescored on d, predefined but not pre-registered, PaD exceeds A_{match} by 0.130547, with descriptive pair-resampling interval $[0.0465, 0.2218]$ (S13.5).

PaD and J have nearby point estimates: PaD minus J is $+0.005507$ on F and -0.006479 on d. Fixed-prediction paired resampling gives 95% descriptive intervals $[-0.0211, +0.0328]$ and $[-0.0382, +0.0243]$, respectively; endpoint-weight intervals also span zero (S13.6). Together with the additive controls, these results support joint access; the mixture-specific comparison is inconclusive in either direction.

6.4 Which evidence changes which decision?

All source comparisons retain full P and use J. The D blocks are single-perturbation phenotypes S, exchange-invariant measurement metadata M, no-interaction expectation H, focal double-response coordinates V, and third-partner context B; FULL retains all five. A protocol fixed after the initial source exploration specifies SH→SMH as the primary M contrast and FULL−V→FULL as the primary V contrast; both are tested on the same records (S13.6–S13.7).

Table 2: Source contributions depend on retained context. Each row adds the indicated block, refitting the same J procedure. Counts concern 21 annotated relations, not independent replicates; FULL−V is already at ceiling.

Contrast	Δd-AUROC	ΔF-AUROC	Correct / 21
SH → SMH (primary M)	+0.180110	+0.134435	$18 \rightarrow 21$
FULL−M → FULL	+0.017493	-0.000972	$20 \rightarrow 21$
FULL−V → FULL (primary V)	+0.021380	+0.037253	$21 \rightarrow 21$

The primary M d increment persists under resampling of the evaluation records by gene endpoint or by unordered pair, with predictions held fixed: intervals $[0.1239, 0.2390]$ and $[0.0881, 0.2806]$. Both primary source increments exceed all 100 corresponding pair-orbit block mismatches, each refitted from scratch. These controls test source–pair correspondence, not conditional independence, and the M control is defined for the primary contrast only.

With the other sources present, M's d increment shrinks and its F increment is near zero (table 2); the full-context M d interval under pair resampling is $[-0.0046, 0.0408]$. Thus its large primary increment does not transfer unchanged to the complete input set.

V's observed F increment has intervals $[-0.0075, 0.0903]$ under endpoint weighting and $[-0.0184, 0.1001]$ under pair resampling; removing DIE2 reverses its sign. Its unchanged 21/21 direction count is a ceiling: without B, SMH→SMHV instead changes 21/21 to 20/21, a difference set by the retained context.

Whether M and V serve different tasks remains unresolved: pair-resampling intervals for the combined selectivity contrast and full-context M's $\Delta d - \Delta F$ are $[-0.0118, 0.0830]$ and $[-0.0019, 0.0395]$ (S13.7). The evidence locates context-dependent predictive contributions.

7 Discussion and conclusion

Independent scoring can lose conditional information even when each channel is modeled nonlinearly. PaD connects that restriction to an achieved, noise-tolerant risk separation in simulation. The measured panels locate the empirical benefit in joint access, while source-level analyses show why aggregate improvement is not enough to explain the value of an assay.

The sharp theorem concerns a shared-background, invariant-D subfamily, whereas both measured panels retain odd D coordinates. An invariant M block does not make the complete D invariant. The real-data evidence is development evidence: the 21 annotated relations share genes, and the training history of the Costanzo fingerprint encoder is not fully reconstructed (S6). The source analyses concern predictive use, not biological mechanism.

For researchers integrating perturbation screens, measurement records should not be discarded merely because they lack a directional sign. Their contribution must be evaluated against the other retained inputs and the intended decision. Relation ranking, d-AUROC and annotated direction signs answer different questions; none substitutes for the others. Perturbation evidence should be evaluated by the decisions it changes, with each contribution attached to its retained context.

REPRODUCIBILITY STATEMENT

The experiment repository is available for review through an anonymous mirror at `anonymous.4open.science/r/PaD-6F27`, which exposes neither commit history nor a hosting account. The reusable `pad-causal` toolkit (version 0.1.1) applies the same learning rule to prepared perturbation evidence through Python and command-line interfaces, with model serialization. Its built wheel is bundled in that repository as `tools/pad_causal-0.1.1-py3-none-any.whl` and installs with `python -m pip install ./tools/pad_causal-0.1.1-py3-none-any.whl`. Appendix S11 distinguishes experiment reproduction from toolkit use and provides the corresponding commands.

The supplementary package supplies proofs, feature definitions, frozen inputs and predictions, figure sources, and deterministic numerical checks (S1–S9). Its source-analysis code refits J and the source ablations, reproduces block mismatches, and regenerates all fixed-prediction intervals and leave-one-gene sensitivities (S13.6–S13.7). Verification and refitting commands are identified separately. The authors checked the proofs line by line and take responsibility for the research decisions, analyses and final text. The verifier checks numerical claims and theory-source preservation, not proof correctness by itself. PDF and Markdown are generated from the same manuscript sources.

ETHICS STATEMENT

This work analyses previously published yeast perturbation datasets and simulated data. It involves no human subjects, no personally identifiable information and no new organism experiments. The measured panels are used under the terms of their original releases and are cited at first use. The method predicts pathway order from assay evidence and makes no clinical or diagnostic claim; the paper states where its predictive results do and do not license mechanistic interpretation.

