FOUR POINTS OR NOTHING: WHAT A BATCH CAN AND CANNOT MEASURE AT ITS OWN LEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Campaigns in protein engineering and molecular design repeatedly ask a local question at the current best candidate: do these two changes interact here? We show that a batch can be full rank for the second-order interaction model and still contain no assumption-free measurement of a single one of the local interactions at its own leads: global identifiability does not imply local measurability. The estimand is the four-point contrast of a double-mutant cycle, so a model-free *linear* estimator of it, unbiased under every response function, exists if and only if all four points have been observed, and is then unique in that class. Auditing twelve batch rules on four real near-complete landscapes shows this is not an artefact of classical criteria: determinantal, gradient-embedding and redundancy-aware information batches reach the same near-zero certification as Bayesian D-optimality by three different routes, while naive exploitation certifies the most by accident. Three structural causes explain it, two computable from the proposed batch alone: separation to minimum Hamming distance two, occupancy below a classical extremal threshold, and rectangles away from the leads. An exact bias identity then gives a parameter-free boundary between direct measurement and the *exact* second-order projection, which a fitted model tracks but does not obey. It holds on 93 of 93 coordinate pairs, and a factorial sweep crossing it in both arguments puts the oracle crossing where the identity does.

1 INTRODUCTION

Many scientific campaigns are a loop: choose a batch from a combinatorial space, measure it, fit a model, choose again. Directed evolution (Wittmann et al., 2021; Yang et al., 2025), lead optimization (Kramer, 2019) and perturbation screening (Horlbeck et al., 2018; Xu et al., 2024) all have this shape, and machine learning supplies the batch step with a rich toolkit: upper-confidence bounds, diversity and determinantal batches (Kathuria et al., 2016; Bıyık et al., 2019), gradient- and Fisher-embedding batches (Ash et al., 2020; 2021) and redundancy penalties (Kirsch et al., 2019). Practitioners ask a question these rules do not target: at the current best candidate, do two changes interact *here*? Biophysics answers it with a double-mutant cycle (Carter et al., 1984; Horovitz & Fersht, 1990), four variants and an alternating sum; medicinal chemistry calls the same four-molecule object a matched molecular square (Kramer et al., 2015; Kramer, 2019; Gogishvili et al., 2021); genetics uses the same contrast to define a genetic interaction (Mani et al., 2008).

This paper asks whether that toolkit buys those four points. It does not, and the failure inverts the usual intuition (Figure 1): the more carefully a rule is engineered to be statistically efficient, the less often it can measure a local interaction. On the one landscape where the second-order interaction block is identifiable outright, five of our twelve rules identify *all* of it and certify *none* of the local interactions at their own leads. The reason is structural: a rectangle holds two points agreeing on every coordinate but one, deliberate near-duplication that diversity objectives penalise and a determinant has no reason to prefer, while naive exploitation concentrates on a near-subcube and closes rectangles by accident.

A batch *certifies* a local interaction when all four points it is a function of have been observed, so it can be read off without fitting anything; certification says the question is answerable, not that the answer is large, and we report it for the batch alone and with the campaign's history. The estimand is not a matter of taste either: local and background-averaged interactions can disagree in sign, and

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.

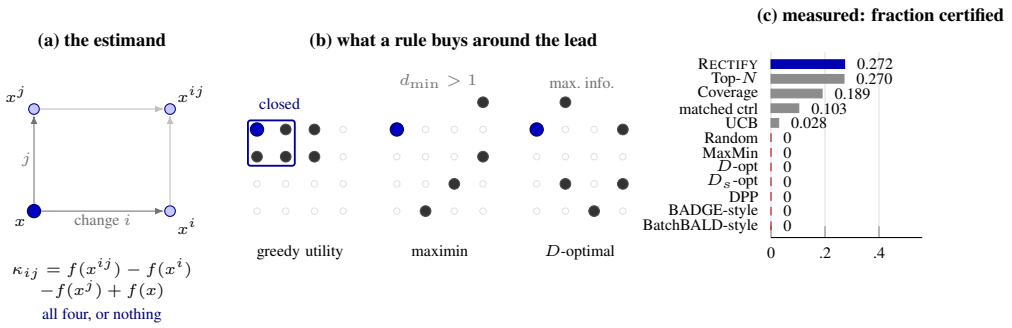


Figure 1: **The estimand fixes the design.** *(a)* the local interaction at a lead x cancels every additive effect, and a model-free *linear* estimator of it exists only if all four points have been observed (Proposition 1). *(b)* utility-greedy selection closes rectangles by accident, while maximin and a D-optimal design on the full second-order model spread and close none. *(c)* the consequence on Poelwijk (Section 5.1).

a large literature documents that epistasis is idiosyncratic to the background (Poelwijk et al., 2016; Sailer & Harms, 2017; Johnson et al., 2019; Lyons et al., 2020; Bakerlee et al., 2022; Park et al., 2022).

Contributions.

- **A measurable paradox: global identifiability and local measurability come apart.** Where Birkes et al. (1976) characterise which contrasts an incomplete layout can estimate, we fix the estimand first and ask what a *chosen batch* buys. The two properties are stated under different assumptions, which is the point rather than a sleight of hand: rank is rank of the second-order model, certification asks for unbiasedness under *every* response. A model-free linear unbiased estimator of κ_{ij} exists iff all four corners have been observed, is then unique in that class with variance $4\sigma^2$ whatever the higher-order structure, and none is cheaper (Lemma 1 and Proposition 1). The twelve rules separate with no overlap: those certifying anything sit at rank share at most 0.500, those certifying nothing at 0.731 or above (Table 13).
- **The failure has computable geometry, on the rules this field actually builds.** Where Fontana et al. (2022) enumerate the circuits of the additive model, we measure what determinantal, gradient-embedding and redundancy-aware information batches buy against that standard alongside classical criteria. Separation to $d_{\min} \geq 2$ forecloses the measurement (Proposition 3); occupancy above an extremal threshold *forces* a rectangle, while below it, under uniform occupancy, the expected count falls quartically (Proposition 2); and a batch can build rectangles in quantity and place none at its leads. The first two are computable before any assay, and with K held fixed it is occupancy against the threshold, not the level count, that predicts certification (Section 5.2).
- **An exact bias identity, and the oracle boundary it defines.** Where Shyamal et al. (2025) design for a global bounded-degree model, we price modelling at a named background: the squared bias of the *exact* second-order projection at a uniformly drawn rectangle is $4c_ic_j$ times the higher-order variance involving both toggled coordinates (Proposition 4), so that projection beats the contrast iff $V_{\geq 3}^{(ij)} < \sigma^2(1 - 1/m_i)(1 - 1/m_j)$ (Corollary 1). A fitted model pays estimation variance and shrinks, making the boundary necessary and not sufficient for it; we test the identity, the boundary and the fitted comparison separately and report where each fails (Section 5.2).

2 Related work

The estimand, the four-point contrast and the estimability theory are classical. The delta is one row (Table 16): no previous work asks what an *automated batch rule* buys against a named local estimand, nor gives a certificate computable on a proposed batch before measurement.

Optimal design, estimability and the four-point move. Designing for a parameter subset while treating the rest as nuisance is D_s-optimality (Atkinson & Cox, 1974; Pukelsheim, 2006); that the

design follows from the estimand is c-optimality (Elfving, 1952), and Proposition 1 is its degenerate case for a local contrast, where the optimal design is the support of the functional. Local contrast versus regression coefficient is likewise established (Egami & Imai, 2019; Dasgupta et al., 2015), as is estimability in incomplete layouts: connectedness (Eccleston & Hedayat, 1974), estimable interaction contrasts spanned by *loop* contrasts of dimension the cycle rank (Birkes et al., 1976), resolution theory on unconfounded two-factor interactions (Fries & Hunter, 1980; Wu & Wu, 2002; Cheng et al., 1999). The four-point $(+1, -1, -1, +1)$ move is the basic Markov-basis move of the two-way no-interaction model (Diaconis & Sturmfels, 1998), minimal-support kernel elements are *circuits* (Fontana et al., 2022), over a product space these are the 2×2 minors of a tensor's flattenings (Bernardi et al., 2018; Landsberg, 2011; Hara et al., 2010; Aoki & Takemura, 2005), which is why Lemma 1's bipartition ranges over blocks, and Proposition 3(ii) is the mixed-alphabet Singleton bound (Singleton, 1964). We claim the audit, the diagnostics and Proposition 4, not this theory.

Batch acquisition. Diversity and determinantal batches (Kathuria et al., 2016; Bıyık et al., 2019), gradient- and Fisher-embedding batches (Ash et al., 2020; 2021) and redundancy-aware ones (Kirsch et al., 2019) maximise a global criterion; we audit one of each. Hernández-Lobato et al. (2014) and Neiswanger et al. (2021) acquire to learn a named property rather than an optimum, Xie et al. (2026) trade regret against statistical power, closest to us Shyamal et al. (2025) build randomised factorial designs for combinatorial interventions under a budget but target a global bounded-degree model, and Jain et al. (2024); Xu et al. (2024); Zhu et al. (2026); Soper et al. (2025) actively choose perturbation pairs. None targets the identifiability of a named local contrast.

Interactions in combinatorial biology and chemistry. Double-mutant cycles (Carter et al., 1984; Horovitz & Fersht, 1990; Horovitz, 1996; Horovitz et al., 2019), non-additivity analysis (Kramer et al., 2015; Kramer, 2019; Gogishvili et al., 2021; Guasch et al., 2024) and genetic-interaction scores (Mani et al., 2008; Costanzo et al., 2016) all compute this four-term contrast, and a cycle is *constructed* rather than mined, so four-point design is the oldest idea here; free-energy-perturbation networks even buy redundant cycles for closure checks (Wang et al., 2015). What has not been done is to ask what an *automated* rule buys against that standard, which combinatorially complete landscapes (Poelwijk et al., 2019; Wu et al., 2016; Johnston et al., 2024; Lite et al., 2020) make it possible to simulate; higher-order structure on them is quantified by Brookes et al. (2022), the local-versus-averaged distinction by Weinreich et al. (2013), and directed-evolution training sets are designed for fitness, not identifiability (Wittmann et al., 2021; Yang et al., 2025; Tran et al., 2026). On the modelling side Lengerich et al. (2020); Hooker (2007) fix the additive-versus-interaction convention, the pairwise interaction index is *defined* as a four-point second difference (Grabisch & Roubens, 1999), query complexity for sparse spectra is studied by Amrollahi et al. (2019); Kang et al. (2024), and Kuskova et al. (2026) give a pre-fit rank diagnostic.

3 SETUP: TWO INTERACTIONS, NOT ONE

Space, batch, pool. The design space is a product $\mathcal{X} = B_1 \times \cdots \times B_K$ with $m_k = |B_k|$ levels at coordinate k, and measuring a point returns $y(x) = f(x) + \varepsilon_x$ with $\mathbb{E}[\varepsilon_x] = 0$, $\mathrm{Var}(\varepsilon_x) = \sigma^2$ and ε_x independent across distinct measured points; every statement below uses this model. A campaign may assay only points of a *pool* $\mathcal{P} \subseteq \mathcal{X}$, all of $\mathcal{X}$ under site saturation or a catalogue of what can be bought, and a *batch* is a set $L \subseteq \mathcal{P}$ of N distinct points chosen by an acquisition rule from a surrogate. For a level $b'_i \neq x_i$ write x^i for the point equal to x except at i and x^{ij} for the point differing at both; $\{x, x^i, x^j, x^{ij}\}$ is a *rectangle* $\mathsf{rect}_{ij}(x; b'_i, b'_j)$.

Definition 1 (Local interaction)**.** The local interaction of changes i and j at the background x is the mixed second difference

$$\kappa_{ij}(x; b'_i, b'_j) \;=\; f(x^{ij}) - f(x^i) - f(x^j) + f(x). \tag{1}$$

Its sign says whether the pair is super- or sub-additive *at this background* but does not by itself order the four variants. With $\Delta_i = f(x^i) - f(x)$ and $\Delta_j = f(x^j) - f(x)$ the single-change effects, $f(x^{ij}) - f(x^i) = \Delta_j + \kappa_{ij}$ and $f(x^{ij}) - f(x^j) = \Delta_i + \kappa_{ij}$, so the double change beats both singles exactly when $\kappa_{ij} > \max\{-\Delta_i, -\Delta_j\}$ and beats the lead when $\kappa_{ij} > -(\Delta_i + \Delta_j)$: a campaign needs κ_{ij} *with* both single-change effects, and all four come off the same four measurements, so the rectangle supplies the whole decision, not only its non-additive part.

Second-order functional ANOVA instead defines a *background-averaged* θ_{ij} by projecting f onto the additive and pairwise subspaces (Lengerich et al., 2020); the two coincide only when f has no structure of order three or higher, and the applied literature is explicit that the gap is large in practice (Poelwijk et al., 2016; Sailer & Harms, 2017; Johnson et al., 2019; Lyons et al., 2020; Bakerlee et al., 2022; Park et al., 2022). Both matter; only κ_{ij} answers a question asked at a named background.

Additive nuisance, and the design matrices. With $\mathcal{A} = \{a \mid a(x) = a_0 + \sum_k a_k(x_k)\}$ the additive functions, $c \in \mathbb{R}^L$ is an *additive-free contrast* if $\sum_{x \in L} c_x a(x) = 0$ for every $a \in \mathcal{A}$, so $c^\top y$ estimates something independent of every main effect. With A_L, Z_L the additive and pairwise-cell model matrices on the batch and $G_L = (I - P_{A_L})Z_L$, the *interaction information* $\mathcal{I}(L) = \frac{1}{2}\log\det(I + \sigma^{-2} G_L^\top G_L)$ is the Bayesian D_s-criterion an information-optimal design maximises.

4 What a batch can and cannot certify

Assumptions. Throughout: the noise model of Section 3; batches of *distinct* design points; and, in Proposition 1, estimators *linear* in y and unbiased for κ_{ij} under *every* $f : \mathcal{X} \to \mathbb{R}$. Replication does not change the estimand, which is a function of four *distinct* cells however many measurements each receives, though with replicates a generic additive-free contrast can be supported on two cells. Proofs are in Section C; Section F says which statements are additionally checked by exhaustive computation.

Two classical facts set the floor (Section B). No nonzero additive-free contrast on $\mathcal{X}$ has support of size one, two or three, and every support-four contrast has weights $(+1, -1, -1, +1)$ on $\{x, x^{S_1}, x^{S_2}, x^{S_1 \cup S_2}\}$ for a nontrivial bipartition (S_1, S_2) of the coordinates on which its four points differ, the rectangle of Definition 1 being $S_1 = \{i\}, S_2 = \{j\}$ (Lemma 1: the degree-two part of the Segre ideal described by the 2×2 minors of a tensor's flattenings (Bernardi et al., 2018; Landsberg, 2011), transported to the kernel of the additive design matrix by Diaconis & Sturmfels (1998), whose support-four solutions are circuits (Fontana et al., 2022; Aoki & Takemura, 2005)). Four is not a convenient number but the floor. The bipartition ranges over coordinate *blocks*, so for $K \geq 3$ two-coordinate rectangles are a minority of the minimal measurements, only 24% of the 100 on 2^4; a block rectangle has all six pairwise distances at least two, so Proposition 3(i) below does not exclude it, and Section 5.1 counts how many our batches hold (Remark 1).

Proposition 1 (The local interaction is estimable exactly when its rectangle is measured)**.** *Fix a rectangle* $\mathsf{rect}_{ij}(x; b_i', b_j')$*, let* κ *be its local interaction* (1)*, and call an estimator* model-free *if it is unbiased for* κ *under* every *response* $f : \mathcal{X} \to \mathbb{R}$*.*

1. *A model-free linear estimator of* κ *exists from a batch* L *if and only if all four corners lie in* L*.*
2. *In that case the contrast* $\hat{\kappa} = y(x^{ij}) - y(x^i) - y(x^j) + y(x)$ *is the* only *model-free linear estimator, so it is trivially the minimum-variance one among them, with* $\mathrm{Var}(\hat{\kappa}) = 4\sigma^2$ *regardless of* f*.*
3. $\hat{\kappa}$ *is exactly unbiased for* κ *under every* f*, whatever its additive part and whatever its structure of order three or higher.*

The quantifier does the work, and without it the proposition is false: under a second-order model κ_{ij} does not depend on the background, so any full-rank design identifies it, a rectangle-free resolution-V fraction included. A rule certifying nothing therefore buys no assumption-free measurement, which is not the same as leaving its own model unidentified.

A batch without the rectangle does not get a weaker version of the same quantity. At a fixed pair (i, j) and background g the additive-free contrasts available are exactly the cycle space of the bipartite incidence graph on $B_i \sqcup B_j$ (Birkes et al., 1976), and a contrast on a cycle of length 2ℓ has variance $2\ell\sigma^2$ and loads the alternating sum of the cells it visits, a single local interaction only for $\ell = 2$ (Proposition 5). A rectangle-free batch need not have any such contrast at all, and when it does it spans at least six cells, is noisier in proportion, and pools several local interactions (Section E).

Proposition 2 (When a rectangle appears by accident)**.** *Fix* (i, j) *and a background* g*, and let* n_g *be the number of batch points with that background. If* n_g *exceeds the Zarankiewicz number* $z(m_i, m_j; 2, 2)$*, the largest number of filled cells an* $m_i \times m_j$ *array can have with no filled* 2×2 *subarray, then some rectangle at* g *is complete. Conversely there exist configurations of*

$z(m_i, m_j; 2, 2)$ *points at* g *with no complete rectangle. Reiman's bound gives* $z(m, n; 2, 2) \le \min\{\bar{z}(m, n), \bar{z}(n, m)\}$ *with* $\bar{z}(m, n) = \frac{1}{2}\big[m + \sqrt{m^2 + 4mn(n-1)}\big]$*, the two orientations differing when* $m \neq n$*, so* $z(m, m; 2, 2) = \Theta(m^{3/2})$*, the matching constructions being classical (Kővári et al., 1954; Reiman, 1958; Erdős et al., 1966; Brown, 1966; Füredi & Simonovits, 2013).*

The extremal problem is Zarankiewicz's, the bound Reiman's, the asymptotics Kővári et al. (1954); Table 14 reports per rule how often a batch exceeds the threshold. Above it a rectangle is *forced*; below it the proposition says only that rectangle-free configurations exist, not that rectangles are unlikely, and the probability statement needs an occupancy model. With n_g cells occupied uniformly at random in the $m \times m$ array, the expected number of filled 2×2 subarrays is exactly $\binom{m}{2}^2 \binom{n_g}{4} / \binom{m^2}{4} = \Theta(n_g^4/m^4)$, quartic in occupancy and quartic-inverse in levels. A batch spreads N points over $\prod_{k \notin \{i,j\}} m_k$ backgrounds, so accidental rectangles are cheap for binary coordinates and very expensive for the twenty-valued protein case (Section 5.2).

Proposition 3 (Separation forecloses two-coordinate certification, up to a capacity)**.** *A two-coordinate rectangle contains two points at Hamming distance one. Hence (i) if the minimum pairwise Hamming distance of a batch satisfies* $d_{\min}(L) \ge 2$*, then* L *contains no two-coordinate rectangle and, by Proposition 1, certifies no local interaction of Definition 1 at all, whatever criterion produced it. (ii) Separation has a hard capacity:* $d_{\min}(L) \ge 2$ *forces* $N \le \min_k \prod_{j \neq k} m_j = \prod_j m_j / \max_k m_k$*, so no larger batch can be separated. Part (ii) is the mixed-alphabet Singleton bound at distance two (Singleton, 1964), stated here for design spaces and tight for every level vector (Section C).*

Proposition 3(i) is the most useful line in the paper, because $d_{\min}$ is computable from a proposed batch before a single measurement: it turns the audit into a prediction. Maximin separates by construction and D-optimality does so unasked, two points at Hamming distance one having nearly collinear rows in $[A, Z]$, while D_s-optimality does not, so this is a property of one criterion, not of optimal design. Part (ii) makes separation a small-batch phenomenon, so the two propositions divide the budget axis between them (Section 5.2).

What modelling costs instead. A campaign that cannot measure κ_{ij} will estimate it from a fitted model, so choosing needs the model's error too. Write $f = \sum_U f_U$ for the functional-ANOVA decomposition under the uniform product measure and $P_2 f = \sum_{|U| \le 2} f_U$ for the best second-order approximation, and draw a rectangle uniformly, x uniform on $\mathcal{X}$ and b_i', b_j' uniform among the other levels.

Proposition 4 (Second-order model bias at a random rectangle)**.** *Write* $c_k = m_k/(m_k - 1)$.

1. $\kappa_{ij}(f_U) = 0$ *identically unless* $\{i, j\} \subseteq U$*: the mixed second difference annihilates every component that does not involve both toggled coordinates.*
2. *Distinct components give uncorrelated contrasts, and for* $\{i, j\} \subseteq U$*,* $\mathbb{E}\big[\kappa_{ij}(f_U)^2\big] = 4\, c_i c_j\, \|f_U\|^2$*. Hence*

$$\mathbb{E}\Big[\big(\kappa_{ij}(f) - \kappa_{ij}(P_2 f)\big)^2\Big] = 4\, c_i c_j \sum_{U \supseteq \{i,j\},\, |U| \ge 3} \|f_U\|^2 \;=:\; 4\, c_i c_j\, V_{\ge 3}^{(ij)}. \tag{2}$$

The squared bias of the *best possible* second-order description of f at a random rectangle is therefore an exact multiple of the higher-order variance involving both toggled coordinates, and with $\mathrm{Var}(\hat{\kappa}) = 4\sigma^2$ that gives a boundary with no free parameter.

Corollary 1 (When modelling can win)**.** *At a uniformly random rectangle, the exact second-order projection has smaller mean squared error than the four-point contrast if and only if*

$$V_{\ge 3}^{(ij)} \;<\; \sigma^2 \Big(1 - \tfrac{1}{m_i}\Big)\Big(1 - \tfrac{1}{m_j}\Big). \tag{3}$$

A fitted model also pays estimation variance, so (3) *is necessary, not sufficient, for a* fitted *second-order model to win.*

Corollary 1 is an *oracle* boundary in three senses. It compares against $P_2 f$, not any fitting procedure: a fitted model pays estimation variance on top of this bias, which hurts it, and shrinks, which helps

it, so the boundary is necessary but not sufficient for it. It averages over uniformly drawn rectangles, whereas a campaign asks at its leads. And $V_{\geq 3}^{(ij)}$ is a property of the landscape, which a prospective campaign does not know. Section 5.2 tests the identity, the boundary and the fitted comparison separately, where each could have failed.

5 EXPERIMENTS

RECTIFY, a constructive control. Auditing a requirement needs a rule that tries to meet it, so we include a deliberately minimal one as a *constructive control*, not a proposed acquisition function: given an acquisition score μ and uncertainty s, RECTIFY spends a fraction ρ of the batch buying the four corners of high-scoring rectangles and the rest on the highest-μ points left. It is purely combinatorial, in classical terms Fedorov-style augmentation (Mitchell, 1974; Cook & Nachtsheim, 1980; Miller & Nguyen, 1994) restricted to rectangle-completing swaps (Algorithm 1). Its role is to answer one question: does deliberately buying the estimand's support change what a batch can answer? It does, and it still loses to plain greedy utility on that metric (Section 6).

Protocol. Everything is fixed before any rule selects, so all face an identical task. A pool is drawn from the landscape, n_{seed} random pool points are measured with noise σ, one random forest fitted to them supplies (μ, s) to every rule, and the *leads* are the top-m_{lead} pool points under that surrogate. The *estimand set* Q holds local interactions at a lead whose four corners exist in the landscape, so the truth is known; it is subsampled to $|Q|/m_{\text{lead}}$ per lead, fixed once per seed and shared by every rule, and an estimand is *purchasable* when its corners lie in the pool. Each rule then selects N points, measured with noise, and one capacity-matched ridge on the full second-order basis is refitted on seed set plus batch. Certification is credited to the *batch alone* and decision metrics use only combinations the rule did not measure. Five seeds; $\pm$ is the standard error.

Rules, control and landscapes. Twelve rules. Nine classical: random selection, greedy utility (Top-N), UCB, maximin Hamming diversity, greedy level coverage, greedy Bayesian D- and D_s-optimal designs (Atkinson & Cox, 1974; DuMouchel & Jones, 1994) computed on the *full second-order model* so the interactions we report as uncertified are in the design's own model, RECTIFY, and a control reproducing its acquisition score and distance profile while removing rectangles. That control is rectangle-*reducing* rather than rectangle-free and imperfect in two tabulated ways (Table 6), so we read it as indicative.

Three modern batch families, each a *principle-preserving adaptation* to a regression surrogate rather than a reimplementation, since the published forms assume a neural network and, for two of them, classification. A *determinantal* batch: greedy MAP inference for a quality-weighted DPP, $L = \operatorname{diag}(q)\, S \operatorname{diag}(q)$, $q_i = e^{\beta \mu_i / 2}$, S an RBF similarity on one-hot features (Kathuria et al., 2016; Bıyık et al., 2019). A *gradient-embedding* batch: BADGE (Ash et al., 2020) in its Fisher reading (Ash et al., 2021), where for a linear model on the second-order basis the loss gradient is the feature vector scaled by the residual, so k-means++ on those embeddings picks points both uncertain and spread. A *redundancy-aware information* batch: BatchBALD (Kirsch et al., 2019) under a Gaussian surrogate, where the joint mutual information is $\frac{1}{2} \log\det(\Sigma_L + \sigma^2 I)$ less the per-point noise entropies, so greedy maximisation is a log-determinant step. Section H gives every setting; the "-style" labels are there so no reader takes these for the published algorithms.

The four landscapes (Table 17) are a complete binary 2^{13} fluorescent protein (Poelwijk et al., 2019), two four-site twenty-level proteins (Wu et al., 2016; Johnston et al., 2024) and a three-site twenty-level interface (Lite et al., 2020); PARD3 at 0.849 additive is a negative control we did not construct, and Poelwijk et al. (2019)'s replicates make our headline landscape's noise measured, $\sigma = 0.2568$. On the two 20^4 landscapes the optimal designs' candidate set is capped by a compute budget; shrinking the pool until the cap cannot bind leaves the result unchanged (Section K).

5.1 THE AUDIT: HOW MUCH DOES EACH RULE CERTIFY?

Table 1 is the paper's main measurement; read it one rule at a time, since any grouping into "spreading" and "concentrating" would be ours rather than the data's. The three modern batch rules end where the classical ones end but do not get there the same way: 2 of the three certify at most

10^{-3} (Gradient emb. (BADGE-style), Info. batch (BatchBALD-style)) and the determinantal batch reaches only 0.017, against D-optimal's exact 0.000, but D-optimal separates ($d_{\min} = 2.0$) and none of the three does (1.1, 1.0, 1.1), and D-optimal never exceeds the occupancy threshold while the determinantal batch exceeds it in 60% of runs (Table 14). Nor is the gap noise: paired over 20 landscape-seed pairs the determinantal batch certifies +0.017 more than D-optimal ($p = 0.001$). They share the outcome, not the cause. Greedy utility, which optimises nothing about information, certifies 0.128; the one exception among the spreading rules is greedy coverage at 0.047, driven entirely by 0.189 on the binary landscape.

A campaign also holds its seed data and its lead is already measured, and the union repairs little: Table 1's last column gains at most 0.036 (Greedy utility (Top-N)), with 5 rules still certifying nothing. History helps where rectangles are nearly closed and does nothing where the geometry was never there, which also fixes the requirement's scope: the four corners must be *observed*, not newly bought.

The mechanism is separation, and it is derivable. The zero rows are not a brute empirical fact: Table 1 prints the explanation in its $d_{\min}$ column. Pooling the 1240 batches of the sweeps reported here, $d_{\min} \geq 2$ holds in 91.5% of D-optimal, 95.4% of maximin and 4.6% of D_s-optimal runs against 0.0% of every other rule's; among the 249 separated batches the number containing any rectangle is 0 and the mean fraction certified exactly 0.0000, as Proposition 3 requires, and the asset pipeline aborts rather than emit a number if that invariant is violated. The zero therefore follows from Proposition 3 once one observes that D-optimality separates, and that observation is the finding. Separated batches are not blind either: over 93 batches they hold 0 two-coordinate but 3067 *block* measurements (Remark 1), only 1.3% of which touch a lead (Table 12).

Global identifiability does not imply local measurability. It would be a misreading to take the zeros for incompetence. On the binary landscape $p_{\text{int}} = 78$ against $N = 160$, so the full second-order interaction block is attainable and 5 of the 12 rules attain it (Gradient emb. (BADGE-style), Bayesian D-optimal, Bayesian D_s-optimal, MaxMin diversity, Random), each certifying at most 0.000 at its own leads. The separation is clean rather than merely negative: the 5 rules that certify anything sit at rank share at most 0.500 (Matched control), the 7 that certify nothing at 0.731 or above (Determinantal (DPP)), nothing between, Spearman -0.90 ($p = 0.000$, Table 13). That a *gradient-embedding* batch sits at full rank with zero certification is the sharpest form of the point, since identifying that block is what the family is built for. We claim the paradox only where it can arise: on the twenty-level landscapes $p_{\text{int}} = 2166$ against a few hundred points, no rule exceeds a seventh of the block. Nor do we rank rules by the reliability of their claims, since rules measuring 0.434 and 0.000 of their own top claims are not comparable on that axis (Section I); we compare *estimators* on a fixed estimand set (Table 17).

5.2 Measure or model: three tests of one boundary

Corollary 1 makes three separable claims, and conflating them is the easiest way to overstate it: the *identity* predicts the bias of $P_2 f$ from higher-order variance; the *oracle boundary* says that bias beats $4\sigma^2$ exactly when $V_{\geq 3}^{(ij)} < \sigma^2(1 - 1/m_i)(1 - 1/m_j)$; and the *fitted comparison*, which a campaign actually sees, is the one the boundary does not govern, since a fitted model pays estimation variance and also shrinks. We test each where it can fail.

One pair at a time. One number per landscape is four binary calls, which a fair coin matches with probability $1/16$. The corollary is indexed by (i, j), so we evaluate both sides for every coordinate pair: 93 pairs over 4 landscapes, 27 predicted to favour measuring and 66 modelling, so both sides are populated. Against brute-force enumeration of rectangles that exist in each landscape the identity holds to a median 0.9% and a worst 5.5%, and the predicted side of 2σ is the observed side for 93 of 93 pairs ($p = 10^{-28}$, Table 18).