AI USE STATEMENT

AI tools assisted with language polishing, data verification, and formatting checks.

REFERENCES

Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5):052802, 2015. doi: 10.1103/PhysRevE.91.052802. URL https://doi.org/10.1103/PhysRevE.91.052802.

Alexis Battle, Martin C. Jonikas, Peter Walter, Jonathan S. Weissman, and Daphne Koller. Automated identification of pathways from quantitative genetic interaction data. *Molecular Systems Biology*, 6:379, 2010. doi: 10.1038/msb.2010.27. URL https://doi.org/10.1038/msb.2010.27.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161. URL https://doi.org/10.3390/e16042161.

C. F. Blum, N. Heramvand, A. S. Khonsari, and M. Kollmann. Experimental noise cutoff boosts inferability of transcriptional networks in large-scale gene-deletion studies. *Nature Communications*, 9:133, 2018. doi: 10.1038/s41467-017-02489-x. URL https://doi.org/10.1038/s41467-017-02489-x.

Brielin C. Brown, Alex Tokolyi, John A. Morris, Tuuli Lappalainen, and David A. Knowles. Large-scale causal discovery using interventional data sheds light on gene network structure in K562 cells. *Nature Communications*, 16:9628, 2025. doi: 10.1038/s41467-025-64353-7. URL https://doi.org/10.1038/s41467-025-64353-7.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.

785–794. ACM, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Michael Costanzo, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420, 2016. doi: 10.1126/science.aaf1420. URL https://doi.org/10.1126/science.aaf1420.

José A. R. Fonollosa. Conditional distribution variability measures for causality detection. In Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu (eds.), *Cause Effect Pairs in Machine Learning*, The Springer Series on Challenges in Machine Learning, pp. 339–347. Springer, Cham, 2019. doi: 10.1007/978-3-030-21810-2_12. URL https://doi.org/10.1007/978-3-030-21810-2_12.

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

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(4):111–152, 2013. URL https://jmlr.org/papers/v14/hyvarinen13a.html.

Martin C. Jonikas, Sean R. Collins, Vladimir Denic, Eugene Oh, Erin M. Quan, Volker Schmid, Jimena Weibezahn, Blanche Schwappach, Peter Walter, Jonathan S. Weissman, and Maya Schuldiner. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, 323(5922):1693–1697, 2009. doi: 10.1126/science.1167983. URL https://doi.org/10.1126/science.1167983.

P. Kemmeren, K. Sameith, L. A. L. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014. doi: 10.1016/j.cell.2014.02.054. URL https://doi.org/10.1016/j.cell.2014.02.054.

Chohee Kim, Mihaela van der Schaar, and Changhee Lee. Discovering features with synergistic interactions in multiple views. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24562–24583. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24ag.html.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1452–1461. PMLR, 2015. URL https://proceedings.mlr.press/v37/lopez-paz15.html.

Florian Markowetz, Dennis Kostka, Olga G. Troyanskaya, and Rainer Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–i312, 2007. doi: 10.1093/bioinformatics/btm178. URL https://doi.org/10.1093/bioinformatics/btm178.

Aditya Krishna Menon and Robert C. Williamson. Bayes-optimal scorers for bipartite ranking. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 68–106. PMLR, 2014. URL https://proceedings.mlr.press/v35/menon14.html.

Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969. URL https://mitpress.mit.edu/9780262534772/perceptrons/.

Hilary Phenix, Katy Morin, Cory Batenchuk, Jacob Parker, Vida Abedi, Liu Yang, Lioudmila Tepliakova, Theodore J. Perkins, and Mads Kærn. Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Computational Biology*, 7(5):e1002048, 2011. doi: 10.1371/journal.pcbi.1002048. URL https://doi.org/10.1371/journal.pcbi.1002048.

Hilary Phenix, Theodore Perkins, and Mads Kærn. Identifiability and inference of pathway motifs by epistasis analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(2):025103, 2013. doi: 10.1063/1.4807483. URL https://doi.org/10.1063/1.4807483.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Michael A. Savageau. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 25(3):365–369, 1969. doi: 10.1016/S0022-5193(69)80026-3. URL https://doi.org/10.1016/S0022-5193(69)80026-3.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(57):1531–1565, 2006. URL https://www.jmlr.org/papers/v7/sonnenburg06a.html.

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. InfMasking: Unleashing synergistic information by contrastive multimodal interactions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/119809.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. Technical Report arXiv:1004.2515, arXiv, 2010. URL https://doi.org/10.48550/arXiv.1004.2515.

Torsten Wörtwein, Lisa Sheeber, Nicholas Allen, Jeffrey Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4681–4696. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.344. URL https://aclanthology.org/2022.findings-emnlp.344/.

Marinka Žitnik and Blaž Zupan. Gene network inference by probabilistic scoring of relationships from a factorized model of interactions. *Bioinformatics*, 30(12):i246–i254, 2014. doi: 10.1093/bioinformatics/btu287. URL https://doi.org/10.1093/bioinformatics/btu287.