Crossing the boundary on purpose. A prediction never evaluated near its own boundary is cheap, and both sides of this one can be moved. On a synthetic 5^5 space we sweep the order-≥ 3 variance over 7 values and the noise over 5, so $r = V_{\geq 3}^{(ij)}/T_{ij}$ runs from 0.00 to 94 and crosses 1 in both arguments (Figure 2, left). The oracle projection changes side where Corollary 1 says it does in 33

Table 1: **The audit.** Left: the fraction of purchasable local interactions at the shared leads each *batch* certifies, per landscape (five seeds, ± s.e.). Right, averaged over landscapes: the fraction of coordinate pairs admitting *some* additive-free contrast, i.e. positive cycle rank in Proposition 5, far weaker than certification; minimum pairwise Hamming distance; interaction information $\mathcal{I}(L)$; and "cum.", the fraction certified by the seed history *with* the batch.

	fraction certified by the batch ↑				averaged diagnostics			
Rule	Poelwijk	GB1	TrpB	ParD3	cycles	$d_{\min}$	$\mathcal{I}(L)$	cum.↑
Random	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.08	1.0	285	0.000
MaxMin diversity	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.00	2.8	292	0.000
Coverage	0.189 ±0.02	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.81	1.0	232	0.066
Bayesian D-optimal	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.00	2.0	309	0.000
Bayesian D_s-optimal	0.000 ±0.00	0.026 ±0.01	0.029 ±0.02	0.001 ±0.00	0.52	1.0	251	0.015
Determinantal (DPP)	0.000 ±0.00	0.026 ±0.01	0.032 ±0.01	0.010 ±0.00	0.24	1.1	219	0.023
Gradient emb. (BADGE-style)	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.12	1.0	266	0.000
Info. batch (BatchBALD-style)	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.000 ±0.00	0.28	1.1	248	0.000
Greedy utility (Top-N)	0.270 ±0.02	0.065 ±0.01	0.067 ±0.01	0.109 ±0.02	0.32	1.0	200	0.164
UCB	0.028 ±0.02	0.048 ±0.01	0.023 ±0.01	0.000 ±0.00	0.34	1.0	202	0.033
Matched control	0.103 ±0.01	0.003 ±0.00	0.003 ±0.00	0.004 ±0.00	0.66	1.0	215	0.049
RECTIFY (control)	0.272 ±0.03	0.045 ±0.01	0.044 ±0.01	0.077 ±0.01	0.86	1.0	200	0.143

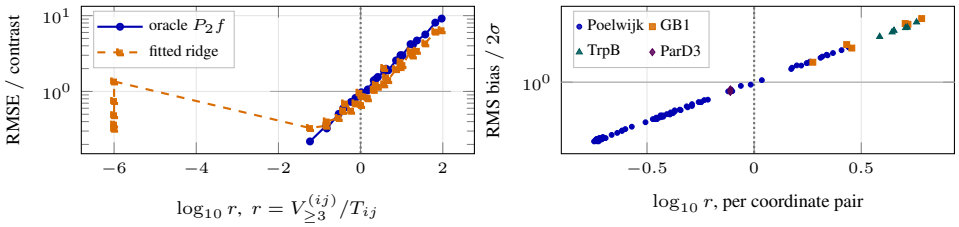


Figure 2: **The boundary, crossed on purpose.** Theory's boundary at $\log_{10} r = 0$ (dotted), empirical tie at 1 (horizontal). **Left:** the factorial sweep, 7 order-$\geq$ 3 shares by 5 noise levels, five seeds; the oracle projection changes side at the predicted boundary, the fitted ridge later because it shrinks. **Right:** all 93 coordinate pairs; predicted and observed sides agree for 93.

of 35 cells, the two exceptions at $r = 1.03, 1.04$ being within sampling error of the boundary; the fitted ridge crosses later, between $r = 1.50$ and 2.08 rather than 1.04 and 1.46, because shrinkage buys it a margin an unbiased projection does not have. Moving σ instead is sharper, since σ is a stated setting on three of the four landscapes: re-evaluating across the noise sweep already in hand takes r from 0.04 to 99.5, predicts a change of side for both landscapes inside the range, and matches the fitted comparison in 7 of 10 cells (Table 19); the three disagreements fail in the two directions named above and no others.

What this licenses. Evaluated at the leads on the *same* local interactions, the oracle prediction agrees with the fitted comparison on all 4 landscapes, the closest call being ParD3 at 0.77, where modelling's measured margin is indeed smallest (Table 17). Read with the three tests above, that is corroboration of an oracle boundary under a distribution shift it does not cover, not a prospective decision rule: $V_{\geq 3}^{(ij)}$ is a property of the landscape, which a campaign does not know before measuring it. Three things could carry the verdict instead of the boundary, and none does: the order-$\geq$ 3 share is a residual collinear with the ridge's parameter-to-observation ratio, which the threshold does not involve; the landscapes are near-complete, and recomputing the decomposition on the largest fully observed sub-product-space of each, and after deleting a further 5, 10 and 20 percent of cells, changes the predicted winner in 0 of 17 variants; and the sign survives 5 ridge penalties and a cross-validated ridge. Detail in Section N.

Which regime makes it worse. The first objection to Table 1 is arithmetic: with K=4, m=20 and N=240 a coordinate pair leaves 400 backgrounds, so a spreading rule has under one point per background whatever its criterion. Sweeping N by a factor of 16× on an 8^4 space does not

rescue it and separates the three mechanisms (Section P): maximin keeps $d_{\min}$=2 and *exactly* zero rectangles to N=400; D-optimal's occupancy climbs only from 2 to 17, never reaching z, and the 128 accidental rectangles it holds at N=960 still certify 0.000; D_s-optimal concentrates past z at *every* budget, building 643 rectangles at the largest and certifying zero because they are not at the leads. The modern rules split across the same three, and only the first two are computable from the batch alone.

Proposition 2 is easy to read as saying that more levels per coordinate means less certification, and a sweep over m shows exactly that; but that sweep shrinks K as m grows, so the trend is an association. Holding $K = 4$ and sweeping m in two arms, one at a fixed budget and one at a fixed fraction of the space, settles it: over 63 rule-by-m cells certification is unrelated to the level count ($\rho = -0.06$, $p = 0.63$) and strongly related to occupancy against the threshold ($\rho = 0.80$, $p = 10^{-15}$; Table 21), so the apparent level effect is what a fixed budget does to occupancy in a growing space. The purchasable fraction meanwhile scales as the cube of pool coverage (Figure 3).

6 Discussion, limitations and conclusion

Batch rules miss this estimand for reasons never visible in the criterion they optimise: diversity separates, which is incompatible with the near-duplication a local contrast needs; D-optimality spreads and holds occupancy under the threshold; D_s-optimality and the modern information batches concentrate where their criterion finds information rather than at the leads. None of this makes them bad at what they were built for; it makes them the wrong instrument for a question asked at a named background, where a campaign reporting a local interaction from such a batch reports model output, not measurement.

The operating envelope. Certification guarantees one thing: that the pairwise question at a named lead is answerable from the batch without assuming a model, for *linear* estimators unbiased under every response on designs without replication (Proposition 1). Proposition 3 forecloses the pairwise estimand only, Proposition 4 concerns the exact projection at a uniformly random rectangle rather than a fitted model at the leads, and the landscapes are biological and retrospective, the noise homoscedastic, the decision metric one round at fixed leads.

It guarantees nothing else, and six results mark where it stops (Section Q has each in full). *(i)* Buying the geometry deliberately does not buy more of it: RECTIFY certifies 0.110 against greedy utility's 0.128 (paired -0.018, $p = 0.000$), and its knob moves *which* pairs are certified, not how many. *(ii)* The rectangle-reducing control has the best held-out regret of every rule it is run on, 1.46 against 1.96 (paired -0.509, $p = 0.014$), but it is not matched, costing up to 31% of the acquisition score on the twenty-level landscapes (Table 6); it does not show the geometry is harmful, it shows these experiments give no evidence that buying it pays for itself on a one-round fixed-lead metric. *(iii)* Certified interactions do not transfer to unmeasured backgrounds, worst where order-≥ 3 structure is largest, the regime where local measurement is most worth having. *(iv)* The reliability advantage does not survive high noise on one of two landscapes, the group means inverting at 2.0$\times$ the measured σ on POELWIJK, since an unbiased contrast has variance $4\sigma^2$ and does not shrink. *(v)* Certification is bit-for-bit identical under a random-forest refit, but claim quality and regret are not (Table 7). *(vi)* A 5×4 stress grid puts RECTIFY's regret behind the best other rule in 20 of 20 cells (Section J).

Conclusion. Three things. Experimental design should be judged against the estimand a campaign intends to answer, not only a global information criterion. Global identifiability under a model and model-free local measurability are distinct resources, and optimising the first drives the second to exactly zero for five of our twelve rules. And higher-order structure prices the alternative: the exact second-order projection is worth its bias only while $V_{\geq 3}^{(ij)} < \sigma^2(1 - 1/m_i)(1 - 1/m_j)$, a boundary that survives being crossed deliberately in both of its arguments, and that a fitted model tracks but does not obey. Ask of a proposed batch not how informative it is, but which questions it makes directly answerable.

AI USE STATEMENT

In this work we used generative AI tools for the following tasks with required disclosure: implementing methods, generating synthetic data sets, providing feedback on the experimental design, cleaning and reformatting datasets, and interpreting results. Specifically, an AI coding assistant was used to write and refactor the simulation, design and diagnostic code, to write the scripts that download and parse the public landscapes, and to generate the figures and the table bodies from the result files; all such code was reviewed line by line by the authors, and every structural claim in Section 4 is additionally covered by an exhaustive brute-force verification script whose output is reported in Section F. AI tools were used to help formulate and refine the framing of the contribution, to suggest which controls and negative controls the claims required, and to critique drafts as adversarial reviewers; the authors then verified each suggestion against the literature and the data, and several AI-suggested claims were discarded after that check, including an earlier and incorrect claim that rectangle contrasts span the estimable interaction contrasts of an incomplete design. We have not used generative AI tools to produce any experimental measurement, and no figure in this paper is an AI-generated image: the data panels are drawn from the result files by the plotting code, and the schematic panels of Figure 1 are hand-authored vector drawings. Additionally, we used generative AI tools for tasks with recommended disclosure: literature search and identification of related work, drafting and editing prose for readability, formatting references, proposing the title, and preparing slides for internal meetings. All references were then opened and checked by the authors against the published record, and several citations proposed by AI tools were found to be inaccurate and corrected or removed. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including its text, figures and claims.

REPRODUCIBILITY STATEMENT

Section 5 states the full protocol, and Section H lists every hyper-parameter, budget, seed-set size, noise level and rule setting used in each figure and table. The four landscapes are public and Section G gives their provenance, licences, the exact files used and the completeness and variance shares we computed from them. Section C proves every statement in Section 4, and Section F reports the exhaustive computational verification of each one over small product spaces, including the counts in Table 4 and the cross-check against a published circuit count.

ETHICS STATEMENT

This work is a retrospective computational study of four public, previously published biological landscape datasets; no new experiments on any organism were performed and no human subjects or personal data are involved. The datasets are redistributed by their authors under CC0 or CC BY licences, and we used them within those terms. Our method allocates measurement effort in molecular and protein design campaigns, so the dual-use considerations are those of protein engineering generally: a method that makes interaction effects cheaper to measure could in principle accelerate the design of harmful as well as beneficial molecules. The contribution here is an identifiability argument and an audit of existing selection rules rather than a capability for designing any specific molecule, and it provides no sequence, compound or protocol; we judge the marginal risk to be low relative to the tools already in public use.

REFERENCES

Bruce J. Wittmann, Yisong Yue, and Frances H. Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 12(11):1026–1045.e7, 2021. doi: 10.1016/j.cels.2021.07.008.

Jason Yang, Ravi G. Lal, James C. Bowden, Raul Astudillo, Mikhail A. Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H. Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, 2025. doi: 10.1038/s41467-025-55987-8.

Christian Kramer. Nonadditivity analysis. *Journal of Chemical Information and Modeling*, 59(9): 4034–4042, 2019. doi: 10.1021/acs.jcim.9b00631.

Max A. Horlbeck, Albert Xu, Min Wang, Neal K. Bennett, Chong Y. Park, Derek Bogdanoff, Britt Adamson, Eric D. Chow, Martin Kampmann, Tim R. Peterson, et al. Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967.e22, 2018. doi: 10.1016/j.cell.2018.06.010.

Zuheng Xu, Moksh Jain, Ali Denton, Shawn Whitfield, Aniket Didolkar, Berton Earnshaw, and Jason Hartford. Automated discovery of pairwise interactions from unstructured data. *arXiv preprint arXiv:2409.07594*, 2024. URL https://arxiv.org/abs/2409.07594.

Tarun Kathuria, Amit Deshpande, and Pushmeet Kohli. Batched Gaussian process bandit optimization via determinantal point processes. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/a1d7311f2a312426d710e1c617fcbc8c-Abstract.html.

Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*, 2019. URL https://arxiv.org/abs/1906.07975.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with Fisher embeddings. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8927–8939, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/4afe044911ed2c247005912512ace23b-Abstract.html.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Paul J. Carter, Greg Winter, Anthony J. Wilkinson, and Alan R. Fersht. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell*, 38(3):835–840, 1984. doi: 10.1016/0092-8674(84)90278-2.

Amnon Horovitz and Alan R. Fersht. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *Journal of Molecular Biology*, 214(3):613–617, 1990. doi: 10.1016/0022-2836(90)90275-Q.

Christian Kramer, Julian E. Fuchs, and Klaus R. Liedl. Strong nonadditivity as a key structure-activity relationship feature: Distinguishing structural changes from assay artifacts. *Journal of Chemical Information and Modeling*, 55(3):483–494, 2015. doi: 10.1021/acs.jcim.5b00018.

Dea Gogishvili, Eva Nittinger, Christian Margreitter, and Christian Tyrchan. Nonadditivity in public and inhouse data: Implications for drug design. *Journal of Cheminformatics*, 13(1):47, 2021. doi: 10.1186/s13321-021-00525-z.

Ramamurthy Mani, Robert P. St. Onge, John L. Hartman, Guri Giaever, and Frederick P. Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008. doi: 10.1073/pnas.0712255105.

Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. The context-dependence of mutations: A linkage of formalisms. *PLOS Computational Biology*, 12(6):e1004771, 2016. doi: 10.1371/journal.pcbi.1004771.

Zachary R. Sailer and Michael J. Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017. doi: 10.1534/genetics.116.195214.

Milo S. Johnson, Alena Martsul, Sergey Kryazhimskiy, and Michael M. Desai. Higher-fitness yeast genotypes are less robust to deleterious mutations. *Science*, 366(6464):490–493, 2019. doi: 10.1126/science.aay4199.

Daniel M. Lyons, Zhengting Zou, Haiqing Xu, and Jianzhi Zhang. Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. *Nature Ecology and Evolution*, 4 (12):1685–1693, 2020. doi: 10.1038/s41559-020-01286-y.

Christopher W. Bakerlee, Alex N. Nguyen Ba, Yekaterina Shulgina, Jose I. Rojas Echenique, and Michael M. Desai. Idiosyncratic epistasis leads to global fitness-correlated trends. *Science*, 376 (6593):630–635, 2022. doi: 10.1126/science.abm4774.

Yeonwoo Park, Brian P. H. Metzger, and Joseph W. Thornton. Epistatic drift causes gradual decay of predictability in protein evolution. *Science*, 376(6595):823–830, 2022. doi: 10.1126/science.abn6895.

David Birkes, Yadolah Dodge, and Justus Seely. Spanning sets for estimable contrasts in classification models. *The Annals of Statistics*, 4(1):86–107, 1976. doi: 10.1214/aos/1176343349.

Roberto Fontana, Fabio Rapallo, and Henry P. Wynn. Circuits for robust designs. *Statistical Papers*, 63(5):1537–1560, 2022. doi: 10.1007/s00362-021-01285-6.

Divya Shyamal, Jiaqi Zhang, and Caroline Uhler. Probabilistic factorial experimental design for combinatorial interventions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55474–55492, 2025. URL `https://proceedings.mlr.press/v267/shyamal25a.html`.

Anthony C. Atkinson and D. R. Cox. Planning experiments for discriminating between models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(3):321–334, 1974. doi: 10.1111/j.2517-6161.1974.tb01010.x.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006. ISBN 9780898716047. doi: 10.1137/1.9780898719109.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. doi: 10.1214/aoms/1177729442.

Naoki Egami and Kosuke Imai. Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, 114(526):529–540, 2019. doi: 10.1080/01621459.2018.1476246.

Tirthankar Dasgupta, Natesh S. Pillai, and Donald B. Rubin. Causal inference from 2^k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):727–753, 2015. doi: 10.1111/rssb.12085.

J. A. Eccleston and A. Hedayat. On the theory of connected designs: Characterization and optimality. *The Annals of Statistics*, 2(6):1238–1255, 1974. doi: 10.1214/aos/1176342876.

Arthur Fries and William G. Hunter. Minimum aberration 2^{k-p} designs. *Technometrics*, 22(4): 601–608, 1980. doi: 10.1080/00401706.1980.10486210.

Huaiqing Wu and C. F. J. Wu. Clear two-factor interactions and minimum aberration. *The Annals of Statistics*, 30(5):1496–1511, 2002. doi: 10.1214/aos/1035844985.

Ching-Shui Cheng, David M. Steinberg, and Don X. Sun. Minimum aberration and model robustness for two-level fractional factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):85–93, 1999. doi: 10.1111/1467-9868.00164.

Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998. doi: 10.1214/aos/1030563990.

Alessandra Bernardi, Enrico Carlini, Maria Virginia Catalisano, Alessandro Gimigliano, and Alessandro Oneto. The hitchhiker guide to: Secant varieties and tensor decomposition. *Mathematics*, 6(12):314, 2018. doi: 10.3390/math6120314.

J. M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011. ISBN 9780821869079. doi: 10.1090/gsm/128.

Hisayuki Hara, Satoshi Aoki, and Akimichi Takemura. Minimal and minimal invariant Markov bases of decomposable models for contingency tables. *Bernoulli*, 16(1):208–233, 2010. doi: 10.3150/09-BEJ207.

Satoshi Aoki and Akimichi Takemura. Markov chain Monte Carlo exact tests for incomplete two-way contingency tables. *Journal of Statistical Computation and Simulation*, 75(10):787–812, 2005. doi: 10.1080/00949650410001690079.

Richard C. Singleton. Maximum distance q-nary codes. *IEEE Transactions on Information Theory*, 10(2):116–118, 1964. doi: 10.1109/TIT.1964.1053661.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/6488484c982e9af5c35689523ba1abfe-Abstract.html.

Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8005–8015, 2021. URL https://proceedings.mlr.press/v139/neiswanger21a.html.

Hongrui Xie, Junyu Cao, and Kan Xu. Adaptive combinatorial experimental design: Pareto optimality for decision-making and inference. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3466–3474, 2026. URL https://proceedings.mlr.press/v300/hongrui26a.html.

Moksh Jain, Alisandra Kaye Denton, Shawn T. Whitfield, Aniket Rajiv Didolkar, Berton Earnshaw, and Jason Hartford. Active learning to discover pairwise genetic interactions via representation learning. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design (GEM)*, 2024. URL https://openreview.net/forum?id=YwAfzDGiqG.

Haonan Zhu, Mary Silva, Jose Cadena, Braden Soper, Michał Lisicki, Braian Peetoom, Sergio E. Baranzini, Shivshankar Sundaram, Priyadip Ray, and Jeff Drocco. Deep active learning-based experimental design to uncover synergistic genetic interactions for host-targeted therapeutics. *Bioinformatics Advances*, 6(1):vbaf228, 2026. doi: 10.1093/bioadv/vbaf228.

Braden Soper, Michal Lisicki, Mary Silva, Jose Cadena, Haonan Zhu, Shivshankar Sundaram, Priyadip Ray, and Jeff Drocco. Knowledge graph-aided Bayesian active learning for top-K genetic interaction discovery. *Scientific Reports*, 15(1):31196, 2025. doi: 10.1038/s41598-025-13972-7.

Amnon Horovitz. Double-mutant cycles: A powerful tool for analyzing protein structure and function. *Folding and Design*, 1(6):R121–R126, 1996. doi: 10.1016/S1359-0278(96)00056-9.

Amnon Horovitz, Rachel C. Fleisher, and Tridib Mondal. Double-mutant cycles: New directions and applications. *Current Opinion in Structural Biology*, 58:10–17, 2019. doi: 10.1016/j.sbi.2019.03.025.

Laura Guasch, Niels Maeder, John G. Cumming, and Christian Kramer. From mundane to surprising nonadditivity: Drivers and impact on ML models. *Journal of Computer-Aided Molecular Design*, 38(1):26, 2024. doi: 10.1007/s10822-024-00566-0.

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

Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K. Dahlgren, Jeremy Greenwood, Donna L. Romero, Craig Masse, Jennifer L. Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L. Mobley, William L. Jorgensen, Bruce J. Berne, Richard A. Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015. doi: 10.1021/ja512751q.

Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10(1):4213, 2019. doi: 10.1038/s41467-019-12130-8.

Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016. doi: 10.7554/eLife.16965.

Kadina E. Johnston, Patrick J. Almhjell, Ella J. Watkins-Dulaney, Grace Liu, Nicholas J. Porter, Jason Yang, and Frances H. Arnold. A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences*, 121(32):e2400439121, 2024. doi: 10.1073/pnas.2400439121.

Thuy-Lan V. Lite, Robert A. Grant, Isabel Nocedal, Megan L. Littlehale, Monica S. Guo, and Michael T. Laub. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife*, 9:e60924, 2020. doi: 10.7554/eLife.60924.

David H. Brookes, Amirali Aghazadeh, and Jennifer Listgarten. On the sparsity of fitness functions and implications for learning. *Proceedings of the National Academy of Sciences*, 119(1):e2109649118, 2022. doi: 10.1073/pnas.2109649118.

Daniel M. Weinreich, Yinghong Lan, C. Scott Wylie, and Robert B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics and Development*, 23(6):700–707, 2013. doi: 10.1016/j.gde.2013.10.007.

Vincent Q. Tran, Matthew Nemeth, Liam J. Bartie, Sita S. Chandrasekaran, Alison Fanton, Hyungseok C. Moon, Brian L. Hie, Silvana Konermann, and Patrick D. Hsu. Rapid directed evolution guided by protein language models and epistatic interactions. *Science*, 392(6798):eaea1820, 2026. doi: 10.1126/science.aea1820.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2402–2412, 2020. URL `https://proceedings.mlr.press/v108/lengerich20a.html`.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125.

Andisheh Amrollahi, Amir Zandieh, Michael Kapralov, and Andreas Krause. Efficiently learning Fourier sparse set functions. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/c77331e51c5555f8f935d3344c964bd5-Abstract.html.

Justin S. Kang, Yigit E. Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the Möbius transform. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46160–46202, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/520b379123d16e41f85472e766846486-Abstract-Conference.html.

Valentina Kuskova, Dmitry Zaytsev, and Michael Coppedge. When are neural interaction discoveries real? Identifiability, recoverability, and a pre-fit diagnostic. *arXiv preprint arXiv:2606.08390*, 2026. URL https://arxiv.org/abs/2606.08390.

Tamás Kővári, Vera T. Sós, and Pál Turán. On a problem of K. Zarankiewicz. *Colloquium Mathematicum*, 3(1):50–57, 1954. doi: 10.4064/cm-3-1-50-57.

István Reiman. Über ein problem von K. Zarankiewicz. *Acta Mathematica Academiae Scientiarum Hungaricae*, 9(3–4):269–273, 1958. doi: 10.1007/BF02020254.

Paul Erdős, Alfréd Rényi, and Vera T. Sós. On a problem of graph theory. *Studia Scientiarum Mathematicarum Hungarica*, 1:215–235, 1966. URL https://users.renyi.hu/~p_erdos/1966-06.pdf.

William G. Brown. On graphs that do not contain a Thomsen graph. *Canadian Mathematical Bulletin*, 9(3):281–285, 1966. doi: 10.4153/CMB-1966-036-2.

Zoltán Füredi and Miklós Simonovits. The history of degenerate (bipartite) extremal graph problems. In *Erdős Centennial*, volume 25 of *Bolyai Society Mathematical Studies*, pp. 169–264. Springer, Berlin, Heidelberg, 2013. ISBN 9783642392856. doi: 10.1007/978-3-642-39286-3_7.

Toby J. Mitchell. An algorithm for the construction of "D-optimal" experimental designs. *Technometrics*, 16(2):203–210, 1974. doi: 10.1080/00401706.1974.10489175.

R. Dennis Cook and Christopher J. Nachtsheim. A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, 22(3):315–324, 1980. doi: 10.1080/00401706.1980.10486162.

Alan J. Miller and Nam-Ky Nguyen. Algorithm AS 295: A Fedorov exchange algorithm for D-optimal design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(4):669–677, 1994. doi: 10.2307/2986264.

William DuMouchel and Bradley Jones. A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, 36(1):37–47, 1994. doi: 10.1080/00401706.1994.10485399.

A NOTATION

Table 2 collects every symbol used in the paper, with the section in which it is introduced. The three that carry the most weight are κ_{ij}, the local interaction of Definition 1; $d_{\min}$, the minimum pairwise Hamming distance of a batch, which by Proposition 3 decides on its own whether any local interaction can be certified; and $\mathcal{I}(L)$, the interaction information that an information-optimal design maximises.

Table 2: Every symbol used in the paper, in order of first appearance.

Symbol	Meaning
$\mathcal{X} = B_1 \times \cdots \times B_K$	product design space, $m_k = \lvert B_k \rvert$ levels at coordinate k
K, m	number of coordinates; common number of levels when all m_k are equal
$f : \mathcal{X} \to \mathbb{R}$	the response; measured with independent noise of variance σ^2
$\mathcal{P} \subseteq \mathcal{X}$	pool: the points a campaign may assay ($\mathcal{P} = \mathcal{X}$: anything synthesisable)
$L \subseteq \mathcal{P}$, $N = \lvert L \rvert$	the batch chosen by an acquisition rule, and its size
x^i, x^{ij}	x with coordinate i (resp. i and j) changed to a stated alternative
x^S	x with every coordinate in the block S changed to a stated alternative
$\mathsf{rect}_{ij}(x; b'_i, b'_j)$	the rectangle $\{x, x^i, x^j, x^{ij}\}$
$\kappa_{ij}(x; b'_i, b'_j)$	local interaction, the mixed second difference (1)
θ_{ij}	background-averaged (functional-ANOVA) pairwise interaction
$\mathcal{A}$, A	additive functions on $\mathcal{X}$, and their model matrix
Z	pairwise-interaction model matrix (reference coding)
p_{main}, p_{int}	$1 + \sum_k (m_k - 1)$ and $\sum_{i<j} (m_i - 1)(m_j - 1)$
G_L	$(I - P_{A_L}) Z_L$, the interaction design after projecting out additive nuisance
$\mathcal{I}(L)$	interaction information $\frac{1}{2} \log\det(I + \sigma^{-2} G_L^\top G_L)$
$d_{\min}(L)$	minimum pairwise Hamming distance within the batch
g, n_g	a background for the coordinates outside (i, j), and the batch points carrying it
H, $\text{comp}(H)$	bipartite level-incidence graph at (i, j, g), and its number of components
$z(m, n; 2, 2)$	Zarankiewicz number: most cells of an $m \times n$ array with no filled 2×2
μ, s	surrogate mean and uncertainty supplied to every rule
β	optimism weight in the score $\mu + \beta s$
ρ, J	fraction of the batch RECTIFY spends on geometry; its block order
n_{seed}, m_{lead}	shared seed-set size; number of leads shared by all rules
Q	estimand set: local interactions at the leads whose truth is known

B CLASSICAL STATEMENTS USED IN THE MAIN TEXT

Section 4 paraphrases the following two classical statements and one counting remark. Both statements are proved in Section C, and both were also verified by exhaustive computation (Section F).

Lemma 1 (Minimal additive-free measurements; classical)**.** *Let A be the model matrix of $\mathcal{A}$ on $\mathcal{X}$. (i) No nonzero c with $c^\top A = 0$ has support of size 1, 2 or 3. (ii) If $\lvert\text{supp}(c)\rvert = 4$ then, up to scale, c has weights $(+1, -1, -1, +1)$ and its support is $\{x, x^{S_1}, x^{S_2}, x^{S_1 \cup S_2}\}$ for a nontrivial bipartition (S_1, S_2) of the set of coordinates on which the four points differ, where x^S toggles every coordinate in S. Taking $S_1 = \{i\}$, $S_2 = \{j\}$ gives the rectangle of Definition 1.*

Remark 1 (Most minimal measurements are not two-coordinate rectangles). The bipartition in Lemma 1(ii) is over coordinate *blocks*, and for $K \geq 3$ the two-coordinate ones are a minority: only 24% of the 100 minimal measurements of 2^4 (Section F). A block rectangle with $\min(\lvert S_1\rvert, \lvert S_2\rvert) \geq 2$ has all six pairwise distances at least two, so Proposition 3(i) does not exclude it; Section 5.1 measures how many our batches hold.

Proposition 5 (What a rectangle-free batch gets instead; classical)**.** *Fix a coordinate pair (i, j) and a background g for the remaining coordinates, and let H be the bipartite graph on $B_i \sqcup B_j$ with one edge per point of L with background g. The additive-free contrasts supported on those points are exactly the cycle space of H, of dimension $\lvert E\rvert - \lvert V\rvert + \text{comp}(H)$ (Birkes et al., 1976). A contrast supported on a cycle of length 2ℓ has variance $2\ell\,\sigma^2$, and its interaction loading is the alternating sum of the 2ℓ cells the cycle visits, which equals a single local interaction only for $\ell = 2$, the rectangle.*

C PROOFS

Proof of Lemma 1. Let $c \neq 0$ satisfy $c^\top A = 0$ with support $S = \mathrm{supp}(c)$, $|S| = s$. The columns of A are the all-ones vector together with, for each coordinate k and each level $v \in B_k$, the indicator of $\{x : x_k = v\}$. Hence $c^\top A = 0$ says exactly: $\sum_{x \in S} c_x = 0$, and for every coordinate k and level v, $\sum_{x \in S:\, x_k = v} c_x = 0$. In words, for every coordinate the partition of S by the level it takes at that coordinate must be blockwise zero-sum.

(i) If $s = 1$ the first condition forces $c = 0$. If $s = 2$, write $S = \{x, x'\}$ with $c_{x'} = -c_x \neq 0$. For any coordinate k, if $x_k \neq x'_k$ then the block of x is a singleton and zero-sum forces $c_x = 0$; so $x_k = x'_k$ for all k, i.e. $x = x'$, a contradiction. If $s = 3$, consider any coordinate k. The level-partition of the three points is either a single block, or a block of two and a singleton, or three singletons; the latter two force a weight to vanish. So every coordinate is constant on S, hence all three points coincide, a contradiction.

(ii) Let $s = 4$. By the same argument no coordinate may split S as $3 + 1$ or $1 + 1 + 1 + 1$, so every coordinate is either constant on S or splits it into two pairs, each pair summing to zero. Normalise $c_{x^1} = 1$. Pick any non-constant coordinate; it pairs x^1 with some x^t and the weights satisfy $c_{x^t} = -1$ and $c_{x^u} + c_{x^v} = 0$ for the other two, so the multiset of weights is $\{1, -1, \lambda, -\lambda\}$. If every non-constant coordinate induced the same pairing, then the two points paired with each other would agree at every coordinate, contradicting distinctness; so two different pairings occur, and matching them forces $\lambda = 1$, giving weights $(+1, -1, -1, +1)$. Label the two positive points x and w and the two negative points u, v. Each non-constant coordinate must pair the points so that each pair sums to zero, i.e. it pairs $\{x, u\}, \{w, v\}$ or $\{x, v\}, \{w, u\}$; the pairing $\{x, w\}, \{u, v\}$ is excluded because those blocks sum to ± 2. Let S_1 be the coordinates of the first type and S_2 those of the second. Both are nonempty (else two points coincide) and disjoint, their union is the set D of non-constant coordinates, and by construction $u = x^{S_1}$, $v = x^{S_2}$ and $w = x^{S_1 \cup S_2}$, which is the claimed form. □

Proof of Proposition 1. (1) The functional κ in (1) is a fixed linear combination of the four values $f(x), f(x^i), f(x^j), f(x^{ij})$ with nonzero coefficients. Suppose $\hat{\kappa} = \sum_{z \in L} c_z\, y(z)$ is unbiased for every f, i.e. $\sum_z c_z f(z) = \kappa(f)$ identically. Evaluating both sides at the indicator function $f = \mathbf{1}_{\{z_0\}}$ for each $z_0 \in \mathcal{X}$ gives c_{z_0} on the left (or 0 if $z_0 \notin L$) and the coefficient of $f(z_0)$ in κ on the right. The latter is nonzero at each of the four corners, so each corner lies in L with the corresponding coefficient, and c vanishes elsewhere. Conversely if the four corners lie in L then the contrast is available and unbiased.

(2) The argument in (1) determines the coefficient vector on every point, so the unbiased *linear* estimator is unique; its variance is $\sigma^2 \|c\|_2^2 = 4\sigma^2$, and uniqueness makes it trivially the minimum-variance one within that class. We claim nothing about nonlinear estimators: the setup assumes only independent noise of variance σ^2, and without a distributional assumption there is no Rao-Blackwell argument to extend the statement. Under Gaussian noise the contrast is the least-squares estimator in the saturated model on those four points and is then UMVU by the standard argument.

(3) Unbiasedness in (1) was established for all f simultaneously, with no model assumed; in particular the estimator is exact for any additive part, since $c^\top A = 0$ by Lemma 1(ii) with $S_1 = \{i\}, S_2 = \{j\}$, and for any structure of order three or higher, since κ is defined directly from f rather than from a projection of it. □

Proof of Proposition 5. Restrict attention to the points of L sharing the background g off $\{i, j\}$; on that set every additive function reduces to $c' + a_i(x_i) + a_j(x_j)$, so the model matrix is the vertex-edge incidence matrix of the bipartite graph H (one edge per point). The kernel of the transpose of a graph's incidence matrix is its cycle space, of dimension $|E| - |V| + \mathrm{comp}(H)$; see Birkes et al. (1976) for the statistical statement in terms of loops, and any algebraic graph theory text for the graph statement. A cycle of length 2ℓ uses 2ℓ edges, so its contrast has ± 1 on 2ℓ points and variance $2\ell\sigma^2$. Writing the cycle as $(a_1, b_1), (a_2, b_1), (a_2, b_2), \ldots, (a_\ell, b_\ell), (a_1, b_\ell)$, its loading on the saturated cell parameters is the alternating sum of those 2ℓ distinct cells. For $\ell = 2$ that sum is $(+, -, -, +)$ on the four cells of one rectangle, which is exactly one local mixed second difference; for $\ell > 2$ the 2ℓ cells do not form a rectangle, so the contrast is a strict mixture of local interactions at that background and identifies none of them individually. □

Proof of Proposition 2. Fix (i, j) and background g and consider the $m_i \times m_j$ array whose filled cells are the levels (x_i, x_j) of the batch points with background g. A complete rectangle at g is exactly a filled 2×2 subarray. By definition $z(m_i, m_j; 2, 2)$ is the largest number of filled cells with no such subarray, so $n_g > z$ forces one, and by maximality there is a configuration with z cells and none. The stated bound is Reiman's (Reiman, 1958); for $m = n$ it gives $z(n, n; 2, 2) \leq \frac{1}{2}n(1 + \sqrt{4n - 3}) = \Theta(n^{3/2})$, with equality when a projective plane of the appropriate order exists. The asymptotic order is Kővári et al. (1954); see Füredi & Simonovits (2013) for the history and the exact statements. □

Proof of Proposition 3. (i) The corners x and x^i of $\mathsf{rect}_{ij}(x; b'_i, b'_j)$ agree on every coordinate except i, so any batch containing a two-coordinate rectangle contains a pair at Hamming distance one and has $d_{\min} = 1$. Contrapositively $d_{\min}(L) \geq 2$ leaves no rectangle, and Proposition 1(1) then denies an unbiased linear estimator of every local interaction. (ii) Fix a coordinate k and let π_k delete it. If $\pi_k(u) = \pi_k(v)$ for distinct $u, v \in L$ then u and v agree off k and differ at k, so they are at Hamming distance one, contradicting $d_{\min} \geq 2$. Hence π_k is injective on L and $N \leq \prod_{j \neq k} m_j$; minimising the bound over k deletes the largest coordinate. This is the puncturing argument for the Singleton bound (Singleton, 1964) at distance two, over a mixed alphabet. □

Tightness, for every level vector. Let $k^* = \arg\max_k m_k$ and take

$$C = \big\{x \in \mathcal{X} \;:\; x_{k^*} \equiv \textstyle\sum_{j \neq k^*} x_j \pmod{m_{k^*}}\big\}.$$

The coordinates off k^* may be chosen freely and then determine x_{k^*}, so $|C| = \prod_{j \neq k^*} m_j$, which meets the bound. If $u, v \in C$ differed in exactly one coordinate k, then $k = k^*$ is impossible because x_{k^*} is a function of the others, and for $k \neq k^*$ we would need $u_{k^*} - v_{k^*} \equiv u_k - v_k \pmod{m_{k^*}}$ with $0 < |u_k - v_k| < m_k \leq m_{k^*}$, so the right side is nonzero while the left is zero. Hence $d_{\min}(C) \geq 2$. No arithmetic condition on the m_k is required. We stress that this is a modular *encoding* of the free coordinates, not a group homomorphism: the levels of coordinate j are the integers $0, \ldots, m_j - 1$ and the sum is reduced modulo m_{k^*}, which is well defined as a map of sets for any level vector but is a homomorphism of groups only when every m_j divides m_{k^*}. The argument above uses only the encoding. With all levels equal it is the familiar sum-zero code of size m^{K-1}. Greedy maximin does not reach the capacity, which is what Table 3 measures.

Proof of Proposition 4 and Corollary 1. Write $c_k = m_k/(m_k - 1)$ and recall that f_U has zero mean in each coordinate of U.

(1) Annihilation. κ_{ij} is the composition $\Delta_i \Delta_j$ of two difference operators, $\Delta_i g(x) = g(x^i) - g(x)$. If $i \notin U$ then f_U does not depend on x_i and $\Delta_i f_U = 0$; likewise for j. So $\kappa_{ij}(f_U) = 0$ unless $\{i, j\} \subseteq U$.

(2) The isometry. Fix $U \supseteq \{i, j\}$ and abbreviate $a = f_U(x^{ij})$, $b = f_U(x^i)$, $c = f_U(x^j)$, $d = f_U(x)$, so $\kappa_{ij}(f_U) = a - b - c + d$. Under the stated sampling each of the four corners is marginally uniform on $\mathcal{X}$: x_i is uniform and b'_i is uniform on the $m_i - 1$ other levels, so $\Pr[b'_i = v] = \sum_{u \neq v} \frac{1}{m_i} \cdot \frac{1}{m_i - 1} = \frac{1}{m_i}$. Hence $\mathbb{E}[a^2] = \mathbb{E}[b^2] = \mathbb{E}[c^2] = \mathbb{E}[d^2] = \|f_U\|^2 =: V$. For the cross terms, condition on the coordinates outside i. Since $\sum_u f_U(u, \cdot) = 0$,

$$\mathbb{E}\big[f_U(x_i, \cdot) f_U(b'_i, \cdot)\big] = \frac{1}{m_i(m_i - 1)} \sum_u f_U(u, \cdot)\Big(\sum_v f_U(v, \cdot) - f_U(u, \cdot)\Big) = -\frac{1}{m_i - 1}\,\mathbb{E}\big[f_U^2\big],$$

so $\mathbb{E}[bd] = \mathbb{E}[ac] = -V/(m_i - 1)$ and, by the same argument in coordinate j, $\mathbb{E}[cd] = \mathbb{E}[ab] = -V/(m_j - 1)$. The two remaining pairs differ in both coordinates and factorise, giving $\mathbb{E}[ad] = \mathbb{E}[bc] = V/\big((m_i - 1)(m_j - 1)\big)$. Writing $r_i = 1/(m_i - 1)$, $r_j = 1/(m_j - 1)$ and collecting,

$$\mathbb{E}[\kappa_{ij}(f_U)^2] = 4V + 2Vr_i + 2Vr_i + 2Vr_j + 2Vr_j + 2Vr_ir_j + 2Vr_ir_j$$

factors directly, and is also what one gets by grouping as $\mathbb{E}[(a+d)^2] + \mathbb{E}[(b+c)^2] - 2\mathbb{E}[(a+d)(b+c)]$ and using the same six expectations gives $\mathbb{E}[\kappa_{ij}(f_U)^2] = 4V(1 + r_i)(1 + r_j) = 4c_ic_j\|f_U\|^2$.

Orthogonality. For $U \neq W$, at least one coordinate k lies in exactly one of them, say $k \in U \setminus W$. Conditioning on all coordinates except x_k and any toggled level at k, f_W is constant

in x_k while $\sum_u f_U(u,\cdot) = 0$, so the conditional expectation of the product vanishes; hence $\mathbb{E}[\kappa_{ij}(f_U)\kappa_{ij}(f_W)] = 0$. Summing the isometry over $U \supseteq \{i,j\}$ with $|U| \geq 3$ gives (2).

(3) Pair averaging. A component f_U with $|U| = k$ contributes to exactly $\binom{k}{2}$ of the $\binom{K}{2}$ coordinate pairs and to no others, by part (1).

Corollary 1. The projection $P_2 f$ is deterministic given f, so its mean squared error at a random rectangle is exactly the squared bias (2), while $\mathrm{Var}(\hat{\kappa}) = 4\sigma^2$ with zero bias by Proposition 1(2). Comparing and dividing by $4c_i c_j$ gives (3). □

Every step is checked numerically in Section F: parts (1), (2), (3), the orthogonality claim and the identity itself agree with brute-force enumeration to machine precision over nine level vectors, equal and mixed.

D WHERE SEPARATION DIES, AND WHAT REPLACES IT

Proposition 3(ii) says separation is a small-batch phenomenon. Table 3 traces both mechanisms across the budget axis on the 8^4 space of the batch-size sweep, where the capacity is $8^4/8 = 512$ points; three seeds per row, and the columns are the batch's minimum pairwise Hamming distance, its total rectangle count and its maximum per-background occupancy against Reiman's *bound* $\bar{z} = 25.5$, which for this array exceeds the exact Zarankiewicz number $z(8,8;2,2) = 24$; the guarantee direction is unaffected, but the plotted line is the bound and we label it as such.

Two things are worth reading off it. Greedy maximin is separated with *exactly* zero rectangles up to $N = 400$ and loses separation at $N = 440$, short of the capacity 512 because a greedy farthest-point rule is not an optimal distance-two code; the moment it does, its occupancy jumps past z and rectangles appear in quantity. D-optimality behaves in the opposite way: it gives up separation already at $N = 240$, so Proposition 3 says nothing about it beyond that point, and yet it still holds zero rectangles up to $N = 480$ because its occupancy never approaches z. The two propositions are therefore not two readings of one fact: they are different mechanisms that produce the same zero in different budget regimes, and the audit of Section 5.1 sits squarely in the first.

Table 3: Separation against occupancy on 8^4, three seeds. Capacity for $d_{\min} \geq 2$ is 512 points (Proposition 3(ii)); Reiman's bound is $\bar{z} = 25.5$. Every row with $d_{\min} \geq 2$ has exactly zero rectangles, as Proposition 3(i) requires.

Rule	N	$d_{\min}$	rectangles	$n_g^{\max}$
MaxMin diversity	240	2.0	0	8
	280	2.0	0	8
	320	2.0	0	8
	360	2.0	0	8
	400	2.0	0	8
	440	1.0	81	28
	480	1.0	768	63
	512	1.0	1323	64
Bayesian D-optimal	120	2.0	0	4
	240	1.0	0	7
	480	1.0	0	10
	960	1.0	120	17

E A RECTANGLE-FREE BATCH THAT STILL MEASURES SOMETHING

Proposition 5 is not vacuous, and it is worth exhibiting the case that rules out the stronger claim one might be tempted to make. In a 3×3 layout let the measured cells be $(1,1),(1,2),(2,2),(2,3),(3,3),(3,1)$. No two rows share two columns, so there is no complete 2×2 subarray and, by Proposition 1, no local interaction is estimable. Yet the incidence graph is a single six-cycle, so its cycle rank is $6 - 6 + 1 = 1$, and the contrast

$$+f(1,1) - f(1,2) + f(2,2) - f(2,3) + f(3,3) - f(3,1)$$

annihilates every additive $f(a, b) = c + \alpha_a + \beta_b$: the α terms cancel in pairs by row and the β terms by column. So a rectangle-free batch can measure a quantity free of additive nuisance. That quantity, however, has variance $6\sigma^2$ rather than $4\sigma^2$ and loads on three distinct interaction cells at once, so it does not answer the question of Definition 1 for any single pair of changes. This is why the paper's claim is about the estimand and not about additive-free measurement in general.

F COMPUTATIONAL VERIFICATION OF EVERY STRUCTURAL CLAIM

Because Section 4 carries the paper, each statement was checked by exhaustive computation in addition to being proved. All checks pass. Two caveats on what "exhaustive" means here: the enumeration of dependent four-subsets in Table 4 is exact rational arithmetic, while the classification of each one as a block rectangle uses a numerical null space with tolerance 10^{-8}; and Proposition 2 is not on the list below, because it is Reiman's bound applied to a definition and what we verify computationally is only the construction of rectangle-free configurations below the bound, not its tightness.

- Lemma 1(i): over every level vector listed in Table 4, enumeration of all subsets of size $1, 2, 3$ found no additive-free contrast with full support.
- Lemma 1(ii): over the same level vectors, computed in exact rational arithmetic, every dependent four-subset of the additive model matrix has a one-dimensional kernel with weights ± 1 up to scale and is a block rectangle for a nontrivial bipartition, with no exceptions. Counts in Table 4; the 2^4 total agrees with the 100 support-four circuits reported by Fontana et al. (2022), who do not note that only 24 of them are two-coordinate rectangles. The corresponding share for 3^3 is 33%. The enumeration also has a closed form, which reproduces every row of Table 4: choosing the common background off D and the two toggled level values on it gives

$$\#\{\text{minimal measurements}\} = \frac{\prod_k m_k}{4} \sum_{D \subseteq [K],\, |D| \geq 2} \left(2^{|D|-1} - 1\right) \prod_{k \in D} (m_k - 1),$$

  whose $|D| = 2$ term counts the two-coordinate rectangles. For 2^4 this is $4(6{\cdot}1{+}4{\cdot}3{+}1{\cdot}7) = 100$, of which $4 \cdot 6 = 24$ are classical.
- Proposition 1: for a 4×4 layout, exhaustive search over all subsets of size up to eight confirms that the cheapest unbiased estimator of a fixed local interaction uses four points with $\|c\|^2 = 4$, and that deleting any one of the four corners leaves no unbiased estimator at all.
- Proposition 5: for random incomplete two-way layouts the cycle rank $|E| - |V| + c$ equals the dimension of the kernel of the additive model matrix in 40 of 40 trials; cycle contrasts of length 2ℓ have squared norm 2ℓ and load on exactly one interaction cell if and only if $\ell = 2$.
- Proposition 3(i): over 2^3, 3×2^2, 2^4 and $3^2 \times 2$, every subset of size up to seven was enumerated and every one with $d_{\min} \geq 2$ was checked; across 1,620 separated subsets, none contains a two-coordinate rectangle.
- Proposition 3(ii): on the same spaces plus 4×3, the largest separated subset found by exhaustive search equals the capacity $\prod_k m_k / \max_k m_k$ in all five cases, so the bound is attained and not merely valid. The parity construction of Section C was checked on ten level vectors including four with a composite largest level (6, 10, 12, 20) and three mixed-level spaces: in every case it has exactly $\prod_k m_k / \max_k m_k$ points, $d_{\min} = 2$, and zero two-coordinate rectangles. This is the check that refutes the prime-power hypothesis a reader might expect here, and it is why the appendix claims tightness for every level vector.
- Additive annihilation and the one-pair property: rectangle contrasts annihilate random additive functions to machine precision, and each loads on its own coordinate pair only, over all tested level vectors.
- Non-identifiability is observational: when the projected interaction design is rank deficient we exhibit two parameter vectors giving identical noiseless responses on every measured point and different orderings off the batch.
- Aliasing: a rectangle-free resolution-V design identifies all pairwise interactions of the second-order model yet returns biased interaction estimates once third-order structure is present, while rectangle contrasts recover the local mixed differences exactly on the same response.

Table 4: Exhaustive enumeration of the minimal additive-free measurements, in exact rational arithmetic. "Classical" counts those that change exactly two coordinates; the rest are block rectangles. For $K \geq 3$ the classical ones are at most half, and a strict minority in every case except $2 \times 2 \times 2$, where they are exactly half.

levels	$\lvert\mathcal{X}\rvert$	minimal measurements	classical	block	% classical
2×2	4	1	1	0	100%
3×3	9	9	9	0	100%
4×3	12	18	18	0	100%
$2 \times 2 \times 2$	8	12	6	6	50%
$3 \times 2 \times 2$	12	33	15	18	45%
$2 \times 3 \times 2$	12	33	15	18	45%
$3 \times 3 \times 2$	18	90	36	54	40%
$3 \times 3 \times 3$	27	243	81	162	33%
$2 \times 2 \times 2 \times 2$	16	100	24	76	24%

G Data provenance

All four landscapes are public. We use them as distributed by their authors and compute the completeness figures and functional-ANOVA variance shares of Table 17 ourselves.

- **Poelwijk**: the complete 2^{13} binary landscape of Poelwijk et al. (2019), Supplementary Data of that paper, CC BY 4.0. All 8,192 genotypes present. The two independent replicates in the same supplement give the measured assay noise $\sigma = 0.2568$ on the standardised scale, replicate correlation 0.9899. We use the combined-brightness response, standardised.
- **GB1**: the four-site landscape of Wu et al. (2016) at positions V39, D40, G41, V54, as redistributed in the FLIP benchmark collection; 149,361 of 20^4 variants, 93.4% complete. Underlying data CC BY 4.0.
- **TrpB**: the four-site active-site landscape of Johnston et al. (2024), CC0 1.0; 159,129 of 20^4 variants on the canonical twenty-amino-acid design, 99.5% complete after removing stop codons.
- **ParD3**: the three-site specificity-residue landscape of Lite et al. (2020) from GEO accession GSE153897, CC BY 4.0; 7,882 of 20^3 variants, 98.5% complete. We use the cognate-toxin response.

Responses are standardised to zero mean and unit variance per landscape, so the noise levels σ in Section 5 are on a common scale. Simulated assay noise is added independently to every measurement a rule takes, including the shared seed set.

H Hyper-parameters and protocol settings

Table 5 lists the budget, seed-set size, noise level, number of leads and estimand-set size of every experiment reported in the paper; the paragraphs after it give the surrogate, the refitted model, the rule settings and the synthetic generator.

The level sweep shrinks K as m grows, over $(m, K) \in \{(2, 12), (3, 8), (5, 5), (8, 4), (12, 4), (20, 3)\}$, so the space size stays within a factor of seven; the budget sweep pairs each $N \in \{60, 120, 240, 480, 960\}$ with $n_{\text{seed}} \in \{60, 100, 140, 200, 260\}$; the crossover sweep holds the pairwise share at 0.30 and varies the order-≥ 3 share over $\{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$; the coverage sweep varies the purchasable fraction over $\{0.40, 0.55, 0.70, 0.85, 1.00\}$; and the noise sweep uses $\sigma \in \{0.05, 0.1284, 0.2568, 0.5136, 1.0\}$ on the standardised response, bracketing Poelwijk et al. (2019)'s measured $\sigma = 0.2568$. Table 3 needs no response model, since $d_{\min}$, the rectangle count and the occupancy are properties of the selected points alone.

Surrogate and refitted model. The surrogate shared by all rules is a random forest of 150 trees, minimum leaf size 2, on one-hot features; its per-tree spread supplies s. The capacity-matched refitted model is ridge regression on the full second-order basis $[A, Z]$ with penalty 1.0 and no

Table 5: Settings used in each experiment. N is the batch size, n_{seed} the shared seed-set size, m_{lead} the number of leads, $|Q|$ the target estimand-set size.

Experiment	landscape	N	n_{seed}	σ	m_{lead}	$\lvert Q\rvert$	seeds
Audit, Table 1	Poelwijk	160	120	0.2568	8	480	0–4
Audit, Table 1	GB1	240	180	0.25	8	480	0–4
Audit, Table 1	TrpB	240	180	0.25	8	480	0–4
Audit, Table 1	ParD3	200	150	0.25	8	480	0–4
Level sweep, Figure 5	synthetic m^K	150	140	0.15	6	480	0–4
Budget sweep, Figure 5	synthetic 8^4	60–960	60–260	0.15	6	480	0–2
Crossover, Figure 5	synthetic 5^5	150	140	0.15	6	480	0–4
Coverage sweep, Figure 3	synthetic 5^5	120	120	0.15	6	480	0–4
Noise sweep, Section 6	Poelwijk	160	120	0.05–1.00	8	480	0–2
Noise sweep, Section 6	GB1	240	180	0.05–1.00	8	480	0–2
ρ sweep, Figure 4	synthetic 5^5	120	120	0.15	6	480	0–4
Model control, Table 7	synthetic 5^5	120	120	0.15	6	480	0–4
Capacity, Table 3	synthetic 8^4	120–960	n/a	n/a	n/a	n/a	0–2

intercept column beyond the one in A; the model-class control replaces it with a 200-tree random forest, minimum leaf size 2. Both are identical across rules.

Rule settings. UCB uses $\beta = 2$. Bayesian D- and D_s-optimal greedy selection uses a unit-scale Gaussian prior on both parameter blocks and exact rank-one gain updates; the criterion is computed on the *full* second-order model $[A, Z]$, so the interactions we then report as uncertified are in the design's own model. RECTIFY and the matched control use $\rho = 0.45$, block order $J = 2$, optimism $\beta = 1$, at most 6 alternative levels per coordinate, and a coverage penalty of 1 per previously used coordinate pair. The matched control searches within ± 300 ranks of the acquisition score for each replacement.

RECTIFY in full. Given the acquisition score μ and uncertainty s, RECTIFY walks the pool in decreasing μ, treats each point as an anchor, and buys the four corners of the rectangle at it with the largest mean optimistic score $\mu + \beta s$, less a penalty of 1 for each coordinate pair already covered; it stops when a fraction ρ of the batch has been spent this way and fills the remainder with the highest-μ unused points. Nothing is fitted and nothing is optimised numerically: the rule is a walk plus a lookup (Algorithm 1).

The three modern rules, and how each was adapted. The published forms of these methods assume a neural surrogate, and two of them assume classification, so none is defined as written on a random-forest regression surrogate. Each is therefore a *principle-preserving adaptation*: we state the defining object of the original, the object that plays its role here, and every setting. None is a reimplementation, and the labels in every table say "-style" for the two that need it.

Determinantal (DPP). Original: sample or maximise under a determinantal point process whose kernel encodes quality and similarity (Kathuria et al., 2016; Bıyık et al., 2019). Here: greedy MAP inference for $L = \operatorname{diag}(q)\, S \operatorname{diag}(q)$ with quality $q_i = \exp(\beta \mu_i / 2)$, $\beta = 1$, and $S_{ij} = \exp(-\|\phi(x_i) - \phi(x_j)\|^2 / (2\ell^2))$ an RBF similarity on the one-hot features ϕ, bandwidth ℓ the median pairwise distance over the candidate set. Selection is the standard Cholesky update: the gain of adding i is $\log d_i^2$ with d_i the residual after projecting out the selected columns. Candidate set capped at 2×10^4, half by acquisition score and half uniformly at random.

Gradient embedding (BADGE-style). Original: embed each candidate by the last-layer loss gradient at its most likely label, then seed a batch by k-means++ in that embedding (Ash et al., 2020); the Fisher reading of the same geometry is Ash et al. (2021). Here the surrogate is a regressor, so there is no label to take: for a linear model on the second-order basis the loss gradient at x is $\phi_2(x)$ scaled by the residual, and we use $s(x)\, \phi_2(x)$ with s the surrogate's own predictive standard deviation, so magnitude is uncertainty and direction is feature direction exactly as in the original. k-means++ seeding then selects N points. Candidate set capped at 2×10^4 as above. This is the one place where

"BADGE" would be misleading without the qualifier, since the published embedding is defined through a class label.

Redundancy-aware information (BatchBALD-style). Original: greedily maximise the mutual information between a batch's joint predictions and the parameters, which is submodular (Kirsch et al., 2019). Here the surrogate is a forest ensemble, and we take its Gaussian reading: the joint predictive entropy of a batch is $\frac{1}{2}\log\det(\Sigma_L + \sigma^2 I)$ and the mutual information is that minus the sum of the per-point noise entropies, so greedy maximisation is a log-determinant step on Σ_L. We build Σ from the surrogate's own per-point uncertainty and the same RBF correlation on one-hot features, so no second model is fitted and no posterior sampling is needed. Candidate set capped at 8×10^3, the objective being denser than the other two.

All three see the same pool, seed set, surrogate, leads and estimand set as the other nine rules, and are run at the same budgets (Table 5).

The candidate cap, and why it is not the cause. Greedy Bayesian optimal design costs $O(n_{\text{cand}}\, p^2)$ per selected point with $p = p_{\text{main}} + p_{\text{int}}$, which on a 20^4 space ($p = 2243$) makes the full 149,361-point pool uncomputable. The candidate set of these two rules is therefore capped by one stated compute budget applied through the same formula to every landscape, $n_{\text{cand}} \le 2.5 \times 10^{10}/p^2$ clipped to $[5000,\, 5 \times 10^6]$, half taken by acquisition score and half uniformly at random. That formula leaves the pool *uncapped* on every landscape except the two 20^4 ones: Poelwijk ($p = 92$) and PARD3 ($p = 1141$) are run on the whole pool, as are all synthetic configurations. Since D- and D_s-optimal certify nothing on the uncapped landscapes as well, the cap is not what produces the result; we flag it because it is the one asymmetry in the protocol, it falls on the two rules the paper criticises, and a reader is entitled to check it. No other rule is capped, because none needs to be.

Synthetic landscapes. f is generated as a sum of a centred additive part, a centred pairwise part over all coordinate pairs, and, when the order-≥ 3 share is nonzero, centred three-way tensors; the three parts are each standardised and then mixed to hit the requested variance shares exactly, and f is standardised. Unless stated, the pairwise share is 0.35 in the level sweep and 0.30 elsewhere, and the order-≥ 3 share is 0.20.

I ADDITIONAL RESULTS

The levels sweep, in full, with its confound. Figure 5 varies m over $(m, K) \in \{(2,12), (3,8), (5,5), (8,4), (12,4), (20,3)\}$, which holds the space size within a factor of seven but moves three things at once: m, the number of coordinates K, and the model dimension p_{int}, which rises from 66 to 1,083. The number of backgrounds per coordinate pair falls as K does, which works *against* the trend we report, so the measured decline is if anything conservative; but a clean level effect would need a sweep at fixed K and varying space size, which we did not run. The decline is also not monotone: all three concentrating rules rise at the single step $m = 5 \to 8$, which is exactly where K drops from 5 to 4. Sign error among the most confident claims rises with m for every rule, including those whose certification is identically zero at every m, so nothing in this sweep separates "sign error rises because certification fell" from "sign error rises because the task got harder"; we report the association and claim no mechanism.

Algorithm 1 RECTIFY: rectangle completion at the leads, a wrapper around any acquisition score.

Require: pool $\mathcal{P}$, score μ, uncertainty s, budget N, design fraction ρ; $L \leftarrow \emptyset$

1: **while** $|L| < \rho N$ **do**
2: take the next *anchor* x in decreasing μ not yet in L; among pairs (i, j) and levels (b'_i, b'_j) whose four corners all lie in $\mathcal{P}$, add the corners of the rectangle maximising mean $\mu + \beta s$ less a penalty for pairs already covered
3: **end while**
4: **return** L filled to N with the highest-μ unused pool points

Table 6: How well the matched control actually matches, per landscape. It is intended to hold the acquisition score and the distance profile fixed while removing rectangles. It succeeds on the binary landscape and only partly on the twenty-level ones, where avoiding rectangles forces it away from the top of the pool; and because an unavailable replacement falls back to the original point, it retains some rectangles rather than none. We report this instead of describing the control as matched.

	mean acquisition score			mean Hamming		rectangles		certified	
Landscape	RECTIFY	control	gap	RECTIFY	control	RECTIFY	control	RECTIFY	control
Poelwijk	2.08	2.03	-2%	3.93	4.28	521	48	0.272	0.103
GB1	2.62	1.80	-31%	2.43	3.13	3214	108	0.045	0.003
TrpB	3.19	2.37	-26%	2.31	2.80	3745	88	0.044	0.003
ParD3	1.77	1.36	-24%	2.21	2.63	2457	41	0.077	0.004

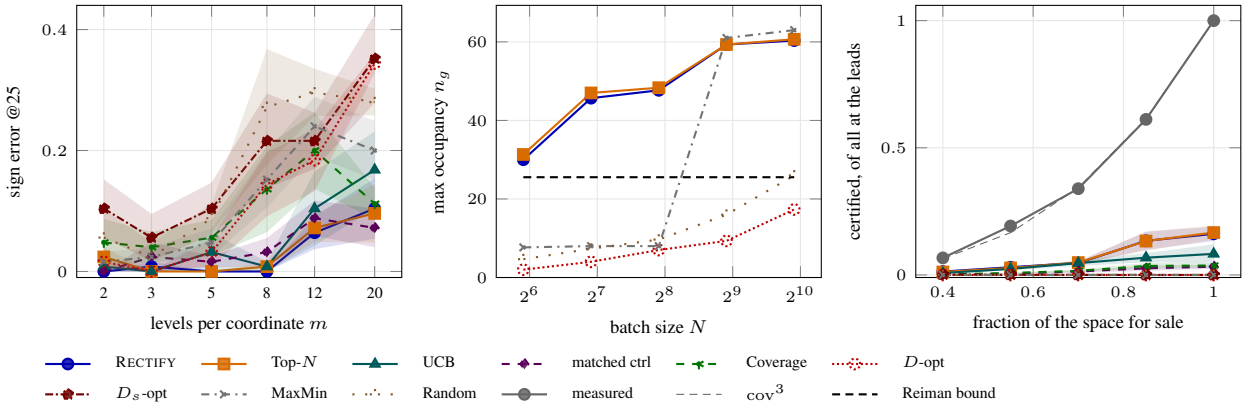


Figure 3: Three further sweeps. **Left:** the consequence of the certification deficit for claims, the sign-error rate among the 25 most confident local interactions each rule reports, against levels per coordinate. **Middle:** the mechanism behind the batch-size panel of Figure 5, the maximum per-background occupancy n_g against Reiman's bound $\bar{z} = 25.5$ (dashed); concentrating rules cross it almost at once while spreading rules stay below it at every budget. **Right:** the catalogue regime. A rectangle needs the lead plus three specific partners, so the fraction of local interactions that are purchasable at all follows cov^3 (dashed), and the share any rule certifies collapses with coverage even though the conditional rate does not.

J THE NEGATIVE RESULTS IN FULL

Held-out decision regret, per rule. Table 8 gives the column that Section 6(ii) rests on. The rectangle-*removing* matched control is the best rule of the nine on every landscape, and RECTIFY is second. Since the control is constructed by taking RECTIFY's batch and swapping points to destroy rectangles while holding the acquisition score as close as we can, this says as directly as our design allows that the rectangle geometry does not pay for itself in downstream decisions at fixed leads. It does not contradict the audit, which is about what a batch can *measure*, but it does bound what measuring buys, and a reader weighing whether to spend budget on geometry should weigh this column first.

An additional phase-grid stress test. Alongside the continuous sweep of Figure 5 we ran a 5×4 grid over the pairwise variance share and the assay noise, on the same 5^5 synthetic generator. We report it because it is unfavourable: RECTIFY's held-out regret is worse than the best other rule in 20 of 20 cells, mean margin -0.216 and worst -0.466 (Table 9). The grid is coarse in the quantity that turned out to matter, the order-≥ 3 share, which is why we replaced it rather than extended it; but replacing an experiment is not a reason to omit its result, and it reinforces Section 6(i) and (ii).

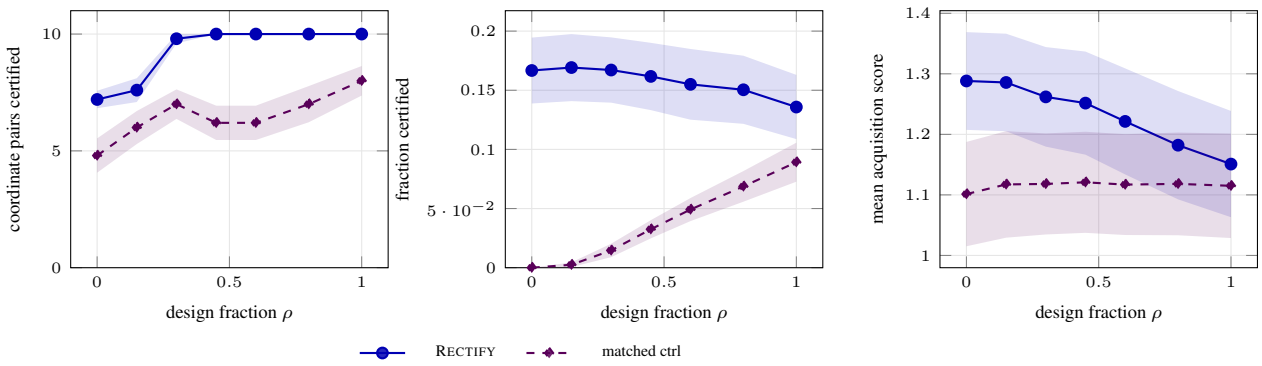


Figure 4: What the single knob ρ actually buys. **Left:** the number of distinct coordinate pairs for which the batch certifies at least one local interaction rises to the maximum of 10. **Middle:** the total fraction certified is flat, so ρ redistributes certification across pairs rather than increasing it. **Right:** the cost, a small monotone decline in the batch's mean acquisition score. The matched control is shown for comparison. Five seeds, shaded band one standard error.

Table 7: Model-class control. Claim quality and held-out decision regret when the refitted model is the second-order ridge and when it is a random forest. Certification is identical in the two halves because it is a property of the batch's geometry and not of any model, which is exactly the point of the audit. The claim-quality and regret columns, by contrast, are *not* stable to the model class: several ranks move, including RECTIFY's own regret rank. We therefore rest the audit on certification, which is model-free by construction, and treat the decision metrics as model-dependent evidence that should not be read as a ranking of rules.

	second-order ridge			random forest		
Rule	sign err @25	RMSE @25	regret	sign err @25	RMSE @25	regret
Random	0.112	0.730	0.897	0.256	0.957	0.867
MaxMin diversity	0.160	0.715	0.902	0.168	0.743	0.699
Coverage	0.080	0.663	0.882	0.080	0.843	0.743
Bayesian D-optimal	0.184	0.877	0.926	0.192	0.912	1.079
Bayesian D_s-optimal	0.216	0.795	1.055	0.248	1.082	0.912
Determinantal (DPP)	0.064	0.612	0.369	0.088	0.933	0.671
Gradient emb. (BADGE-style)	0.144	0.732	1.028	0.176	0.785	0.987
Info. batch (BatchBALD-style)	0.144	0.786	1.136	0.288	0.916	0.884
Greedy utility (Top-N)	0.008	0.477	0.608	0.056	0.766	0.719
UCB	0.032	0.562	0.547	0.104	0.841	0.738
Matched control	0.032	0.594	0.481	0.056	0.679	0.506
RECTIFY (control)	0.008	0.482	0.583	0.040	0.676	0.888

K TWO ABLATIONS THE AUDIT WOULD BE INCOMPLETE WITHOUT

The measure-or-model verdict does not depend on the ridge penalty. The comparison in Table 17 sets one free parameter, the penalty of the refitted second-order ridge, and a verdict that moved with it would be worth nothing. Table 10 sweeps it over 5 settings spanning 0.1 to 1000, and adds a ridge whose penalty is chosen by cross-validation on the batch, which is the fair capacity-matched version of the model arm. The sign of model − contrast is the same in every cell: the model wins on POELWIJK by between -0.28 and -0.30, and loses on GB1 by between +0.52 and +3.58, with the cross-validated ridge giving -0.28 and +2.25. Each row is reported for RECTIFY and for greedy utility, so this table doubles as the second-rule version of Table 17: the certified subset differs between the two rules and the sign does not.

The candidate cap is not what produces the zero. Greedy Bayesian optimal design costs $O(n_{\text{cand}}p^2)$ per selected point, so on the two 20^4 landscapes its candidate set is capped, and on GB1 that cap is 5,000 of 149,361 points. Two of the four landscapes are capped and two are not,

Table 8: Held-out decision regret at the leads, restricted to combinations each rule did not itself measure, averaged over the four landscapes and five seeds (lower is better). The control built to remove rectangles wins.

Rule	regret↓
Matched control	1.46
RECTIFY (constructive control)	1.96
Greedy utility (Top-N)	2.37
UCB	2.40
Random	2.89
Greedy coverage	3.05
Bayesian D_s-optimal	3.06
Bayesian D-optimal	3.11
MaxMin diversity	3.14

Table 9: The phase grid, in full. Held-out regret for RECTIFY against the best of the other eight rules in each cell; a negative margin means RECTIFY is worse. Three seeds per cell.

pairwise share	σ	RECTIFY	best other	its regret	margin
0.05	0.05	0.517	Matched control	0.282	-0.235
0.05	0.15	0.580	UCB	0.352	-0.228
0.05	0.30	0.735	Matched control	0.363	-0.373
0.05	0.60	0.466	Matched control	0.405	-0.061
0.15	0.05	0.631	Matched control	0.394	-0.237
0.15	0.15	0.521	Matched control	0.413	-0.108
0.15	0.30	0.455	Matched control	0.429	-0.026
0.15	0.60	0.586	Matched control	0.445	-0.142
0.30	0.05	0.368	Matched control	0.277	-0.091
0.30	0.15	0.610	Greedy utility (Top-N)	0.484	-0.127
0.30	0.30	0.904	Matched control	0.550	-0.354
0.30	0.60	0.813	MaxMin diversity	0.730	-0.083
0.50	0.05	0.554	Matched control	0.249	-0.305
0.50	0.15	0.521	Matched control	0.356	-0.164
0.50	0.30	0.505	Bayesian D-optimal	0.481	-0.024
0.50	0.60	1.008	Greedy utility (Top-N)	0.589	-0.419
0.70	0.05	0.723	Matched control	0.438	-0.285
0.70	0.15	0.665	Matched control	0.199	-0.466
0.70	0.30	0.575	Matched control	0.415	-0.160
0.70	0.60	1.096	Bayesian D-optimal	0.661	-0.435

which leaves the cap confounded with the landscape in Table 1; the control has to break that confound *within* one landscape.

Two obvious designs do not work. Raising the cap on the full pool costs $\sim 10^{11}$ flops per selected point at $20{,}000$ candidates, and would still compare two capped settings. Randomly subsampling the pool fails for a reason this paper supplies: a rectangle needs the lead plus three specific partners, so availability scales as the cube of coverage, and a $5{,}000$-of-$149{,}361$ pool leaves no purchasable estimand at all. What works is a *complete sub-landscape*: keeping the eight most frequent amino acids at each of the four sites gives an 8^4 space that is fully measured within GB1, where $p = 323$, the cap cannot bind, the pool is complete so estimands exist, and capped versus uncapped is a comparison inside one real landscape. Table 11 reports it, and certification stays at zero for D-optimal, so the zero is a property of the criterion and not of the compute budget we gave it. Two details are worth noting in passing: on this sub-landscape D-optimal has $d_{\min} = 1$ and still certifies nothing, at occupancy far below the threshold, while D_s-optimal runs well above the threshold and also certifies nothing, so the second and third mechanisms of Section 5.2 are both visible here on real data rather than only in the synthetic sweep. Eight levels is as far as this control goes: at twelve the space is $20{,}736$ points with $p = 771$, and greedy selection there is beyond the compute we have.

Table 10: Ridge-penalty ablation. RMSE of the contrast and of the refitted model on the local interactions each batch certified, by penalty α and by rule; Δ = model − contrast, so $\Delta < 0$ means modelling wins. Three seeds. The last row of each block uses a cross-validated penalty. The sign never changes.

		RECTIFY			greedy utility		
Landscape	α	contrast	model	Δ	contrast	model	Δ
Poelwijk	0.1	0.493	0.264	-0.229	0.593	0.253	-0.340
	1	0.493	0.266	-0.228	0.593	0.255	-0.338
	10	0.493	0.266	-0.228	0.593	0.257	-0.336
	100	0.493	0.253	-0.240	0.593	0.249	-0.344
	1000	0.493	0.249	-0.244	0.593	0.246	-0.346
	ridge-CV	0.493	0.265	-0.229	0.593	0.254	-0.339
GB1	0.1	0.519	1.140	+0.621	0.525	0.942	+0.417
	1	0.519	2.343	+1.824	0.525	2.369	+1.844
	10	0.519	3.583	+3.064	0.525	3.818	+3.293
	100	0.519	3.939	+3.420	0.525	4.173	+3.648
	1000	0.519	3.986	+3.467	0.525	4.221	+3.696
	ridge-CV	0.519	2.837	+2.318	0.525	2.716	+2.191

Table 11: Candidate-cap ablation on GB1, three seeds. Upper block: the full 149,361-point pool at the cap the audit uses. Lower block: complete sub-landscapes on which the optimal designs are uncapped and see every candidate.

setting	pool	candidates seen	rule	fraction certified
capped (the audit)	149,361	5,000	Bayesian D-optimal	0.0000
capped (the audit)	149,361	5,000	Bayesian D_s-optimal	0.0258
capped (the audit)	149,361	5,000	greedy utility	0.0654
uncapped	4,096	all	Bayesian D-optimal	0.0000
uncapped	4,096	all	Bayesian D_s-optimal	0.0000
uncapped	4,096	all	greedy utility	0.1201

L BLOCK RECTANGLES: WHAT A SEPARATED BATCH COULD STILL HAVE MEASURED

Proposition 3(i) forecloses two-coordinate rectangles, and by Lemma 1 those are a minority of the minimal additive-free measurements once $K \geq 3$. The rest toggle coordinate *blocks*, and a block rectangle $\{x, x^{S_1}, x^{S_2}, x^{S_1 \cup S_2}\}$ with $\min(|S_1|, |S_2|) \geq 2$ has pairwise distances $\{|S_1|, |S_2|, |S_1| + |S_2|\}$, all at least two. Separation therefore does *not* exclude it. The smallest witness is $\{0000, 1100, 0011, 1111\}$ in 2^4, which has $d_{\min} = 2$, no two-coordinate rectangle, and is annihilated exactly by the additive model matrix; it also lies inside the sum-zero code that Section C uses as the tight case for Proposition 3(ii), so the extremal separated batch is full of them.

This matters because it is the one way the audit's zeros could be an artefact of what we chose to count, and it cuts in the direction that would flatter our thesis. So we counted. For every batch we enumerate all pairs of points, all nontrivial bipartitions of their difference set, and test whether the two remaining corners are present; each quadruple is found once per diagonal and the counts are halved. Table 12 reports the result.

The answer has two halves, and the reviewers who raised this were right that the second one exists. *First*, the two-coordinate count in a separated batch is zero everywhere, with no exception in any batch we enumerated, which is what Proposition 3(i) asserts and all it asserts. *Second*, separated batches do contain block measurements, and the number depends on how deep one is willing to look: with the enumeration capped at $|S_1| + |S_2| \leq 4$ the real landscapes show none, but removing the cap on the 2^{13} landscape turns up 263 per batch, of which 2 touch a lead. The cap matters, so the last block of Table 12 reports the uncapped enumeration separately rather than letting the capped zeros stand as an unconditional claim.

We draw the narrow conclusion. A separated batch is not blind: it can measure whether two *groups* of changes interact, and on a 2^{13} space it can do so a few hundred ways. It cannot measure whether two *single* changes interact, and it almost never reaches the leads even for the coarse question. Coarse and fine are different estimands; Definition 1 names the fine one, and that is the one Proposition 3 forecloses.

Table 12: All support-four additive-free measurements inside each audited batch, averaged over three seeds. "classical" toggles two single coordinates, which is the object of Definition 1; "both ≥ 2" toggles two blocks of size at least two, which separation does *not* forbid; the last two columns restrict each to quadruples with a lead as a corner. "depth" is the cap on $|S_1| + |S_2|$, needed because the bipartition count grows as $2^{|D|-1}$; the final block removes it.

				anywhere		at a lead	
Landscape	Rule	depth	$d_{\min}$	classical	both ≥ 2	classical	both ≥ 2
GB1	MaxMin diversity	6	3.0	0	0	0	0
GB1	Random	6	1.0	0	0	0	0
GB1	Greedy utility	6	1.0	6341	0	954	0
ParD3	Bayesian D-optimal	6	2.0	0	0	0	0
ParD3	MaxMin diversity	6	2.0	0	0	0	0
ParD3	Random	6	1.0	1	0	0	0
ParD3	Greedy utility	6	1.0	5630	0	822	0
Poelwijk	Bayesian D-optimal	4	2.0	0	0	0	0
Poelwijk	MaxMin diversity	4	3.0	0	0	0	0
Poelwijk	Random	4	1.0	0	1	0	0
Poelwijk	Greedy utility	4	1.0	663	3378	147	866
TrpB	MaxMin diversity	6	3.0	0	0	0	0
TrpB	Random	6	1.0	0	0	0	0
TrpB	Greedy utility	6	1.0	8453	0	1140	0
synthetic 8^4	Bayesian D-optimal	4	1.3	0	755	0	2
synthetic 8^4	Bayesian D_s-optimal	4	1.0	677	1250	1	14
synthetic 8^4	MaxMin diversity	4	1.7	256	1160	0	16
synthetic 8^4	Random	4	1.0	18	478	0	0
synthetic 8^4	Greedy utility	4	1.0	17	485	1	22
Poelwijk	Bayesian D-optimal	13 (all)	2.0	0	189	0	0
Poelwijk	MaxMin diversity	13 (all)	3.0	0	339	0	5
all separated	**pooled, 33 batches**	–	≥ 2	0	3067	0	40

M SUPPORTING TABLES FOR THE AUDIT

Global rank against local certification. Table 13 is the column behind Section 5.1's central claim, on the binary landscape, the only one of the four where the second-order interaction block ($p_{\text{int}} = 78$) is small enough relative to the batch ($N = 160$) for full identification to be possible at all. The rank share and the certified fraction are strongly anti-correlated (Spearman -0.90, $p = 0.000$), and the anti-correlation is not driven by one family: a Fisher-embedding batch, a maximin batch, a random batch and both optimal designs all sit at rank one with zero certification.

Occupancy against the extremal threshold. Table 14 reports, per rule, the mean maximum per-background occupancy and how often a run exceeds Reiman's bound. We give it per rule rather than per group because the threshold differs by an order of magnitude between the binary landscape and the twenty-level ones, so any group average mixes incomparable quantities. 3 rules never exceed it on any run (Bayesian D-optimal, MaxMin diversity, Random) and 2 always do.

The bias identity, measured. Table 15 checks Proposition 4 against brute-force enumeration on each landscape. The prediction uses only the landscape's own ANOVA decomposition and the level counts; nothing is fitted to the measurement. The worst disagreement is 4%, on the 2^{13} landscape where the higher-order variance is spread over orders three to six and the pair-weighting $\binom{k}{2}/\binom{K}{2}$ therefore does the most work.

Table 13: Second-order interaction rank as a fraction of p_{int}, against the fraction of local interactions certified at the leads. Binary landscape, five seeds. Every rule that identifies the whole block certifies nothing; every rule that certifies anything identifies at most 0.974 of it.

Rule	$\text{rank}(G_L)/p_{\text{int}}$	fraction certified
Random	1.000	0.000
MaxMin diversity	1.000	0.000
Coverage	0.477	0.189
Bayesian D-optimal	1.000	0.000
Bayesian D_s-optimal	1.000	0.000
Determinantal (DPP)	0.731	0.000
Gradient emb. (BADGE-style)	1.000	0.000
Info. batch (BatchBALD-style)	0.974	0.000
Greedy utility (Top-N)	0.431	0.270
UCB	0.459	0.028
Matched control	0.500	0.103
RECTIFY (control)	0.446	0.273

Table 14: Per-background occupancy against Reiman's bound, averaged over the four landscapes and five seeds. "over" is the percentage of runs in which the batch's maximum occupancy exceeds the bound for that landscape.

Rule	mean $n_g^{\max}$	over (%)
Random	8	0
MaxMin diversity	6	0
Coverage	12	25
Bayesian D-optimal	4	0
Bayesian D_s-optimal	86	30
Determinantal (DPP)	115	60
Gradient emb. (BADGE-style)	18	5
Info. batch (BatchBALD-style)	58	15
Greedy utility (Top-N)	157	100
UCB	151	100
Matched control	47	25
RECTIFY (control)	126	95

N TESTING THE MEASURE-OR-MODEL BOUNDARY

Section 5.2 tests three separable claims. This appendix carries the tables behind them.

The four landscapes, and the weakest of the four tests. Table 17 is the comparison Section 5.2 summarises: the exact variance shares of each landscape, the two sides of Corollary 1 printed separately so the ratio can be checked, and the two estimators evaluated on the same local interactions at the leads. With four landscapes it is four binary calls, which is why the tests below and in Figure 2 move the boundary deliberately instead.

One pair at a time. Corollary 1 is indexed by a coordinate pair, so the natural sample size is pairs, not landscapes. For each (i, j) we compute $V_{\geq 3}^{(ij)}$ from the exact functional-ANOVA decomposition, the threshold $\sigma^2(1 - 1/m_i)(1 - 1/m_j)$, the RMS bias the identity predicts, $\sqrt{4c_i c_j V_{\geq 3}^{(ij)}}$, and the same quantity measured by enumerating up to 20,000 rectangles of that pair which exist in the landscape. Two things can fail independently: the identity, and the side of 2σ the pair falls on.

Two caveats on the order-≥ 3 share. It is a residual, so on the three landscapes without replicates it absorbs assay noise as well as genuine higher-order structure, and only POELWIJK has a measured σ. It is also collinear with the refitted ridge's parameter-to-observation ratio, which orders the model column of Table 17 *better* than the share alone does; the boundary of Corollary 1, which that ratio does not enter, is what breaks the tie between the two explanations. Separately, the contrast column

Table 15: Proposition 4 against measurement. "order ≥ 3 var" is the variance of $f - P_2 f$; "predicted" is $\sqrt{4 c_i c_j V_{\geq 3}^{(ij)}}$ averaged over coordinate pairs; "measured" is the root mean squared κ_{ij} of $f - P_2 f$ over rectangles drawn from the landscape.

Landscape	order ≥ 3 var	predicted RMS bias	measured	ratio
Poelwijk	0.127	0.416	0.412	0.992
ParD3	0.045	0.446	0.432	0.968
GB1	0.377	0.998	1.007	1.009
TrpB	0.387	1.095	1.088	0.994

Table 16: Where this work sits. "Local" means a contrast at a named background rather than a coefficient averaged over backgrounds; "certificate" means a quantity computable on a proposed batch that decides whether the estimand is measurable at all.

Line of work	Target	Local	Audits rules	Certificate
Double-mutant cycles	a local interaction	yes	no	no
Optimal design	model parameters	no	no	partial
Factorial designs, circuits	global contrasts	no	no	no
Active interaction discovery	interacting pairs	sometimes	no	no
This work	a named local interaction	yes	yes	yes

Table 17: **The landscapes, and the boundary they are predicted to fall on.** Left: exact functional-ANOVA variance shares. Middle: the two sides of Corollary 1 printed separately so the ratio can be checked, $V = \mathbb{E}_{ij}[V_{\geq 3}^{(ij)}]$ against $\mathrm{thr} = \mathbb{E}_{ij}[\sigma^2(1 - 1/m_i)(1 - 1/m_j)]$, averaged over coordinate pairs; below one the *exact projection* is predicted to beat the contrast. V differs from the order-≥ 3 share because it weights an order-k component by $\binom{k}{2}/\binom{K}{2}$. Right: the two estimators of the *same* local interactions on the subset RECTIFY*'s* batch certified (RMSE, five seeds); that subset is rule-dependent, so Section K repeats the columns for greedy utility, with the same sign everywhere.

			variance share			Corollary 1			RMSE of κ	
Landscape	variants	$K \times m$	additive	pairwise	ord. ≥ 3	V	thr.	V/thr.	contrast	model
ParD3	7,882	3×20	0.849	**0.105**	0.046	0.0437	0.0564	0.77	0.50	0.33
Poelwijk	8,192	13×2	0.137	**0.736**	0.127	0.0108	0.0165	0.66	0.50	0.27
GB1	149,361	4×20	0.193	**0.429**	0.378	0.2236	0.0564	3.96	0.48	2.15
TrpB	159,129	4×20	0.256	**0.356**	0.388	0.2705	0.0564	4.80	0.54	2.27

of Table 17 sits within 8% of 2σ throughout, which Proposition 1(2) forces under homoscedastic noise; it therefore checks the noise model rather than the landscapes and we do not count it as a test.

Moving the threshold. The boundary is a function of σ, and on three of the four landscapes σ is a stated simulation setting rather than a measurement, so the verdict could turn on that choice. Table 19 re-evaluates it across the noise sweep of Table 5, which spans a factor of twenty in σ and therefore a factor of four hundred in the threshold. Both landscapes are predicted to change side inside the range.

Does incomplete coverage carry the verdict? Three of the four landscapes are near-complete rather than complete, so the decomposition is computed from the cells that exist. If missingness were related to fitness, $V_{\geq 3}^{(ij)}$ could be biased. Table 20 recomputes the verdict on the largest fully observed sub-product-space of each landscape, where no cell is imputed and nothing is averaged over a hole, and after deleting a further 5, 10 and 20 percent of observed cells at random. The deletion ablation is omitted for the binary landscape, which is complete and where the Moebius inversion over 2^{13} subsets is ill-conditioned under deletion.

Table 18: Per-pair test of the identity and of the oracle boundary. "above" counts pairs with $V_{\geq 3}^{(ij)}$ above the threshold, i.e. predicted to favour direct measurement. Relative error is between the identity's predicted RMS bias and the enumerated one. The last column counts pairs whose predicted side of 2σ is the observed side.

Landscape	pairs	above thr.	median |rel. err.| (%)	worst (%)	side correct
GB1	6	6	2.8	5.5	6/6
ParD3	3	0	1.5	2.3	3/3
Poelwijk	78	15	0.8	3.1	78/78
TrpB	6	6	0.9	2.1	6/6

Table 19: The oracle boundary across the noise sweep. V is a property of the landscape; the threshold moves with σ^2. "contrast" and "fitted" are the measured RMSE of the four-point estimator and of the capacity-matched ridge at the leads, three seeds. The three disagreements are the two departures named in Section 5.2 and no others.

Landscape	σ	V	thr.	V/thr.	predicted	contrast	fitted	measured	agree
GB1	0.0500	0.2246	0.0023	99.54	contrast	0.088	2.493	contrast	yes
GB1	0.1284	0.2246	0.0149	15.09	contrast	0.292	2.289	contrast	yes
GB1	0.2568	0.2246	0.0595	3.77	contrast	0.472	2.163	contrast	yes
GB1	0.5136	0.2246	0.2381	0.94	model	1.131	2.366	contrast	no
GB1	1.0000	0.2246	0.9025	0.25	model	2.205	2.315	contrast	no
Poelwijk	0.0500	0.0108	0.0006	17.29	contrast	0.103	0.204	contrast	yes
Poelwijk	0.1284	0.0108	0.0041	2.62	contrast	0.262	0.233	model	no
Poelwijk	0.2568	0.0108	0.0165	0.66	model	0.533	0.241	model	yes
Poelwijk	0.5136	0.0108	0.0659	0.16	model	1.028	0.236	model	yes
Poelwijk	1.0000	0.0108	0.2500	0.04	model	2.101	0.362	model	yes

O THE LEVEL SWEEP WITH THE NUMBER OF COORDINATES HELD FIXED

The sweep of Figure 5 shrinks K as m grows so that the space size and the model dimension stay comparable. That is the right control for "is this a capacity effect?" and the wrong one for "is this a level effect?", because m, K and p_{int} then move together. Table 21 holds $K = 4$ and moves only m, in two arms: one at a fixed budget, where the batch covers a shrinking fraction of a growing space and occupancy falls as N/m^2 against a threshold growing as $m^{3/2}$; and one at a fixed fraction of the space, where occupancy grows as m^2 instead. If the level count were the mechanism both arms would fall. They do not: certification is unrelated to m ($\rho = -0.06$, $p = 0.63$) and strongly related to occupancy against the threshold ($\rho = 0.80$, $p = 10^{-15}$), which is what Proposition 2 predicts and what the confounded sweep could not show.

P WHICH REGIME MAKES IT WORSE: LEVELS, BUDGET, COVERAGE

Section 5.2 states the three mechanisms and the numbers that separate them. This section carries the sweeps behind those numbers.

The sweeps. Levels, or occupancy? Certification falls as m grows in the left panel, from 0.190 to 0.030 for the concentrating rules while the separating rules sit at 0.000 throughout. But that sweep shrinks K as m grows to hold the space comparable, so m, K and p_{int} move together and the trend is an association. Holding $K = 4$ and sweeping m from 5 to 10 in two arms, one at a fixed budget and one at a fixed fraction of the space, settles it: over 63 rule-by-m cells certification is unrelated to the level count ($\rho = -0.06$, $p = 0.63$) and strongly related to per-background occupancy against the Zarankiewicz threshold ($\rho = 0.80$, $p = 10^{-15}$), averaging 0.082 certified above the threshold against 0.003 below it, with m still irrelevant within the cells above ($\rho = 0.08$, $p = 0.66$; Table 21). The mechanism is the one Proposition 2 names, and the apparent level effect is what a fixed budget does to occupancy in a growing space.

Table 20: Sensitivity of the measure-or-model verdict to incomplete coverage. 0 of 17 variants change the predicted winner.

Landscape	variant	cells	complete (%)	V/thr.	predicted
GB1	as reported	149361	93.4	3.96	contrast
GB1	complete subgrid	78030	100.0	3.75	contrast
GB1	drop 5%	141957	88.7	3.94	contrast
GB1	drop 10%	134620	84.1	3.87	contrast
GB1	drop 20%	119607	74.8	3.76	contrast
ParD3	as reported	7882	98.5	0.77	model
ParD3	complete subgrid	4352	100.0	0.61	model
ParD3	drop 5%	7494	93.7	0.71	model
ParD3	drop 10%	7087	88.6	0.66	model
ParD3	drop 20%	6306	78.8	0.51	model
Poelwijk	as reported	8192	100.0	0.66	model
Poelwijk	complete subgrid	8192	100.0	0.66	model
TrpB	as reported	159129	99.5	4.80	contrast
TrpB	complete subgrid	69632	100.0	3.60	contrast
TrpB	drop 5%	151249	94.5	4.76	contrast
TrpB	drop 10%	143438	89.6	4.66	contrast
TrpB	drop 20%	127449	79.7	4.55	contrast

Table 21: Certification and occupancy with $K = 4$ fixed, five seeds. $n_g^{\max}/z$ is the per-background occupancy against Reiman's bound. Greedy utility concentrates and stays above the threshold at every m in both arms; random selection stays below it and certifies nothing. The level count moves by a factor of two without moving certification.

			greedy utility		random	
Arm	m	N	certified	$n_g^{\max}/z$	certified	$n_g^{\max}/z$
fixed-budget	5	150	0.106	1.55	0.003	0.97
fixed-budget	6	150	0.194	1.89	0.000	0.61
fixed-budget	8	150	0.111	1.78	0.000	0.33
fixed-budget	10	150	0.114	1.90	0.000	0.19
fixed-coverage	5	38	0.126	1.23	0.000	0.37
fixed-coverage	6	78	0.144	1.77	0.000	0.39
fixed-coverage	8	246	0.136	2.02	0.000	0.40

Batch size, and three ways to certify nothing. The first objection to Table 1 is arithmetic: with K=4, m=20 and N=240 a coordinate pair leaves 400 backgrounds, so a spreading rule has under one point per background whatever its criterion. Sweeping N by a factor of 16× on an 8^4 space (Figure 5) does not rescue it, and separates three reasons a batch certifies nothing. *(i) No rectangles at all, because the batch separates:* maximin holds $d_{\min}$=2 with *exactly* zero rectangles to N=400 and loses separation only at N=440, short of the capacity $8^4/8 = 512$ of Proposition 3(ii) because a greedy rule is not an optimal distance-two code (Section D). *(ii) Too few, because occupancy stays under the threshold:* D-optimal stops separating at N=240, so Proposition 3 is silent beyond that, yet its occupancy climbs only from 2 to 17 and never reaches z, so no rectangle is ever forced. It certifies 0.000 even at N=960, where it does hold 128 accidental rectangles, 23 of them anchored at a high-scoring point: at the top of the budget range it is leaving mechanism (ii) for mechanism (iii) below, which is the handover the sweep is there to show. *(iii) Plenty of rectangles, in the wrong place:* D_s-optimal concentrates past z at *every* budget and builds 643 rectangles at the largest, yet certifies exactly zero, placing them where its criterion finds information rather than at the leads. The three modern rules split across the same three: the Fisher-embedding batch certifies nothing until N=960, exactly where its occupancy crosses z; the determinantal batch crosses earliest, at N=240, and certifies the most, 0.254; and the redundancy-aware batch reaches 1.9 times z yet certifies nothing at any budget. Certifying is thus two properties, having rectangles and having them where the question is asked, and only (i) and (ii) are computable from the batch alone.

Pool coverage. The purchasable fraction scales as the cube of pool coverage (Figure 3): with a sparse catalogue, widen the pool rather than enlarge the batch.

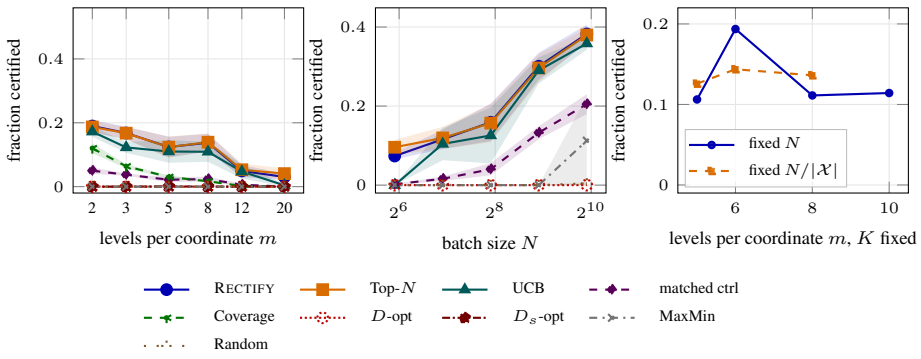


Figure 5: All synthetic, band one standard error; five seeds for the left panel, three for the middle. **Left:** levels per coordinate with K shrinking to hold the space comparable, the confounded sweep; the separating rules sit at zero throughout. **Middle:** batch size over a factor of 16×; the gap widens rather than closing, the two optimal designs staying at zero for opposite reasons (Section 5.2), and the seed set grows with N too. **Right:** the same axis with $K = 4$ held fixed, greedy utility in both arms; certification does not fall with m once the confound is removed (Table 21). The shared legend is for the left and middle panels.

Q THE OPERATING ENVELOPE IN FULL

Section 6 states six results that mark where certification stops. This section gives the detail behind each.

(i) The knob does not buy more certification. RECTIFY is behind greedy utility on 3 of the four landscapes, and certification is flat in ρ: 0.167 at $\rho = 0$, which *is* greedy utility, against 0.162 at $\rho = 0.45$. What ρ buys is *which* pairs are certified, the coordinate pairs reached rising from 7.2 to 10.0 of 10 for a 2.8% cut in acquisition score (Figure 4).

(ii) How far from matched the control is. Table 6 reports it per landscape: the acquisition-score gap runs from 2% on the binary landscape to 31% on the twenty-level ones, the mean Hamming profile also moves, and the control is rectangle-reducing rather than rectangle-free, cutting the rectangle count by at least 91%. A comparison that changes three things at once cannot attribute the regret difference to one of them; we read Table 8 as an intervention stress test.

(iv) Where the reliability advantage inverts. On POELWIJK the group means invert at 2.0× the measured noise, beyond which random and maximin report more reliable signs: a certified contrast has variance $4\sigma^2$ with no shrinkage, while a regularised model degrades gracefully. On GB1 no inversion occurs anywhere out to $\sigma = 1.00$. This is the same trade Corollary 1 prices, seen from the reliability side rather than the error side, and it is why the envelope is stated in terms of both.